Good Advisor for Source Localization: Using Large Language Model to Guide the Source Inference Process

Dongpeng Hou[1,2], **Wenfei Wei**[3], **Chao Gao**[3], **Xianghua Li**[3] and **Zhen Wang**[1,2,3*]
[1]School of Cybersecurity, Northwestern Polytechnical University
[2]School of Mechanical Engineering, Northwestern Polytechnical University
[3]School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University
w-zhen@nwpu.edu.cn

Abstract

With the rapid development of AI large model technology, large language models (LLMs) provide a new solution for source localization tasks due to the deep linguistic understanding and generation capabilities. However, it is difficult to understand complex propagation patterns and network structures when LLMs are directly applied to source localization, resulting in limited accuracy of source localization. Meanwhile, the high-dimensional embedding of the textual representation introduces significant amounts of redundant features, which also reduces its efficiency in source localization task to some extent. To solve the above problems, this paper proposes a multi-modal fusion framework for rumor source localization, namely Contrastive Rumor Source Localization via LLM (CRSLL), based on the idea of contrastive learning. Specifically, the framework constructs propagation embeddings by comprehensively capturing both propagation dynamics and user profile features, adopts a contrastive learning approach to enhance the representation ability of comment embeddings of rumor cascades by differentiating them from non-rumor cascade comments, filters out invalid features through a differentiable masking strategy, and fuses comment modality embeddings with propagation embeddings through an attention mechanism, so as to better capture the multi-modal data interactions. It is worth mentioning that the framework uses LLM as a good "advisor" to provide a rich deep semantic representation, which improves the accuracy of rumor source localization. The code is available at https://github.com/cgao-comp/CRSLL.

1 Introduction

The wide usage of social media has brought both convenience and potential risks to everyone's lives [Meel and Vishwakarma, 2020]. One key issue that has gained significant attention from the government is the spread of rumors. Various fast-spreading rumors have led to significant economic losses [Depoux *et al.*, 2020]. Therefore, it is crucial to identify the rumor sources to prevent further damage [Jiang *et al.*, 2019].

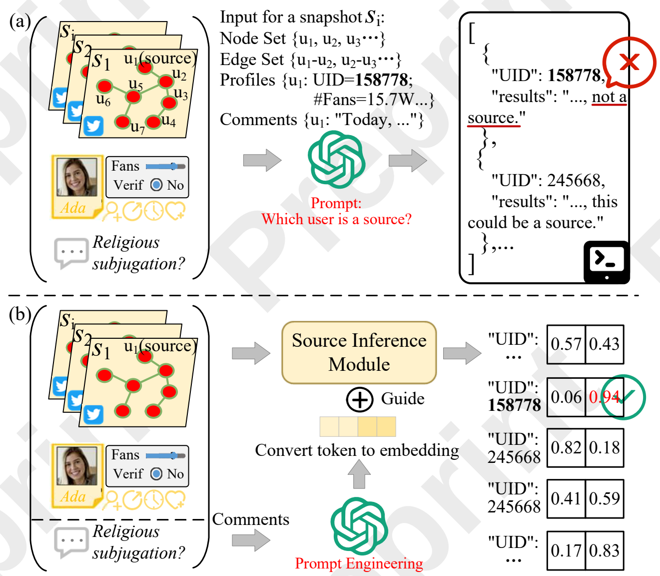

Figure 1: The illustration of the role of large language models (LLMs) in the source localization task. In this case, the LLM (a) fails to output the correct judgment of the real propagation source, but (b) assists the source inference module to judge correctly by providing informative source analysis from the textual comment.

In the field of rumor source localization, the widely employed methods are primarily based on graph theory algorithms [Wang *et al.*, 2017; Hou *et al.*, 2024a] and Graph Neural Network (GNN) algorithms [Dong *et al.*, 2019; Hou *et al.*, 2025], which mainly adopt the snapshot propagation cascades without textual information [Jiang *et al.*, 2016; Jin and Wu, 2021]. In practice, these traditional methods have preliminarily demonstrated their effectiveness. However, with the rapid development of the internet, an increasing amount of textual comments are generated by social media users in the propagation process. The textual comments are not only the direct carrier for users to express their opinions [Yang *et al.*, 2019] and emotions [Zhang *et al.*, 2021], but also potentially serve as a critical channel for the propagation and diffusion of rumors. When confronting the ever-increasing volume and significance of textual comments, the traditional methods cannot fully utilize this textual information. Therefore, how to leverage the potential of textual comments for effectively locating the source becomes a critical question.

