



 Latest updates: <https://dl.acm.org/doi/10.1145/3746252.3761146>

RESEARCH-ARTICLE

Structure-prior Informed Diffusion Model for Graph Source Localization with Limited Data

HONGYI CHEN, Tsinghua University, Beijing, China

JINGTAO DING, Tsinghua University, Beijing, China

XIAOJUN LIANG, Peng Cheng Laboratory, Shenzhen, Guangdong, China

YONG LI, Tsinghua University, Beijing, China

XIAOPING ZHANG, Tsinghua University, Beijing, China

Open Access Support provided by:

Tsinghua University

Peng Cheng Laboratory



PDF Download
3746252.3761146.pdf
15 February 2026
Total Citations: 1
Total Downloads: 110

Published: 10 November 2025

[Citation in BibTeX format](#)

CIKM '25: The 34th ACM International Conference on Information and Knowledge Management
November 10 - 14, 2025
Seoul, Republic of Korea

Conference Sponsors:
SIGWEB
SIGIR

Structure-prior Informed Diffusion Model for Graph Source Localization with Limited Data

Hongyi Chen*
Shenzhen Key Laboratory of
Ubiquitous Data Enabling, Shenzhen
International Graduate School
Tsinghua University
Shenzhen, China
chenhy23@mails.tsinghua.edu.cn

Jingtao Ding
Department of Electronic Engineering
Tsinghua University
Beijing, China
dingjt15@tsinghua.org.cn

Xiaojun Liang
Pengcheng Laboratory
Shenzhen, China
liangxj@pcl.ac.cn

Yong Li
Department of Electronic Engineering
Tsinghua University
Beijing, China
liyong07@tsinghua.edu.cn

Xiao-Ping Zhang†
Shenzhen Key Laboratory of
Ubiquitous Data Enabling, Shenzhen
International Graduate School
Tsinghua University
Shenzhen, China
xpzhang@ieee.org

Abstract

Source localization in graph information propagation is essential for mitigating network disruptions, including misinformation spread, cyber threats, and infrastructure failures. Existing deep generative approaches face significant challenges in real-world applications due to limited propagation data availability. We present SIDSL (Structure-prior Informed Diffusion model for Source Localization), a generative diffusion framework that leverages topology-aware priors to enable robust source localization with limited data. SIDSL addresses three key challenges: unknown propagation patterns through structure-based source estimations via graph label propagation, complex topology-propagation relationships via a propagation-enhanced conditional denoiser with GNN-parameterized label propagation module, and class imbalance through structure-prior biased diffusion initialization. By learning pattern-invariant features from synthetic data generated by established propagation models, SIDSL enables effective knowledge transfer to real-world scenarios. Experimental evaluation on four real-world datasets demonstrates superior performance with 7.5-13.3% F1 score improvements over baselines, including over 19% improvement in few-shot and 40% in zero-shot settings, validating the framework's effectiveness for practical source localization. Our code can be found here.

CCS Concepts

• **Networks** → *Network dynamics*; • **Information systems** → *Social networks*; **Data mining**.

*Also with Pengcheng Laboratory.

†Corresponding author



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

CIKM '25, Seoul, Republic of Korea

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2040-6/2025/11

<https://doi.org/10.1145/3746252.3761146>

Keywords

Diffusion Model; Graph Source Localization; Information Diffusion

ACM Reference Format:

Hongyi Chen, Jingtao Ding, Xiaojun Liang, Yong Li, and Xiao-Ping Zhang. 2025. Structure-prior Informed Diffusion Model for Graph Source Localization with Limited Data. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*, November 10–14, 2025, Seoul, Republic of Korea. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3746252.3761146>

1 Introduction

Graph information propagation issues have far-reaching societal consequences. Quickly identifying disruption sources is critical for impact mitigation. By analyzing snapshots of affected networks, we can trace spread origins, essential for managing disease outbreaks [3, 25], network security [15], and infrastructure failures [1].

Early methods [18, 21, 23, 27, 45, 46] for source localization in graphs rely on topology-derived metrics applicable only to specific propagation patterns like the Susceptible-Infected (SI) or Independent Cascade (IC) models. Wang et al. [36] introduced label propagation based on source centrality, but with limited expressiveness and scalability in encoding topological information. Data-driven methods [4, 9, 33] overcome these limitations as they directly learn graph neural networks (GNNs) to capture propagation processes but neglect propagation indeterminacy [20]. Recent deep generative models, including variational autoencoders [20], normalization flows [39], and diffusion models [11, 40], quantify this indeterminacy by learning empirical data distributions, advancing state-of-the-art performance.

However, collecting real-world propagation data for deep generative methods is difficult and costly, requiring models that adapt to limited data environments. This poses three main challenges. **First, real-world graphs exhibit unknown propagation patterns that are challenging to characterize with limited data.**

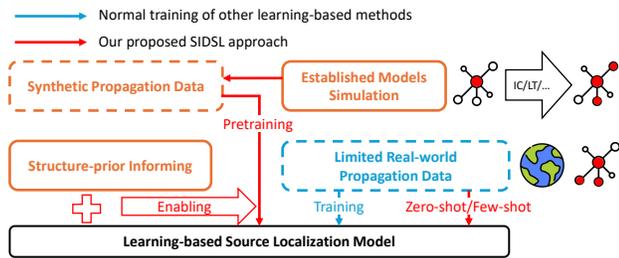


Figure 1: The proposed SIDSLS approach.

Existing methods [4, 20, 33, 40] rely purely on data for pattern understanding, limiting generalization to unseen scenarios. **Second, complex interrelations between propagation patterns and graph topology are difficult to capture with limited data.** Existing methods require extensive labeled data from target networks to account for the structural heterogeneity’s impact on propagation patterns. This dependence causes underperformance with limited training samples, hindering practical applicability. **Third, inherent class imbalance between source and non-source nodes becomes more harmful under data scarcity, compromising identification accuracy.** This challenge remains unaddressed by existing solutions and requires comprehensive handling of both abundant and limited data scenarios through novel approaches that prevent model degeneration towards non-source predictions.

Therefore, in this paper, we propose a novel generative diffusion framework, namely **Structure-prior Informed Diffusion Model for Source Localization (SIDSLS)**, which leverages topology-aware priors to achieve robust source localization with limited propagation data collected from real environments. SIDSLS employs a denoising diffusion model guided by structural priors to predict source node distributions from graph topology and observed node states.

To tackle the first challenge, we leverage graph structure-based source estimations, generated through graph label propagation, which identify potential sources based on locally highest infection values. Integrated into the denoising network, these topology-aware priors provide stable guidance across different propagation patterns, enhancing generalization to unseen propagation scenarios. To address the second challenge, we design a propagation-enhanced conditional denoiser with a GNN-parameterized label propagation module (GNN-LP) that combines label propagation for tracing infection pathways with GNN for efficient topology-aware feature extraction, enabling effective learning of topology-propagation relationships from limited data. To tackle the third challenge, we develop a structure-prior biased diffusion process that comprehensively handles the fundamental source/non-source imbalance by initializing from structure-based estimations, providing a complete solution that advances imbalance resolution in both data-rich and data-scarce scenarios.

The above designs work together in our structure-prior informed diffusion framework to capture stable structural patterns independent of specific propagation dynamics, allowing our model to learn pattern-invariant features using synthetic propagation data simulated from established models [13, 14, 16, 22]. This allows effective knowledge transfer to real-world scenarios through few-shot or zero-shot learning (Figure 1).