In the field of text analysis, pre-trained language models like BERT [Devlin, 2018] are commonly used for their deep linguistic understanding ability. However, small language models like BERT have limitations in domain-specific background knowledge, particularly the lack of external background knowledge or experiences (e.g., historical facts), which led to limited accuracy. As a powerful tool, LLMs like GPT-4 can capture and analyze fine-grained segments (such as historical factual contexts) based on extensive knowledge bases and inference capabilities [Achiam *et al.*, 2023]. Like the application of LLMs in the rumor detection field [Lai *et al.*, 2024], an LLM can be used to address the rumor source localization problem by using snapshots containing textual comments as input. However, as can be seen from Fig. 1(a), the LLM fails to output the correct judgment of the real propagation source if directly using it as a predictor. This suggests that the LLMs are not suitable decoders for rumor source localization tasks, which may not fully comprehend the intricate propagation patterns and network structures embedded in cascade data. In summary, integrating textual comments and snapshots of propagation cascades to enhance the accuracy of rumor source localization is a challenge.

In this paper, we propose a novel framework for rumor source localization, Contrastive Rumor Source Localization via LLM (CRSLL), which uniquely integrates user comments and propagation dynamics for source identification. As illustrated in Fig. 1(b), unlike prior works that do not consider user comments data, we leverage LLMs with prompt engineering to analyze whether each comment indicates the user is a potential rumor source [Hu *et al.*, 2024]. These generated comment analyses replace raw comments as the advisor for downstream localization tasks. To effectively embed the analysis information, we adopt contrastive learning to distinguish rumor-relevant comment patterns from non-rumor comment patterns, and introduce a differentiable Gumbel-Softmax masking mechanism to filter out noise and retain discriminative features. In parallel, for propagation modeling, we construct cascade dynamics and user profile features and use a GNN to learn propagation-aware embeddings. Finally, a cross-modal attention mechanism fuses the comment and propagation signals, enabling the model to identify the sources more accurately and robustly. The major contributions are as follows:

- **LLM Advisor for Source Localization**: Instead of directly using an LLM as predictors for localization, we use LLMs as advisors and analyze whether comments could be potential sources via prompt reasoning. After integrating the comment, dynamics, and profiles mode, LLM provides indirect but more interpretable guidance for the small model and consistently outperforms direct LLM based predictions across multiple datasets.
- **Effective Processing for Comment Analysis**: We consider the quality of LLM-generated analysis, leveraging contrastive learning to enhance the representational ability of analysis embeddings, and implementing a differentiable masking technique to filter out invalid features, thereby improving the robustness.
- **Complete Propagation Datasets in Real-World Scenarios**: We expand propagation datasets containing user profiles, raw comments, and LLM-generated comment analyses. And the ablation study demonstrates the usefulness of these features in datasets.

2 Related Work

In the field of rumor propagation analysis, there are two basic and intertwined issues: how rumor propagates through networks and how to locate the source of rumor. Propagation models describe the mechanisms of rumor propagation and they are capable of providing simulated data for source localization. Conversely, source localization methods aim to reverse this process, leveraging observed propagation patterns to locate the sources of rumor. These two concepts exist in a mutually dependent relationship. Therefore, we conduct a comprehensive review of related work in propagation models and source localization methods.