Our contributions are summarized as follows:

- (1) We propose SIDSLS, a structure-prior informed diffusion framework that uniquely integrates structural priors with diffusion models to effectively address the challenge of limited labeled data in source localization. Unlike existing approaches requiring abundant training data, our framework leverages domain knowledge and topology-aware priors for robust generalization to unknown propagation patterns in real-world low-data scenarios.
- (2) We introduce a series of innovative techniques, including a propagation-enhanced conditional denoiser with GNN-LP module and structure-prior biased denoising, which work synergistically to handle structural heterogeneity and class imbalance issues.
- (3) We evaluate SIDSLS’s performance across four real-world datasets and demonstrate its superior effectiveness in source identification tasks. Our method achieves superior performance across four real-world datasets, outperforming baselines by 7.5-13.3% in F1 scores, with over 19% improvement in few-shot and 40% in zero-shot settings. Additional evaluations further validate SIDSLS’s effective synthetic-to-real transfer capability in improving sample efficiency and reducing training time.

2 Related Works

2.1 Source Localization

Source localization, the inverse problem of information propagation in networks, infers initial sources from diffused observations, with applications such as rumor source identification and power grid failure detection [28, 30]. Early approaches are rule-based and rely on metrics or heuristics derived from the network’s topology for source identification [26, 45, 46]. For instance, Shah and Zaman [26] developed a rumor-centrality-based maximum likelihood estimator for the Susceptible-Infected (SI) [16] model. These methods, however, inadequately encode topology. Subsequently, deep learning methods emerged [4, 8, 18, 21, 33, 36], but most struggled to model source location uncertainty in stochastic propagation. To overcome this, deep generative models have been adopted [11, 20, 35, 39, 40]. SLVAE [20] utilizes Variational Auto-Encoders (VAEs) and optimizes the posterior for better prediction. However, VAEs struggle with complex propagation patterns. DDMSL [40] models the Susceptible-Infected-Recovered (SIR) [16] infection process into Diffusion Models (DM) [7, 38], and designs a reversible residual block based on Graph Convolutional Networks (GCNs) [17]. However, it requires additional intermediate propagation data and cannot be generalized to other propagation patterns. In contrast to existing methods that demand extensive training data, our novel framework leverages domain knowledge for robust localization with minimal labeled data—a significant advance for real-world scenarios. A comparison of typical methods is in Appendix B.

2.2 Typical Propagation Models

Information propagation models estimate information spread in networks, applied in event prediction [43], adverse event detection [34], disease spread prediction [31], and so on. Two main categories exist: infection models and influence models. Infection models such as Susceptible-Infected (SI) and Susceptible-Infected-Susceptible (SIS) manage state transitions in networks [13, 16]. Infected nodes

attempt to infect neighbors with probability β , while in SIS models, infected nodes revert to susceptible with probability λ . The Susceptible-Infected-Recovered (SIR) model adds a recovered state. Influence models include Independent Cascade (IC) and Linear Threshold (LT) [14]. In IC, newly activated nodes attempt once to activate neighbors based on edge weights. LT models activate nodes when accumulated neighbor influence exceeds a threshold.

3 Preliminaries

3.1 Problem Formulation

Our research problem is formulated as follows. Given an undirected social network $\mathcal{G} = (V, E)$ where V is the node set, E is the edge set, and $Y = \{Y_1, \dots, Y_{|V|}\}$ is an infection state of all nodes in \mathcal{G} , which describes that a subset of nodes in \mathcal{G} have been infected. Each $Y_i \in \{1, 0\}$ denotes the infection state of node $v_i \in V$, where $Y_i = 1$ indicates that v_i is infected and otherwise $Y_i = 0$ indicates it is uninfected. We aim to find the original propagation source \hat{X} from the propagated observation Y , so that the loss with the ground Truth source set $X^* \in \{1, 0\}^{|V| \times 1}$ is minimized, i.e. $\hat{X} = \operatorname{argmin}_X \|X - X^*\|_2^2$. To account for the uncertainty in source localization, we need to construct a probabilistic model $P(X|Y, \mathcal{G})$, which can be used to sample for the final prediction.

3.2 Label Propagation based Source Identification

In realistic situations, the intractable propagation process does not have an explicit prior, and it is also challenging to value appropriate parameters for the pre-selected underlying propagation model. To address this, Wang et al. [36] propose Label Propagation based Source Identification (LPSI). Since LPSI investigates the same problem as our research, we use it as a baseline to compare performance. LPSI captures source centrality characteristics in the method design. The centrality of sources shows that nodes far from the source are less likely to be infected than those near it [27], which can also be observed in the real-world data by our analysis in the Appendix E. Based on these ideas, they propose to perform label propagation on the observation state of the network. By setting $Y[Y = 0] = -1$ and $\mathcal{Z}^{t=0} \leftarrow Y$, the iteration of label propagation and the convergence states are as follows:

$$\mathcal{Z}_i^{t+1} = \alpha \sum_{j:j \in \mathcal{N}(i)} S_{ij} \mathcal{Z}_j^t + (1 - \alpha) Y_i. \quad (1)$$

\mathcal{Z} finally converges to: $\mathcal{Z}^* = (1 - \alpha)(I - \alpha S)^{-1} Y$, where $S = D^{-1/2} A D^{-1/2}$ is the normalized weight matrix of graph \mathcal{G} , α is the fraction of label information from neighbors, and $\mathcal{N}(i)$ stands for the neighbor set of the node i . After obtaining the converged label matrix \mathcal{Z}^* , one node can be identified as a source when its final label is larger than its neighbors. While node labels alone cannot fully capture complex structural information, this method still effectively identifies structural patterns related to source centrality (Appendix E). In our work, we leverage the idea of LPSI to generate structural guidance for efficient source identification, and details can be found in Section 4.

4 SIDSL: the Proposed Method

4.1 Diffusion Model for Source Localization

To capture the indeterminacy of the ill-posed localization problem, it is essential to build a probabilistic model that can also leverage the topological information in the graph structure. We consider using the generative diffusion model (DM) framework to tackle this requirement by modifying it as a source predictor, which classifies nodes as source or non-source. The diffusion model aims to learn source distributions by gradually adding Gaussian noise in samples in the forward process and learning to reverse this process using denoising networks conditioned on observations and graph structure.

The forward process of the diffusion model is to gradually corrupt the initial source labels ($X_0 = X$) by adding Gaussian noise over n timesteps, which can be formulated as:

$$p(X_{1:n}|X_0, Y, \mathcal{G}) = \prod_{t=1}^n p(X_t|X_{t-1}, Y, \mathcal{G}), \quad (2)$$

$$p(X_t|X_{t-1}, Y, \mathcal{G}) = \mathcal{N}(X_t; \sqrt{1 - \beta_t} X_{t-1}, \beta_t I),$$

where β_t is the noise schedule, X_0 the initial source labels, Y the observations, \mathcal{G} the graph structure, and X_t the noisy features at timestep t . At $t = n$, X_t becomes pure Gaussian noise, while the reverse process leverages Y and \mathcal{G} to guide denoising.

In the reverse process, it gradually transforms pure Gaussian noise into the original source labels, generating samples $\hat{X} \in [0, 1]^{|V| \times 1}$ representing source probabilities, with final predictions obtained by thresholding. To enhance stability across different propagation patterns, we leverage graph structure-based source estimations as conditional priors. These priors $X_{est} = \mathcal{Z}^*$, standing for potential sources, are obtained through label propagation (Section 3.2), aggregating infection values through the graph's normalized weight matrix to capture source centrality and structural importance.

The reverse process can be formulated as:

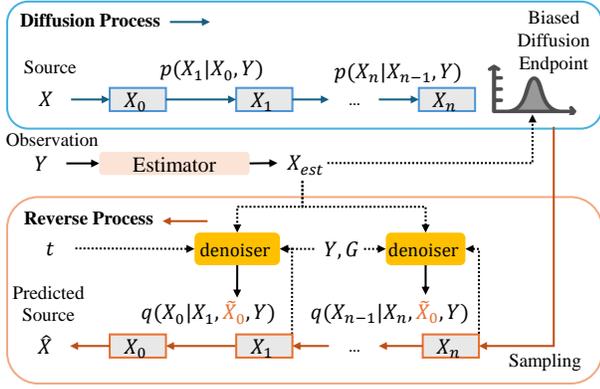
$$q(X_{n-1:0}|X_n, Y, \mathcal{G}) = \prod_{t=n}^1 q(X_{t-1}|X_t, Y, \mathcal{G}), \quad (3)$$

$$q(X_{t-1}|X_t, Y, \mathcal{G}) = \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t, Y, \mathcal{G}, X_{est}), \sigma_t^2 I),$$

where μ_θ is parameterized by a denoising network f_θ that predicts the mean of the Gaussian distribution and σ_t is the predicted variance. The iterative diffusion process enables fine-grained prior knowledge integration through progressive denoising. By incorporating topology-aware priors X_{est} , our framework combines structural knowledge with data-driven learning, enhancing generalization across propagation scenarios.

4.2 Structure-prior Biased Denoising Process

Source nodes constitute a small minority of total graph nodes [2]. With limited training data, this severe class imbalance leads models to degenerate towards predicting all nodes as non-sources. Following Han et al. [6], we propose a structure-prior biased denoising scheme that sets the prior to structure-based source estimations rather than random noise (Figure 2). This leverages two insights:


Figure 2: The framework of SIDL.

(1) structure-based estimations capture inherent source properties like centrality, providing more informative priors than random noise, and (2) initializing denoising from these estimations biases exploration toward regions with higher source likelihood. Formally, this approach decomposes the posterior as $p(X_0|Y, \mathcal{G}) \propto p_{\text{prior}}(X_{\text{est}}|Y, \mathcal{G}) \cdot p(X_0|X_{\text{est}})$, where the structural prior p_{prior} anchors learning in high-probability source regions, enabling the denoiser $p(X_0|X_{\text{est}})$ to focus on lower-variance refinements that are learnable with scarce data. This modified forward process shifts the diffusion trajectory from converging to noise toward structure-informed regions, while preserving generative correctness as our sampling process matches the modified forward posterior’s functional form. These topology-aware priors prevent model collapse while reducing data requirements for accurate source learning.

Specifically, we first modify the mean of the diffusion endpoint (or reverse starting point) as the graph structure-based source estimation $X_{\text{est}} = X_{\text{est}}(Y, \mathcal{G})$ instead of using standard Gaussian noise, i.e.:

$$p(X_n|Y, \mathcal{G}) = \mathcal{N}(X_{\text{est}}(Y, \mathcal{G}), I). \quad (4)$$

According to the original notation in Ho et al. [7], the Markov transition should be modified as:

$$p(X_t|X_{t-1}, Y, \mathcal{G}) = \mathcal{N}(\sqrt{1-\beta_t}X_{t-1} + (1-\sqrt{1-\beta_t})X_{\text{est}}, \beta_t I), \quad (5)$$

which derives the closed-form distribution with arbitrary t :

$$p(X_t|X_0, Y, \mathcal{G}) = \mathcal{N}(\sqrt{\alpha_t}X_0 + (1-\sqrt{\alpha_t})X_{\text{est}}(Y, \mathcal{G}), (1-\alpha_t)I), \quad (6)$$

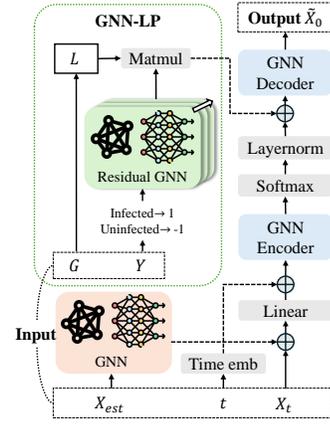
where $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$.

In the reverse denoising process, the reverse Markov denoiser $q(X_{t-1}|X_t, Y, \mathcal{G})$ recovers the original data. DM framework trains the parameterized denoiser to fit the ground truth forward process posterior:

$$\begin{aligned} p(X_{t-1}|X_t, X_0, Y, \mathcal{G}) &= p(X_{t-1}|X_t, X_0, X_{\text{est}}(Y, \mathcal{G})) \\ &= \mathcal{N}(\tilde{\mu}(X_t, X_0, X_{\text{est}}(Y, \mathcal{G})), \tilde{\beta}_t I), \end{aligned} \quad (7)$$

where

$$\begin{aligned} \tilde{\mu}(X_t, X_0, X_{\text{est}}(Y, \mathcal{G})) &:= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} X_0 + \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t} X_t + \\ &\quad \left(1 + \frac{(\sqrt{\bar{\alpha}_t} - 1)(\sqrt{\alpha_t} + \sqrt{\bar{\alpha}_{t-1}})}{1-\bar{\alpha}_t}\right) X_{\text{est}}, \quad (8) \\ \tilde{\beta}_t &:= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t. \end{aligned}$$


Figure 3: The architecture of the denoising network.

This formulation requires that the denoiser f_θ learns to predict X_0 given the modified trajectory, ensuring the sampling posterior $q_\theta(X_{t-1}|X_t, X_{\text{est}})$ matches the ground-truth form $p(X_{t-1}|X_t, X_0, X_{\text{est}})$ with X_0 replaced by the denoiser’s prediction. The denoising network outputs the estimated source vector $\tilde{X}_0 := f_\theta(X_t, X_{\text{est}}, Y, \mathcal{G}, t)$ to calculate the posterior for step-by-step denoising. The denoising network f_θ can be trained by the simple L2 loss function [7]:

$$L(\theta) = \mathbb{E}_{X_0 \sim p(X_0|\cdot), t, \epsilon} \|X_0 - f_\theta(X_t, t, \cdot)\|_2^2, \quad (9)$$

where \cdot represents the conditional inputs.

4.3 Propagation-enhanced Conditional Denoiser

4.3.1 Denoising Network Architecture. The architecture of our denoising network is shown in Figure 3.

Encoding the noisy input and soft labels. The pre-estimated X_{est} is forwarded through a multi-layer GNN to capture the hidden message with graph structural information. Subsequently, it is added to the noisy input X_t and passed through a linear layer. The final input for the GNN encoder is $Z_e = \text{Linear}(\text{GNN}(X_{\text{est}}) \oplus X_t) \oplus \text{Emb}(t)$, where for the denoising step t , we use the classical sinusoidal embedding [32]. The \oplus indicates element-wise sum. Z_e is then passed through a GCN-based encoder and is smoothed through a softmax function σ and layer normalization: $Z_d = \text{LN}(\sigma(\text{GNN}(Z_e)))$.

Softmax and layer normalization operations are then used to improve the network’s representational capacity and convergence performance, resulting in better performance and faster training [12].

Conditioning. Shown on the left part of the figure, a GCN-based module learns the encoding carrying the source prominence and centrality from the infection state input Y , which will be elaborated on in the next section.

Decoder. Z_d and encoded condition h_{out} are decoded through a GCN-based module, resulting in the estimation for the uncorrupted sample X_0 (i.e. X): $\tilde{X}_0 = \text{GNN}(Z_d, h_{\text{out}})$.

4.3.2 GNN-parameterized Label Propagation(GNN-LP). Our conditioning module takes the observed infection states as input, which contain crucial coupling information between spreading dynamics and network topology. To effectively extract this information, we introduce the GNN-LP module. We first employ label propagation [36], which first marks infected nodes as 1 and uninfected

Algorithm 1 Pretraining on synthetic data and few-shot learning on real data.