2.1 Propagation Models

The study of information propagation in social networks began with simple epidemiological models such as the Susceptible-Infected (SI) model [Yang *et al.*, 2020; Paluch *et al.*, 2021; Zang *et al.*, 2015], the Susceptible-Infected-Recovered (SIR)model [Zhu and Ying, 2014; Tang *et al.*, 2018] and the Susceptible-Infected-Susceptible (SIS) model. These models provide a basic framework for understanding information spread and generate datasets for source localization tasks. However, they were primarily based on the assumption of homogeneity among individuals, where all individuals in the propagation models were assumed to have the same features such as infection and recovery rates. This homogeneity assumption does not accurately reflect the complexity and diversity observed in real-world scenarios. To overcome this limitation, some heterogeneous diffusion models such as the Heterogeneous SI (HSI) and Heterogeneous SIR (HSIR) were introduced [Karrer and Newman, 2010; Ellison, 2020]. These models simulate a more realistic information propagation process by considering differences between individuals. Additionally, there are some influence models such as the Independent-Cascade (IC) and Linear Threshold (LT) [Goldenberg *et al.*, 2001; Granovetter, 1978] were introduced, which highlight the dynamics of mutual influence of information propagation. However, it is important to note that while these models are valuable tools for propagation simulating, they aren't based on real data and don't consider textual comments. Therefore, their applicability in real-world scenarios is limited.

2.2 Source Localization Methods

In real-world scenarios, snapshot data, which captures the state of the network at specific points in time, is easily obtainable. Consequently, there is a significant amount of research on snapshot based source localization. Wang et al. proposed the LPSI method. This method employs a label propagation technique, which is based on source prominence, to locate the source [Wang *et al.*, 2017]. The GCNSI method proposes a GCN based model to locate multiple rumor sources without prior knowledge of the underlying propagation model [Dong *et al.*, 2019]. Furthermore, methods like IVGD [Wang *et al.*, 2022], MCGNN [Shu *et al.*, 2021], and SL_VAE [Ling *et al.*, 2022] build dynamic propagation features prior to source inference. Hou et al. utilize an encoder-decoder framework to learn the influence matrix between any two users, which is then employed in the source inference process [Hou *et al.*, 2023]. Furthermore, Huang et al. address the ill-posed problem of the source localization problem by proposing a two-stage optimization framework, the source localization denoising diffusion model (SL-Diff), which quantifies uncertainty in the propagation process to improve detection accuracy [Huang *et al.*, 2023]. Unlike the above methods, our proposed CRSLL method goes beyond static snapshots and innovatively integrates the consideration of textual comments, leveraging the LLM to guide the localization of rumor sources.

3 Preliminaries

3.1 Propagation Cascades

We obtain K number of available experienced historical propagation cascades $\mathcal{C}_k$=$(\mathcal{V}_k, \mathcal{E}_k, \mathcal{F}_k)$ $(1 \leq k \leq K)$ from Twitter or Weibo platforms, where $\mathcal{V}_k$ is the participant user set with UID in a social media platform, $\mathcal{E}_k$ is the set of participant's directed propagation interaction (including comments or retweets from a user to another), and $\mathcal{F}_k$ is the feature set (i.e., user profiles) for each user, including user description, blue verification status, location, registration date, number of posts, fans list, and followings list.

3.2 Historical Relationship Network

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 $\boldsymbol{\mathcal{G}}$=$(\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{F}})$, which is a union graph by combining structural information of different cascades based on the same UIDs. Sincerely, we pick this idea from the field of diffusion inference [Ramezani *et al.*, 2023], where it is widely used as an intuitive yet effective approach when the underlying network is unknown. Specifically, if different cascades share the same UID, it typically suggest that these cascades are not isolated incidents but rather part of an underlying relationship network, driven by shared interests or topics. By uniting these cascades, a complex historical relationship network emerges where users from various cascades engage with each other either directly or indirectly, based on shared interests or topics. Focusing on this distinct identified area, our research is to locate the sources from a new propagation within $\boldsymbol{\mathcal{G}}$.

3.3 Problem Definition

Having constructed the historical relationship network $\boldsymbol{\mathcal{G}}$=$(\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{F}})$in a social platform, as for a new propagation cascade $C = \{V, E\}$ at a timestamp of a concerned topic spreading in the area $\boldsymbol{\mathcal{G}}$, we conveniently observe an available snapshot V, which only includes the UID of the participants. And we denote the original rumor sources set as $R \subset C$. The goal of our method is to predict a source set $\hat{R}$ which can maximize the indicator like $\frac{\hat{R} \cap R}{\hat{R} \cup R}$ based on the historical prior knowledge $\boldsymbol{\mathcal{G}}$ and a new snapshot V.