Require: Graph $\mathcal{G} = (V, E)$, synthetic data generators \mathcal{M} s (e.g. SIS, IC), few-shot samples \mathcal{D}_{real}

- 1: // Phase 1: Learn from Diverse Propagation Patterns
- 2: Generate synthetic (X^*, Y) pairs using \mathcal{M} s on \mathcal{G}
- 3: Train diffusion model using Eq.9 with (X^*, Y, \mathcal{G})
- 4: // Phase 2: Adapt to Real Scenarios
- 5: **for** each (X^*, Y) in \mathcal{D}_{real} **do**
- 6: Fine-tune model using Eq.9 with (X^*, Y, \mathcal{G})
- 7: **end for**
- 8: // Inference on Real New Cases
- 9: **Input:** Observation Y_{new} on \mathcal{G}
- 10: Compute structure-based prior X_{est} from Y_{new} using Eq.1
- 11: Sample initial noise $X_n \sim \mathcal{N}(0, I)$
- 12: **return** $\hat{X} \leftarrow \text{denoise}(X_n)$ with $(X_{est}, Y_{new}, \mathcal{G})$

as -1 in the observation state Y , resulting in Y^* . The propagation follows the update rule: the label of a node in the next step is a combination of its original label and the sum of normalized labels from its neighbors. We can rewrite this iteration as:

$$\mathcal{Z}_i^{t+1} = \hat{\alpha} Y_i^* + \sigma\left(\sum_{j:j \in \mathcal{N}(I)} \phi(\mathcal{Z}_j^t, S_{ij})\right), \quad (10)$$

where we add non-linear transformations $h(\cdot)$ and $\sigma(\cdot)$ to enhance the expressiveness of the propagation process. During propagation, node labels evolve with the graph structure, reflecting node importance and the relationship between propagation patterns and network structure. We enhance this with GNNs, which learn topology-specific message passing rules to capture how local structures influence propagation. This method’s formulation matches the general GNN [5] framework and is parameterized by a GCN [17] combined with a residual block:

$$\begin{aligned} g(h^{(l)}) &= \sigma(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} \cdot h^{(l)} \cdot w), \\ h^{(0)} &= Y^* U^T, \quad h^{(l+1)} = h^{(l)} + g(h^{(l)}). \end{aligned} \quad (11)$$

Among them, $U \in \mathbb{R}^{C \times 1}$ is the linear transformation, σ is the activation operator PReLU, $h^{(l)}$ stands for the output hidden state of the l -th layer of the GCN, $\tilde{A} = A + I$ is the adjacency matrix with self-loops, and \tilde{D} is the degree matrix of \tilde{A} . The final layer’s output $h^{(l_f)}$ is projected back to dimension 1 and multiplied by the graph’s Laplacian matrix L , i.e. $h_{out} := L \cdot h^{(l_f)}$, which highlights the prominent nodes with higher propagation propagated labels among their neighbors. Through this integration, the module transforms binary infection states Y into continuous representations h_{out} that simultaneously encode both the spreading process and network structure information, enabling effective learning of their complex interactions even with limited training samples.

4.4 Pretrain Using Simulation Data

The proposed structure-prior informed diffusion framework effectively extracts topology-aware features that remain stable across different propagation dynamics. This enables pretraining on synthetic data from established models (IC, LT, SIS) with efficient few-shot adaptation to real scenarios, addressing data-limited source

localization. We illustrate this process in Algorithm 1 and evaluate SIDL’s simulation-to-reality transfer in Section 5.3 by testing two approaches: one using pretrained models with few-shot learning, and another trained exclusively on real-world data.

5 Experiments

5.1 Experiment Settings

Datasets. Datasets. Following Ling et al. [20] and Huang et al. [11], we evaluate SIDL on four real-world propagation datasets containing actual cascades: *Digg*, *Twitter*, *Android*, and *Christianity*. For each cascade, we define source nodes as those infected within the first 10% of the propagation time and use the network’s infection status at 30% of the time as the observation input. Further dataset details are in Appendix A.

Implementation Details. We use a 6:1:1 train/validation/test split for all datasets. The diffusion framework employs $T = 500$ timesteps with linear noise scheduling. The denoising network uses a 2-layer GCN for LPSI estimation X_{est} conditioning, while the GNN encoder/decoder comprises 3-layer GCNs (hidden dimension 128) and the residual GNN uses 2 layers (hidden dimension 8). We use Adam optimizer with learning rate searched from 0.01, 0.005, 0.001, linear decay scheduling, and maximum 500 training epochs.

Baselines. Following previous works [20, 40], we selected two representative heuristic methods, i.e., Netsleuth [23] and LPSI [36], and deep learning methods, i.e., GCNSI [4], TGASI [9] and recent generative deep learning methods SLVAE [20], DDMSL [40] that capture source localization uncertainty. These baselines are all state-of-the-art (SOTA) multi-source localization methods in their domains. Please refer to the Appendix C for specific information on baselines and our method.

Metrics. Following [33], we use three metrics: 1) F1-score (F1): The harmonic mean of precision and recall; 2) Recall (RE): The proportion of correctly identified positive cases (source nodes); and 3) Precision (PR): The proportion of true positives among predicted positive cases. We omit accuracy because it is a misleading metric in highly imbalanced datasets, as it fails to capture the model’s ability to identify the rare positive class.

5.2 Overall Performance on Real-world Datasets

To evaluate the real-world performance of our proposed method against baselines, we conduct direct training and testing on four real-world datasets for all methods. The experimental results are presented in Table 1. Our proposed SIDL outperforms all baselines across nearly all metrics on all datasets. Specifically, SIDL achieves F1 scores that exceed the second-best baseline by 7.5%, 11.0%, 11.8%, and 13.3% on *Digg*, *Twitter*, *Android*, and *Christianity*, respectively, which demonstrates SIDL’s superior ability to predict source nodes despite their sparsity. Our model also excels in recall, with significant gains in *Android* (48.3%) and *Christianity* (12.8%), demonstrating strong ability in identifying true source nodes.

Further, we have the following findings. **First**, compared to purely data-driven generative methods like DDMSL and SLVAE, SIDL demonstrates superior performance, validating the effectiveness of incorporating structural prior information. **Second**, compared to rule-based and learning-based methods, generative

Table 1: Performance evaluation without pretraining under the real-world propagation. Best and second best (F1, RE, PR) are highlighted with bold and underlines respectively.

Datasets	Digg			Twitter			Android			Christianity		
Methods	F1	RE	PR									
Netsleuth	0.006	0.003	0.000	0.160	0.181	0.143	0.142	0.105	0.219	0.128	0.099	0.181
LPSI	<u>0.544</u>	0.516	0.575	0.487	0.495	0.479	0.348	0.517	0.268	0.221	0.282	0.198
GCNSI	0.458	0.411	0.517	0.374	0.352	0.399	0.383	0.474	0.321	0.343	0.321	0.370
TGASI	0.472	0.406	0.564	0.362	0.327	0.405	0.388	0.462	0.335	0.377	0.339	<u>0.423</u>
SLVAE	0.479	0.565	0.416	0.353	0.424	0.302	<u>0.467</u>	<u>0.588</u>	0.387	<u>0.458</u>	<u>0.662</u>	0.351
DDMSL	0.517	<u>0.592</u>	0.459	<u>0.492</u>	<u>0.504</u>	<u>0.481</u>	0.448	0.540	<u>0.432</u>	0.417	0.481	0.368
SIDSL(Ours)	0.585	0.605	<u>0.566</u>	0.546	0.516	0.580	0.522	0.702	0.439	0.519	0.747	0.436
Δ	+7.5%	+2.2%	-1.6%	+11.0%	+2.4%	+20.6%	+11.8%	+48.3%	+1.6%	+13.3%	+12.8%	+3.1%

Table 2: Few-shot learning performance evaluation under real-world propagation, comparing pretraining (P) and no pretraining (NP). Pretraining utilizes simulation data (IC, LT); few-shot learning uses limited real-world data (10%). Rule-based methods are omitted due to their lack of pretraining support. Best and second-best results are bold and underlined, respectively.