4 Method

In this part, the source localization framework incorporating LLM prompt engineering, called CRSLL, is proposed. As shown in Fig. 2, CRSLL innovatively includes five main components: propagation embedding construction, contrastive learning for comment embedding, feature selection with differentiable masking, attention fusion across modalities, and weighted binary classification. In detail, it works as follows: First, it takes traditional snapshot data and textual comments data as inputs, converting them into propagation and comment embeddings, respectively. Then, it utilizes contrastive learning to enhance the representational ability of comment embeddings and conducts feature selection through differentiable masking to improve the quality of high-dimensional comment embeddings, which contain redundant information. Lastly, CRSLL integrates the comment and propagation embeddings by attention mechanism and optimizes source inference with weighted binary classification loss to achieve the rumor source localization task.

4.1 Propagation Embedding Construction

For a new observed propagation snapshot V of the new concerned cascade C, historical relationships in $\boldsymbol{\mathcal{G}}$ are used to extract a knowledge based snapshot subgraph $S = \{\hat{V}, \hat{E}\}$ or adjacency A. To perceive the future potential participants from the historical experience in $\boldsymbol{\mathcal{G}}$, first, we extract additional one-hop relationships of V from the focused area $\boldsymbol{\mathcal{G}}$.

$$\hat{V} = \{u \mid u \in \mathcal{N}^{\boldsymbol{\mathcal{G}}}(v), v \in V\} \cup V, \quad (1)$$

$$\hat{E} = \{(v_i, v_j) \mid v_i, v_j \in \hat{V} \text{ and } (v_i, v_j) \in \boldsymbol{\mathcal{E}}\}, \quad (2)$$

where $\mathcal{N}^{\boldsymbol{\mathcal{G}}}(v)$ is the neighbor set of user v in the historical network $\boldsymbol{\mathcal{G}}$. Then, a single snapshot V can be mapped onto $\boldsymbol{\mathcal{G}}$ and is denoted as $S = \{\hat{V}, \hat{E}, \hat{Y}\}$. Here, $\hat{Y}(v_j) = 1$ indicates that a user v_j has participated in the new cascade C of concerned topic, and $\hat{Y}(v_j) = 0$, otherwise. After obtaining a snapshot subgraph S, first, in the encoder phase, the propagation embedding including propagation dynamic features and user profiles is constructed to better solve the user-level-based source localization task.

Some unique propagation dynamic features were designed for each user and then combined with user profiles to differentiate each unique user. These features include seven explicit dynamic indicators (denoted as H_1-H_7), which are constructed to characterize the propagation dynamic features of an individual. Among them, the ratio of participated neighbors and non-participated neighbors of v_j are shown in Eq. (3) and Eq. (4), respectively.