Dataset		Digg			Twitter			Android			Christianity		
Methods	Config	F1	RE	PR									
GCNSI	P	0.382	0.398	0.367	0.291	0.251	0.346	0.194	0.145	<u>0.293</u>	0.159	0.174	0.146
	NP	0.162	0.201	0.136	0.003	0.015	0.002	<u>0.012</u>	0.029	<u>0.008</u>	<u>0.176</u>	<u>0.179</u>	<u>0.178</u>
TGASI	P	<u>0.406</u>	0.407	<u>0.405</u>	<u>0.292</u>	0.249	<u>0.353</u>	<u>0.212</u>	<u>0.184</u>	0.251	<u>0.206</u>	0.193	<u>0.221</u>
	NP	<u>0.297</u>	<u>0.303</u>	<u>0.291</u>	0.003	0.015	0.001	0.012	<u>0.062</u>	0.006	0.036	0.038	0.035
SLVAE	P	0.315	0.441	0.246	0.195	0.325	0.139	0.036	0.132	0.021	0.171	<u>0.215</u>	0.142
	NP	0.103	0.188	0.071	0.030	0.042	0.023	0.009	0.010	0.008	0.008	0.016	0.005
DDMSL	P	0.348	<u>0.504</u>	0.266	0.250	<u>0.359</u>	0.192	0.109	0.114	0.104	0.165	0.145	0.191
	NP	0.005	0.003	0.015	<u>0.101</u>	<u>0.095</u>	<u>0.108</u>	0.010	0.015	<u>0.009</u>	0.009	0.005	0.045
SIDSL(Ours)	P	0.571	0.623	0.527	0.518	0.503	0.547	0.547	0.567	0.529	0.483	1.000	0.319
	NP	0.384	0.409	0.362	0.120	0.118	0.143	0.018	0.253	0.015	0.432	0.757	0.309
Δ	P	+41%	+24%	+30%	+77%	+40%	+55%	+158%	+208%	+81%	+134%	+365%	+44%
	NP	+29%	+35%	+24%	+19%	+24%	+32%	+50%	+308%	+67%	+145%	+323%	+74%

methods (SIDSL, SLVAE, DDMSL) show stronger performance, highlighting the advantages of effective source distribution modeling. LPSI and other deep learning-based methods perform comparably well, successfully capturing structural characteristics. Netsleuth demonstrates the weakest performance due to limited generalization beyond SI propagation patterns [13, 23]. **Finally**, SIDSL shows the most substantial improvement on the Christianity dataset (13.3% in F1), as a smaller network scale enhances topological feature capture, amplifying structural prior benefits.

5.3 Performance in Low Data Regime

To assess deep learning methods' use of simulation data pretraining for few-shot/zero-shot real-world source localization, we generate pretraining data via standard IC and LT models (1:1 ratio) across four networks. These models capture social network dynamics like real-world interaction randomness and topology-based spread. For few-shot learning (P), we used a shared, random 10% of training data for all methods. We compare performance against models trained on equivalent data volumes without pretraining (NP). For zero-shot learning, models are tested directly after pretraining.

Few-shot analysis. The result of few-shot learning is shown in Table 2. Rule-based methods (Netsleuth and LPSI) are omitted as they are training-free. SIDSL demonstrates superior performance compared to all baseline methods across all datasets, both with and

without pretraining, achieving improvements of 24-365% across all metrics. We observe two key findings. **First**, SIDSL demonstrates enhanced performance advantages when leveraging pretraining, with 8 out of 12 cases showing larger gains with P compared to NP across different datasets and metrics. This pattern of larger improvements with pretraining demonstrates SIDSL's superior ability to effectively leverage information from simulated data to real-world scenarios in limited-data settings. **Second**, baseline methods exhibit varying capabilities in utilizing pretraining. While TGASI shows consistent improvements with pretraining, generative approaches (SLVAE, DDMSL) demonstrate limited gains (e.g., F1 increases from 0.009 to 0.036 for SLVAE, 0.010 to 0.109 for DDMSL on Android). This highlights generative models' sensitivity to distribution shifts, whereas our method achieves robust improvements through effective incorporation of pattern-invariant structural priors.

Zero-shot analysis. The results of zero-shot learning are shown in Table 3. SIDSL significantly outperforms all baseline methods across all datasets, achieving improvements up to 364% in F1 scores. SIDSL achieves particularly strong performance on Android and Christianity datasets (>0.9 recall with reasonable precision). Baseline methods show limited zero-shot capability (F1 < 0.3). Zero-shot SIDSL outperforms rule-based methods (Table 1) in most metrics. These findings demonstrate SIDSL's superior simulation-to-real transfer without fine-tuning.

Table 3: Zero-shot performance evaluation with pretraining under the real-world propagation. Results show performance without finetuning when using simulation data(IC and LT) for pretraining. The rule-based methods are omitted as they do not support pretraining. Best and second best are highlighted with bold and underline respectively.

Dataset	Digg			Twitter			Android			Christianity		
Methods	F1	RE	PR									
GCNSI	0.195	0.273	0.152	<u>0.203</u>	0.162	<u>0.271</u>	0.103	0.079	<u>0.148</u>	0.030	0.060	0.020
TGASI	0.189	0.281	0.142	0.199	0.172	0.236	<u>0.109</u>	<u>0.107</u>	0.110	0.059	0.067	0.052
SLVAE	0.278	0.236	<u>0.337</u>	0.172	0.205	0.148	0.004	0.025	0.002	0.080	<u>0.140</u>	0.056
DDMSL	<u>0.291</u>	<u>0.339</u>	0.264	0.203	<u>0.272</u>	0.162	0.019	0.013	0.038	<u>0.133</u>	0.127	<u>0.147</u>
SIDSL(Ours)	0.539	0.629	0.472	0.509	0.569	0.460	0.506	0.914	0.355	0.465	1.000	0.305
Δ	+85%	+86%	+40%	+151%	+109%	+70%	+364%	+754%	+140%	+250%	+614%	+108%

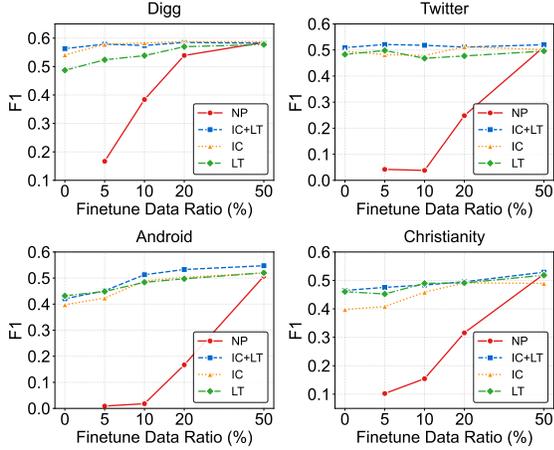


Figure 4: Few-shot performance of SIDSL with different pre-training dataset compositions (IC+LT, IC, LT), then finetuned on real-world data. "NP" denotes "non-pretrained".

In few-shot and zero-shot settings, pretrained SIDSL achieves higher recall than precision gains, effectively identifying true sources but with occasional false positives, especially in Android and Christianity datasets. This suggests strong propagation modeling despite some errors. For misinformation tracing, pretrained SIDSL provides superior source identification. Accuracy can be enhanced using precision-based classifiers with boosting to filter false positives. Without pretraining, SIDSL outperforms other methods on both metrics, proving effective in low-data scenarios.

5.4 Analysis of Pretraining on Simulation Data

To investigate how pretraining on simulation data affects the model's few-shot/zero-shot performance under different data volumes, we evaluate SIDSL's F1 scores across the pretraining dataset with different propagation model combinations (IC, LT, and IC+LT with a ratio of 1:1) and compare them with non-pretrained (NP) SIDSL. We then test them under different ratios of the finetune dataset of real-world propagation data. The results of four datasets are shown in Figure 4.

Finetune data volume analysis. Across all four datasets, models with pretraining consistently outperformed non-pretrained (NP) models in low-data scenarios. Under the IC+LT combination, the

Table 4: Ablation study. The relative performance change of each ablation against SIDSL is reported.

Dataset		Digg			Christianity		
Data	Ablation	F1	RE	PR	F1	RE	PR
100%	SIDSL	100%	100%	100%	100%	100%	100%
	SIDSL w/o SBD	-34%	-47%	-16%	-19%	-49%	+2%
	SIDSL w/o GNN-LP	-7%	-9%	-5%	-4%	-6%	-5%
10%	SIDSL	100%	100%	100%	100%	100%	100%
	SIDSL w/o SBD	-46%	-54%	-36%	-35%	-55%	-24%
	SIDSL w/o GNN-LP	-17%	-16%	-17%	-21%	-34%	-17%

pretrained models achieved optimal F1 scores on Digg and Twitter using just 5% of training data, while NP models required 50%. Similarly, on Android and Christianity datasets, pretrained models reached near-peak performance (within 0.033 and 0.018 F1) with only 20% of the data, compared to NP models, which needed 50%. These findings suggest that aligning pretraining patterns with specific propagation characteristics benefits model performance.

Pattern analysis. On the Digg dataset, SIDSL pretrained with IC+LT and IC configurations outperform those using LT, indicating that IC patterns better align with Digg's propagation characteristics, as also shown in [42]. For Android and Christianity Q&A datasets, IC+LT and LT configurations performed better in few-shot scenarios. This is likely because user participation in these platforms' information cascades (e.g., comments) is driven by cumulative influence from multiple sources (LT) rather than single independent triggers (IC). On Twitter, the IC+LT configuration performs the best, demonstrating that pattern diversity enhances pretraining effectiveness on this dataset.

5.5 Ablation Study

We conduct ablation studies to investigate the significance of key components in SIDSL. For the first ablation model, SIDSL w/o SBD (structure-prior biased diffusion), instead of computing estimation through Equation 1, we directly substitute it with the node having the "highest closeness centrality" in the infection subgraph. For the second ablation, SIDSL w/o GNN-LP, we replace the observation state encoding with a standalone GNN while maintaining an equivalent parameter count. Table 4 presents the ablation results on Digg and Christianity datasets under both full (100%) and limited (10%) training data conditions. Removing either component leads to decreased overall performance under both conditions. There are two more findings: **First**, removing SBD leads to more severe performance degradation compared to removing GNN-LP, with the

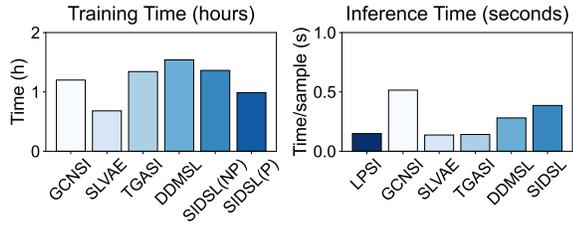


Figure 5: Training and inference time (per sample) comparison on Digg dataset. "NP" denotes the "non-pretrained" version. "P" denotes using the scheme of "pretraining+finetuning", and the overall training time is reported.

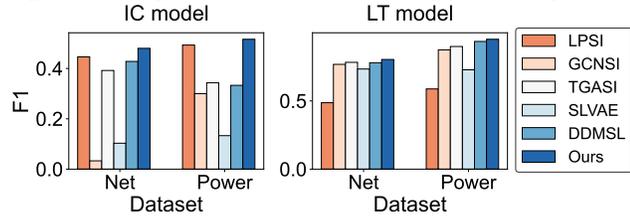


Figure 6: Performance under IC and LT propagation patterns.

impact most pronounced in recall metrics. The substantial decline in recall ($TP/(TP+FN)$) indicates reduced true positive predictions. The decline in recall exceeds that of precision ($TP/(TP+FP)$), implying fewer positive predictions overall ($TP+FP$). This demonstrates SBD’s role in maintaining predictive balance by ensuring adequate positive predictions. **Second**, the impact of removing GNN-LP shows greater impact under limited data (F1 drops: -17% to -21% at 10% vs. -4% to -7% at full data), demonstrating GNN-LP’s effectiveness in low-resource scenarios.

5.6 Time Cost Analysis

We present a detailed comparison of the computational cost of our proposed model against baselines in the Digg dataset, which has one of the largest networks for better comparison. Results are shown in Figure 5. While our model’s original training duration is the second highest, our model requires less time to pretrain on the synthetic data and finetune than initially training on the real data (-25%), which makes the training time the second lowest.

In inference time, our model is the second-highest because of the iterative nature of the DDPM-based denoising process. We opt for DDPM as our foundation due to its classical design and proven effectiveness. Note that SIDS(L)’s architecture is fully compatible with more computationally efficient diffusion variants, such as DDIM [29], which could substantially reduce the current computational overhead. This flexibility, combined with our model’s transfer capabilities, makes SIDS(L) particularly resource-efficient in practical deployments.

5.7 Results on Synthetic Data

It is a common practice in previous works [20, 33, 40] to evaluate methods’ performance on synthetic data with established propagation patterns. Following [20, 33], we conducted experiments on two commonly evaluated networks: *Net Science (Net)* and *Power Grid(Power)*. Using both IC and LT models, we simulated 100 steps until convergence to generate two sets of synthetic data. Strong

baseline models were trained and tested on these datasets, with results shown in Figure 6. Our method outperforms the baselines, achieving F1 score improvements of 7.6% and 4.7% under the IC model, and 2.7% and 1.7% under the LT model on the two datasets, respectively. These results demonstrate our method’s ability to accurately identify sources with established propagation patterns.

6 Conclusion and Future Work

We introduced SIDS(L), a structure-aware diffusion model that significantly enhances source localization in sparse graphs by integrating structural priors. Our framework shows a notable 7.5-13.3% F1 score improvement over leading methods, proving robust in few/zero-shot scenarios with high performance using just 10% of training data. This highlights its strong generalization and value in data-scarce environments. However, the iterative diffusion process introduces computational overhead, potentially limiting real-time applications. While not suited for million-node networks, its principles can be adapted for hierarchical frameworks for large-scale graph analysis. Future work will focus on improving computational efficiency and extending the framework to more source types. These limitations also suggest avenues for research to enhance scalability and robustness.

Appendix

Table 5: Overview of networks. CC: clustering coefficient.

Dataset	Nodes	Edges	Mean Degree	CC
Digg	14511	194405	13.39	0.1353
Twitter	12619	309621	24.52	0.2962
Android	9958	42915	8.62	0.4121
Christianity	2897	30044	20.74	0.6027
Jazz	198	2742	13.84	0.6174
Net Science	1589	2742	1.72	0.6377
Power Grid	4941	6594	1.33	0.0801

A Dataset Description

The study utilizes both real-world and synthetically generated propagation data. This section details the datasets employed in our study, with network statistics summarized in Table 5.

The cascades in real-world propagation datasets represent information diffusion. These include Digg[24], with cascades formed by votes on stories; Twitter[41], where each tweet’s spread is a cascade; and the Stack Exchange communities Android and Christianity [11], where cascades are chronologically ordered posts with identical tags. The top 10% of nodes ranked by influence time were designated as diffusion sources for each cascade, with the top 30% considered as observed influenced nodes.