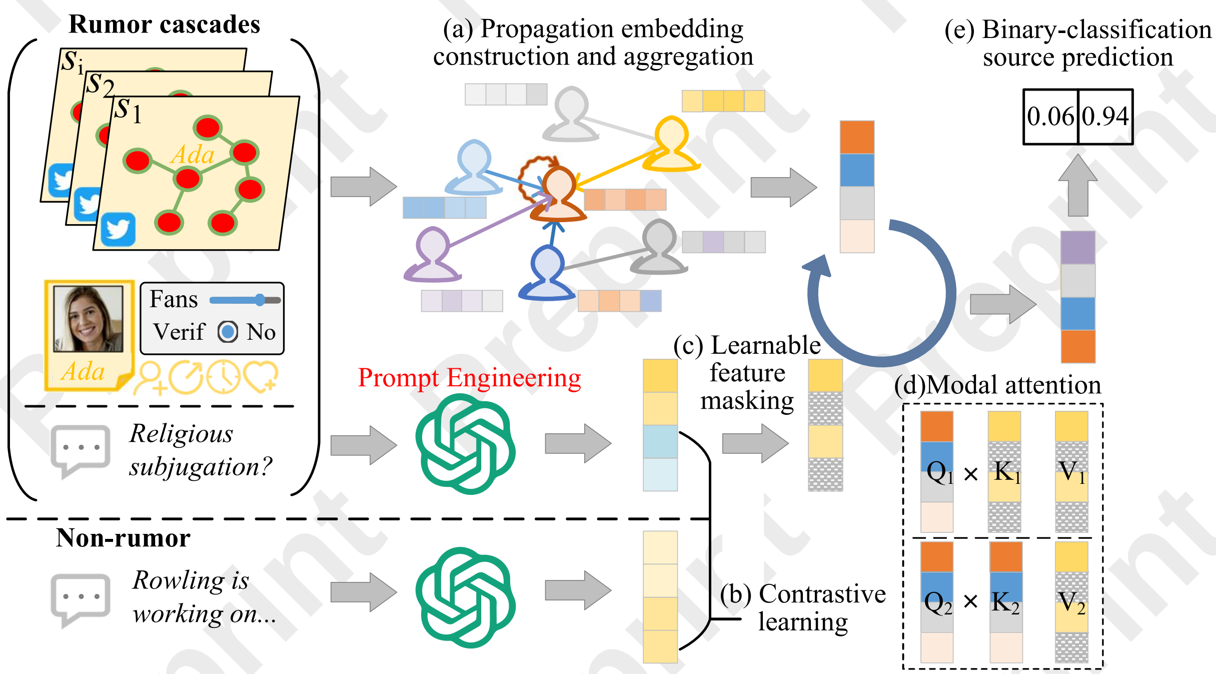

Figure 2: The illustration of CRSLL based on the LLM prompt engineering. (a) Propagation Embedding Construction: The propagation embedding for each user is constructed by dynamically aggregating propagation dynamics and user profile features from their own and their neighbors. (b) Contrastive Learning for Comment Embedding: After getting the source analysis based comment embedding by LLM's prompt engineering, contrastive learning is used to enhance the representational ability of comment embeddings of rumor cascades by differentiating them from non-rumor cascade comments. (c) Feature Selection with Differentiable Masking: Given the high dimensional BERT embeddings with redundancy, a differentiable masking strategy is employed to filter out invalid features in order to enhance the quality of features while preserving the learnable gradients. (d) Attention Fusion Across Modalities: The attention mechanism is applied to combine the comment modality embeddings with the propagation embeddings. (e) Weighted Binary Classification: A weighted binary classification loss is designed to focus on the minority of source users and optimize the source inference process.

$$H_1(v_j) = \frac{\sum_{v_k \in \mathcal{N}^S(v_j)} \hat{Y}(v_k)}{|\mathcal{N}^S(v_j)|}, \tag{3}$$

$$H_2(v_j) = \frac{|\mathcal{N}^S(v_j)| - \sum_{v_k \in \mathcal{N}^S(v_j)} \hat{Y}(v_k)}{|\mathcal{N}^S(v_j)|}. \tag{4}$$

What's more, we also consider the normalized number of participated and non-participated neighbors of v_j, which are shown in Eq. (5) and Eq. (6), respectively.

$$H_3(v_j) = \frac{\sum_{v_k \in \mathcal{N}^S(v_j)} \hat{Y}(v_k)}{\max_{u \in \hat{V}} (\mathcal{N}^S(u))}, \tag{5}$$

$$H_4(v_j) = \frac{|\mathcal{N}^S(v_j)| - \sum_{v_k \in \mathcal{N}^S(v_j)} \hat{Y}(v_k)}{\max_{u \in \hat{V}} (\mathcal{N}^S(u))}. \tag{6}$$

Here, features H_1-H_4 indicate that we are not solely focused on the dynamic ratio of neighbor users. Both the total number of participated and non-participated neighbors emphasize our concern for the precise count of neighbors' states, not just their proportions. For example, considering that a user only has one neighbor and the neighbor participates in the topic, then H_1 is a relatively large feature indicator. However, indicator H_3 of such a user is small. Therefore, both normalized numerical features and proportional features need to be considered. Moreover, the original state $\hat{Y}(v_j)$ of each user in S, whether participated (H_5) or non-participated (H_6), collectively represents the essential property of the individual. Furthermore, we also pay attention to the degree centrality (H_7) [Simmie *et al.*, 2013] which can reflect the celebrity effect in social networks. After these seven propagation dynamic features are obtained, the propagation embedding $H(v_j) \in \mathbb{R}^d$ can be obtained by concatenating the normalized user profile features in $\boldsymbol{\mathcal{F}}(v_j)$.