Synthetic propagation datasets were generated using SIR, IC, and LT models on three real-world networks: Jazz[24], Net Science[24], and Power Grid [37]. The SIR model used constant infection and recovery rates. For IC and LT models, edge propagation probabilities were inversely proportional to the target node’s degree. Simulations ran for 100 steps or until convergence. The Jazz network represents musician collaborations; 5% of nodes were randomly selected as sources. The Net Science network is a coauthorship graph of network scientists, and the Power Grid network maps the Western US

Table 6: Comparison of localization methods.Obs.: Observation input (single/multiple snapshots); PI: Prior-knowledge informed; Patterns: Applicable diffusion patterns.

Category	Method	Indeterminacy	Patterns	Obs.	PI
Rule-based	NetSleuth [23]	✗	SI	single	-
	OJC [44]	✗	SI, SIR	single	-
	LPSI [36]	✗	any	single	-
Data-driven	GCNSI [4]	✗	any	single	✓
	IVGD [33]	✗	IC	single	✗
	SLVAE [20]	✓	any	single	✗
	SLDiff [11]	✗	any	multiple	✗
	TGASI [9]	✗	any	multiple	✗
	DDMSL [40]	✓	any	multiple	✓
	PGSL [39]	✓	any	single	✗
	GINSI [2]	✗	any	single	✗
	NFSL [8]	✗	any	single	✓
	CRSL [10]	✗	any	single	✓
SIDSL (Ours)	✓	any	single	✓	

electrical grid topology; 0.5% of nodes were randomly selected as sources in both datasets.

B Comparison of Multiple Source Localization Methods

Table 6 compares mainstream source localization methods on functionality, requirements, and applications. Our method is more functional, versatile, requires less input, and addresses existing limitations, offering significant practical value. Unlike other diffusion-based methods (DDMSL, TGASI), ours doesn't require propagation data or model parameters pre-localization. PGSL, using a flow-model, is surpassed by our diffusion model's superior distribution modeling. GINSI addresses incomplete data (not our focus), otherwise reducing to a GAT-based baseline like GCNSI. Cross-platform works [19, 35] are orthogonal and not discussed.

C Baselines

We compare SIDSL against state-of-the-art baselines that locate sources from propagation snapshots without requiring knowledge of the underlying propagation pattern.

NetSleuth [23] employs a Minimum Description Length approach for multiple sources, specifically within the Susceptible-Infected (SI) model. **LPSI** [36] detects multiple sources using source prominence and label propagation; its parameter α was tuned from {0.1, 0.3, 0.5, 0.7, 0.9} for each dataset. **GCNSI** [4] utilizes Graph Convolutional Networks for multiple rumor source identification, with settings as per the original paper. **SLVAE** [20] is a probabilistic Variational Autoencoder for source localization that quantifies uncertainty; we used the original implementation and tuned the learning rate (0.001-0.05). **TGASI** [9] offers a sequence-to-sequence framework (GNN encoder, GRU decoder with temporal attention) for multiple rumor source detection, considering heterogeneous user behavior and time-variance, designed for transferability with a unique loss function. **DDMSL** [40] proposes a probabilistic model for source localization and diffusion path reconstruction via a discrete denoising-diffusion model with a reversible residual network.

D Experiments and Implementation Details

We partitioned datasets using a 6:1:1 train/validation/test split. SIDSL employed $T=500$ diffusion timesteps with a linear noise schedule, without classifier-free sampling. Grid search was conducted over encoder/decoder layers {2, 4, 6} and hidden dimensions {32, 64, 128, 256}. After comparing GNN variants (GCN, GIN, GAT, MLP), we selected a 2-layer GCN for LPSI estimation (X_{est}) as the denoising network. The encoder/decoder architecture consisted of 3-layer GCNs with 128-dimensional hidden layers, while the conditioning module employed a 2-layer GCN with 8-dimensional hidden layers. Models were trained using Adam optimizer with learning rates selected from {0.01, 0.005, 0.001} and linear decay, running for 500 epochs (300 for few-shot scenarios) on an RTX 2080 Ti.

E Analyzing Source Centrality in Empirical Data

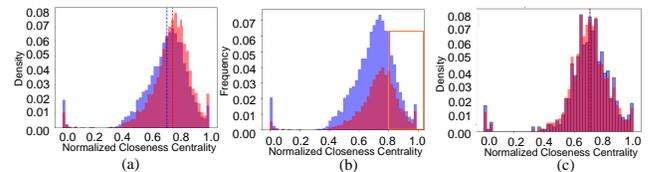


Figure 7: Normalized closeness centrality (CC) on Digg. (a)(b) CC distributions for infected (blue) vs. source (red) nodes: (a) Density plot, with mean CC (dashed lines). (b) Frequency plot, orange box highlights $CC > 0.8$. (c) CC probability density function for ground truth (blue) vs. predicted (red) sources.

The assumption of source centrality is validated by our empirical analysis of the Digg dataset. Figure 7(a) shows the mean normalized closeness centrality (CC) of source nodes is markedly higher than other infected nodes. Furthermore, Figure 7(b) reveals that sources constitute over 63% of nodes with CC scores above 0.8, highlighting their key structural role.

Our proposed SIDSL model is informed by this centrality principle but uses a Graph Neural Network (GNN) for adaptive learning. As shown in Figure 7(c), the CC probability density of SIDSL's predicted sources on Digg (mean 0.7020, std 0.1444) closely matches that of the ground truth sources (mean 0.7044, std 0.1567). This statistical alignment shows our model captures empirical source distributions without introducing bias, demonstrating a robust synthesis of knowledge-guided inference and data-driven learning.

Acknowledgments

This work was supported by the Shenzhen Ubiquitous Data Enabling Key Lab under grant ZDSYS20220527171406015, the Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Endowed Professorship Scheme of the Shenzhen Pengrui Foundation and the National Natural Science Foundation of China under U23B2030.

GenAI Usage Disclosure

The authors used Claude (Anthropic) to assist with specific formatting tasks and language editing in this work. Specifically, Claude

was used to convert pre-existing tabular data into LaTeX table format for presentation purposes and to provide grammar correction and language polishing of author-written text.

No content generation, analysis, interpretation, or creation of new ideas was performed using generative AI tools. All original research, methodology, results, and conclusions presented in this work are entirely the authors' own. The use of Claude was limited to technical formatting assistance and language refinement, similar to using grammar checking tools or word processing software.

References

- [1] Massoud Amin and Phillip F Schewe. 2007. Preventing blackouts. *Scientific American* 296, 5 (2007), 60–67.
- [2] Le Cheng, Peican Zhu, Keke Tang, Chao Gao, and Zhen Wang. 2024. GIN-SD: source detection in graphs with incomplete nodes via positional encoding and attentive fusion. In *Proceedings of the AAAI Conf. on Artificial Intelligence*, Vol. 38, 55–63.
- [3] Jingtao Ding, Chang Liu, Yu Zheng, Yunke Zhang, Zihan Yu, Ruikun Li, Hongyi Chen, Jinghua Piao, Huandong Wang, Jiazhen Liu, et al. 2024. A Comprehensive Survey on Artificial Intelligence for Complex Network: Potential, Methodology and Application. *arXiv preprint arXiv:2402.16887* (2024).
- [4] Ming Dong, Bolong Zheng, Nguyen Quoc Viet Hung, Han Su, and Guohui Li. 2019. Multiple rumor source detection with graph convolutional networks. In *Proceedings of the 28th ACM international Conf. on information and knowledge management*. 569–578.
- [5] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *Int'l Conf. on machine learning*. PMLR, 1263–1272.
- [6] Xizewen Han, Huangjie Zheng, and Mingyuan Zhou. 2022. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems* 35 (2022), 18100–18115.
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.
- [8] Dongpeng Hou, Yuchen Wang, Chao Gao, Xianghua Li, and Zhen Wang. 2024. New Localization Frameworks: User-centric Approaches to Source Localization in Real-world Propagation Scenarios. In *Proceedings of the 33rd ACM Int'l Conf. on Information and Knowledge Management*. 839–848.
- [9] Dongpeng Hou, Zhen Wang, Chao Gao, and Xuelong Li. 2023. Sequential attention source identification based on feature representation. *IJCAI* (2023).
- [10] Dongpeng Hou, Wenfei Wei, Chao Gao, Xianghua Li, and Zhen Wang. [n. d.]. Good Advisor for Source Localization: Using Large Language Model to Guide the Source Inference Process. ([n. d.]).
- [11] Bosong Huang, Weihao Yu, Ruzhong Xie, Jing Xiao, and Jin Huang. 2023. Two-stage denoising diffusion model for source localization in graph inverse problems. In *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*. Springer, 325–340.
- [12] Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. 2023. Normalization techniques in training dnns: Methodology, analysis and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [13] Matt J Keeling and Ken TD Eames. 2005. Networks and epidemic models. *Journal of the royal society interface* 2, 4 (2005), 295–307.
- [14] David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 16th ACM SIGKDD international Conf. on Knowledge discovery and data mining*. 137–146.
- [15] Jeffrey O Kephart and Steve R White. 1993. Measuring and modeling computer virus prevalence. In *Proceedings 1993 IEEE Computer Society Symp. on Research in Security and Privacy*. IEEE, 2–15.
- [16] William Ogilvy Kermack and Anderson G McKendrick. 1927. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* 115, 772 (1927), 700–721.
- [17] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [18] Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. 2010. Finding effectors in social networks. In *Proceedings of the 16th ACM SIGKDD international Conf. on Knowledge discovery and data mining*. 1059–1068.
- [19] Chen Ling, Tanmoy Chowdhury, Jie Ji, Sirui Li, Andreas Züfle, and Liang Zhao. 2024. Source Localization for Cross Network Information Diffusion. In *Proceedings of the 30th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*. 5419–5429.
- [20] Chen Ling, Junji Jiang, Junxiang Wang, and Zhao Liang. 2022. Source localization of graph diffusion via variational autoencoders for graph inverse problems. In *Proceedings of the 28th ACM SIGKDD Conf. on knowledge discovery and data mining*. 1010–1020.
- [21] Wuqiong Luo, Wee Peng Tay, and Mei Leng. 2013. Identifying infection sources and regions in large networks. *IEEE Transactions on Signal Processing* 61, 11 (2013), 2850–2865.
- [22] Zhiqiang Pan, Chen Gao, Fei Cai, Wanyu Chen, Xin Zhang, Honghui Chen, and Yong Li. 2025. On the Cross-Graph Transferability of Dynamic Link Prediction. In *Proceedings of the ACM on Web Conf. 2025*. 4101–4110.
- [23] B Aditya Prakash, Jilles Vreeken, and Christos Faloutsos. 2012. Spotting culprits in epidemics: How many and which ones?. In *2012 IEEE 12th international Conf. on data mining*. IEEE, 11–20.
- [24] Ryan Rossi and Nesreen Ahmed. 2015. The network data repository with interactive graph analytics and visualization. In *Proceedings of the AAAI Conf. on artificial intelligence*, Vol. 29.
- [25] Xiaolei Ru, Jack Murdoch Moore, Xin-Ya Zhang, Yeting Zeng, and Gang Yan. 2023. Inferring patient zero on temporal networks via graph neural networks. In *Proceedings of the AAAI Conf. on Artificial Intelligence*, Vol. 37, 9632–9640.
- [26] Devavrat Shah and Tauhid Zaman. 2011. Rumors in a network: Who's the culprit? *IEEE Transactions on information theory* 57, 8 (2011), 5163–5181.
- [27] Devavrat Shah and Tauhid Zaman. 2012. Rumor centrality: a universal source detector. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international Conf. on Measurement and Modeling of Computer Systems*. 199–210.
- [28] Sushila Shelke and Vahida Attar. 2019. Source detection of rumor in social network—a review. *Online Social Networks and Media* 9 (2019), 30–42.
- [29] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).
- [30] Hongyuan Su, Yu Zheng, Jingtao Ding, Depeng Jin, and Yong Li. 2024. Rumor mitigation in social media platforms with deep reinforcement learning. In *Companion Proceedings of the ACM Web Conf. 2024*. 814–817.
- [31] Yinzhou Tang, Huandong Wang, and Yong Li. 2023. Enhancing Spatial Spread Prediction of Infectious Diseases through Integrating Multi-scale Human Mobility Dynamics. In *Proceedings of the 31st ACM Int'l Conf. on Advances in Geographic Information Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) (SIGSPATIAL '23). ACM, New York, NY, USA, Article 43, 12 pages. doi:10.1145/3589132.3625586
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [33] Junxiang Wang, Junji Jiang, and Liang Zhao. 2022. An invertible graph diffusion neural network for source localization. In *Proceedings of the ACM Web Conf. 2022*. 1058–1069.
- [34] Junxiang Wang and Liang Zhao. 2018. Multi-instance domain adaptation for vaccine adverse event detection. In *Proceedings of the 2018 World Wide Web Conf.* 97–106.
- [35] Zhen Wang, Dongpeng Hou, Shu Yin, Chao Gao, and Xianghua Li. 2024. Joint Source Localization in Different Platforms via Implicit Propagation Characteristics of Similar Topics. In *Proceedings of the Thirty-Third Int'l Joint Conf. on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). Int'l Joint Conf.s on Artificial Intelligence Organization, 2424–2432. doi:10.24963/ijcai.2024/268 Main Track.
- [36] Zheng Wang, Chaokun Wang, Jisheng Pei, and Xiaojun Ye. 2017. Multiple source detection without knowing the underlying propagation model. In *Proceedings of the AAAI Conf. on Artificial Intelligence*, Vol. 31.
- [37] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world' networks. *nature* 393, 6684 (1998), 440–442.
- [38] En Xu, Can Rong, Jingtao Ding, and Yong Li. 2025. A Diffusive Data Augmentation Framework for Reconstruction of Complex Network Evolutionary History. *arXiv preprint arXiv:2501.06485* (2025).
- [39] Xovee Xu, Tangjiang Qian, Zhe Xiao, Ni Zhang, Jin Wu, and Fan Zhou. 2024. PGSL: A probabilistic graph diffusion model for source localization. *Expert Systems with Applications* 238 (2024), 122028.
- [40] Xin Yan, Hui Fang, and Qiang He. 2024. Diffusion model for graph inverse problems: Towards effective source localization on complex networks. *Advances in Neural Information Processing Systems* 36 (2024).
- [41] Cheng Yang, Hao Wang, Jian Tang, Chuan Shi, Maosong Sun, Ganqu Cui, and Zhiyuan Liu. 2021. Full-scale information diffusion prediction with reinforced recurrent networks. *IEEE Transactions on Neural Networks and Learning Systems* 34, 5 (2021), 2271–2283.
- [42] Shiqi Zhang, Jiachen Sun, Wenqing Lin, Xiaokui Xiao, Yiqian Huang, and Bo Tang. 2024. Information Diffusion Meets Invitation Mechanism. In *Companion Proceedings of the ACM on Web Conf. 2024*. 383–392.
- [43] Liang Zhao. 2021. Event prediction in the big data era: A systematic survey. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–37.
- [44] Kai Zhu, Zhen Chen, and Lei Ying. 2017. Catch'em all: Locating multiple diffusion sources in networks with partial observations. In *Proceedings of the AAAI Conf. on Artificial Intelligence*, Vol. 31.
- [45] Kai Zhu and Lei Ying. 2014. Information source detection in the SIR model: A sample-path-based approach. *IEEE/ACM Transactions on Networking* 24, 1 (2014), 408–421.
- [46] Kai Zhu and Lei Ying. 2014. A robust information source estimator with sparse observations. *Computational Social Networks* 1, 1 (2014), 1–21.