Since we have the propagation embedding of users in the topology scenarios, an intuitive strategy to aggregate user features is to use the GNN unit. However, in the aggregation process of single-layer GCN, we observe that celebrities with higher degrees, despite having a larger number of interacting neighbors, often contribute with relatively lower feature weights in the aggregation process, impacting both themselves and their neighboring nodes. This tendency highlights an application challenge of the single-layer GCN module in the source localization field, where the influence of highly connected nodes might be diminished in the aggregation process. Therefore, we propose a self-loop attention based GCN to revise the coefficient weight of propagation dynamic features and profiles $H(v_j)$ of each user during the process of information aggregation. In this way, celebrities can weaken the average of feature influence by its neighbors, so as to better reflect the real-world level of influence of their characteristics during the aggregation process. As shown in Eq. (7), we add a learnable diagonal matrix to personalize the element value on the diagonal of the matrix $\Lambda \in \mathbb{R}^{|\hat{V}|*|\hat{V}|}$. The propagation embedding including propagation dynamics and user profiles can be updated as follows:

$$H \leftarrow \sigma\left[(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}} + \Lambda)HW\right]. \tag{7}$$

where W is the learnable weights in the module. $\tilde{A} = A + I$, where I is an identity matrix. $\tilde{D}$ is the corresponding degree matrix of $\tilde{A}$, and σ is an activation function. The diagonal elements of this matrix are controlled by a multi-head attention mechanism. And a single-layer BP neural network $\vec{a} \in \mathbb{R}^{14}$ is applied for each head of the attention mechanism. To make coefficients easily comparable across all users, we normalize the self-loop attention mechanism.

$$\begin{aligned}\phi(v_j) &= \mathrm{softmax}_{v_k \in \hat{V}}\left(e(v_k)\right) = \frac{\exp\left(e(v_j)\right)}{\sum_{v_k \in \hat{V}} \exp\left(e(v_k)\right)} \\ &= \frac{\exp\left(\mathrm{ReLU}\left(\overrightarrow{a}^T\left[{W_A}^T H(v_j)\right]\right)\right)}{\sum_{v_k \in \hat{V}} \exp\left(\mathrm{ReLU}\left(\overrightarrow{a}^T\left[{W_A}^T H(v_k)\right]\right)\right)},\end{aligned} \tag{8}$$

where W_A is the learnable matrix in the attention module, and $e(v_j)$ is the self-loop attention coefficient of v_j. Furthermore, by diagonalizing the normalized self-loop attention coefficients $\phi(v_j)$ for each user, the matrix Λ can be obtained. Λ allows for the automated and dynamic adjustment of influence coefficients for each node, catering to its unique role in the network.

4.2 Contrastive Learning for Comment Embedding

The propagation embedding H of a propagation cascade C is constructed in the above section. Considering the fact that propagation is user-driven, integrating user profile information can enhance the quality of the embedding. However, comments play a crucial role in the dynamics of information spread, acting as indicators of user engagement and sentiment. Therefore, considering the factor of comments during the source inference process can further improve the detection performance. Considering the powerful expert inference capability of LLMs in analyzing external background knowledge or experience (i.e., historical facts) , which is beyond the capabilities of smaller models like BERT, we opt not to directly use pre-trained BERT for embedding conversion. Instead, we refine the comment modality based on LLM prompt engineering.

Prompt Engineering: Source Analysis Generation

System Prompt: This is a propagation cascade in a social network, involving user IDs and comments. You then need to analyze the reasons whether each comment corresponding to the uid is a source user (the first person to initiate the propagation) or not. Your answer must also be in JSON format.
Context Prompt: All UIDs and comments of a propagation cascade with JSON format.

After converting the source analysis from all comments of a cascade C to the embedding H^* based on the pre-trained BERT, a contrastive learning mechanism between rumors and non-rumors is deployed to enhance the embedding quality by minimizing the distance between positive similar pairs (i.e., comments from other rumor cascades) and maximizing the distance between negative dissimilar pairs (i.e., comments from non-rumor cascades). The procedure is demonstrated as follows:

$$\mathcal{L}_{\mathrm{CL}}(H^*) = -\log\left(\frac{\exp(\mathrm{sim}(H^*, \mathbf{S}^*(H^+)))}{\exp(\mathrm{sim}(H^*, \mathbf{S}^*(H^+)))) + \exp(\mathrm{sim}(H^*, \mathbf{S}^*(H^-)))}\right), \tag{9}$$

where H^+ is the positive pair from the same batch but different from H^*, and H^- is the negative pair from the non-rumor cascades, $\mathbf{S}^*$ denotes the random sampling operator, which selects an instance from the set, $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity evaluation measuring the closeness between two vectors. The contrastive revision effectively refines the differentiation of the embeddings between the rumor comment and non-rumor comment, ensuring that the overall representation quantity of H^*.

4.3 Feature Selection with Differentiable Masking

Considering the high-dimensional embedding of the comment modality, i.e., 768 dimensions of H^*, may introduce redundant or highly correlated features. This can lead to noise and inefficiencies in subsequent modal fusion processes, ultimately degrading the quality of the final representation. To mitigate this issue, we implement a Gumbel-Softmax based feature masking process that enables the differentiable masking of invalid features. More precisely, the binary classification-based decision network Θ determines whether each feature should be masked or retained through a linear transformation that produces logits for each feature of the comment embedding H^*:

$$\psi = \Theta(H^* \in \mathbb{R}^{N\times 768}) \in \mathbb{R}^{N\times 768\times 2}. \tag{10}$$

For a differentiable approximation of discrete feature selection, we employ the Gumbel-Softmax technique:

$$\mathbf{\Psi} = \mathrm{GumbelSoftmax}(\psi, \tau, \mathrm{hard} = \mathrm{True}), \tag{11}$$

where τ denotes the temperature parameter controlling the softness of the output, and 'hard=True' ensures a one-hot vector output during the forward pass while preserving differentiability during the backward pass through the use of a straight-through gradient estimator.

The decision network's output dictates whether features are masked or retained:

$$H^* = \begin{cases} H^*[v][f] \odot (\mathbf{\Psi}[v][f][0] \cdot w), & \text{if } \mathbf{\Psi}[v][f][0] = 1 \\ H^*[v][f] \odot \mathbf{\Psi}[v][f][1], & \text{if } \mathbf{\Psi}[v][f][1] = 1 \end{cases}, \tag{12}$$

where $\odot$ represents element-wise multiplication and w is a decay coefficient less than 1. By introducing Gumbel noise into the logits for feature selection and applying the softmax function, we achieve a continuous, differentiable approximation of the feature decision and masking process for H^*.

4.4 Attention Fusion Across Modalities

After obtaining the propagation embedding H based on the self-loop attention mechanism and comment embedding H^* based on contrastive learning and differentiable masking, a cross-attention mechanism is applied for H and H^* to dynamically adjust the weight of each user.

$$H' = \mathrm{softmax}\left(\mathbf{Q_1}(H^*) \cdot \mathbf{K_1}(H)^T / \sqrt{d}\right) \mathbf{V_1}(H), \tag{13}$$

$$H^{*\prime} = \mathrm{softmax}\left(\mathbf{Q_2}(H) \cdot \mathbf{K_2}(H^*)^T / \sqrt{d}\right) \mathbf{V_2}(H^*), \tag{14}$$

where $\mathbf{Q}(H)$ is the query matrices applied to H, $\mathbf{K}(H)$ is the key matrices applied to H, and $\mathbf{V}(H)$is the value matrices applied to H. d is the dimensionality. Furthermore, the two optimized embeddings are concatenated to assemble a more comprehensive representation containing rich propagation characteristics for further binary classification task.

$$\hat{R} = \mathrm{Softmax}\left(\mathrm{BinaryMLP}\left(\ \mathrm{CAT}(H', H^{*\prime})\right)\right). \tag{15}$$

4.5 Weighted Binary Classification

Further, a loss function is required to realize the parameters optimization of CRSLL. Without loss of generality, $\hat{R}[:,0]$ is denoted to be the predicted probability for the non-source classification and $\hat{R}[:,1]$ to be the probability for the source classification. And the loss $\mathcal{L}$ is used to train the complete process of CRSLL based on the $|V|$-nodes accumulated binary classification task.

$$\begin{aligned} \mathcal{L}(R, \hat{R}) = &-(1 - \frac{\sum R}{|\hat{V}|}) R \log(\hat{R}[:,1]) \\ &- \frac{\sum R}{|\hat{V}|}(1 - R)\log(\hat{R}[:,0]). \end{aligned} \tag{16}$$

5 Experiments

5.1 Experimental Setup

We used three datasets collected from two real-world social platforms, Weibo and Twitter, for source localization, namely Weibo [Ma *et al.*, 2017], Twitter15, and Twitter16 [Liu *et al.*, 2015; Ma *et al.*, 2016]. Furthermore, we have crawled user profile information for each user based on the UID in the propagation cascade, achieving user profile alignment on the social platform [Hou *et al.*, 2024b]. And the comments information in the cascades are integrated. The relevant information of the three datasets is shown in Tab. 1. To demonstrate the validity and novelty of the proposed localization methods, we consider TGASI [Hou *et al.*, 2023], IVGD [Wang *et al.*, 2022], SL_VAE [Ling *et al.*, 2022], GCSSI [Dong *et al.*, 2022], MCGNN [Shu *et al.*, 2021], GIN-SD [Cheng *et al.*, 2024b], and HFSD [Cheng *et al.*, 2024a] for comparison. What's more, we also use the common language model, including pre-trained BERT [Devlin *et al.*, 2019], GPT-4o, and GPT-4 [OpenAI, 2022].

Statistic	**Twitter15**	**Twitter16**	**Weibo**
#users	480,987	289,675	2,856,741
#users in $\mathcal{G}$	480,405	289,504	2,856,519
#relations in $\mathcal{G}$	565,948	334,603	3,508,596
#cascades	1490	818	4664
#rumors	372	207	2244
#non-rumors	744	410	2082
#comments	16,428	11,240	61,247

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

And to demonstrate the source prediction performance of all methods rigorously, the widely used standard F1-score [Sokolova *et al.*, 2006] is chosen as the evaluation metric [Wang *et al.*, 2023; Hou *et al.*, 2023].

$$\text{F1-score} = \frac{2 * Precision * Recall}{Precision + \ Recall} \tag{17}$$

where the *Precision* is the ratio of the ground-truth sources in the predictions. The *Recall* is the ratio of successful predictions in the ground-truth sources.

In our experiments, we employ a 10-fold cross-validation strategy to divide the training and test datasets. Further, CRSLL utilizes the training dataset for learning, and then the final result is output by averaging the prediction across each fold in the test dataset. For optimization, the Adam optimizer is used, configured with a learning rate of 0.0005 for all model parameters. In the loss function, both α and β are set to 0.5.

5.2 Overall Experimental Results

The source detection performance based on the real-world dataset is illustrated in Tab. 2. GPT-4 outperforms all other deep learning based comparison methods, exhibiting the best detection performance among all SOTA methods. This underscores the significant potential of LLMs in the field of propagation source localization when many kinds of social context information such as text are available.

Then, compared with the optimal baseline GPT-4, CRSLL exhibits an average improvement of 62.3% in real-world datasets. There are three key reasons for the significant improvement in real-world datasets: (1) The self-loop attention-based mechanism refines the aggregation strategy for each user, improving the representation quality of propagation dynamics and profile. (2) The contrastive learning from non-rumor comments enhances the representation quality of rumor comments, and the learnable feature masking module further removes the redundant features of high-dimensional comment embedding. (3) Different modalities are dynami-

Datasets	*Twitter15*	*Twitter16*	*Weibo*
Vanilla BERT	0.202	0.217	0.188
GIN-SD [Cheng *et al.*, 2024b]	0.575	0.583	0.566
HFSD [Cheng *et al.*, 2024a]	0.489	0.502	0.477
TGASI [Hou *et al.*, 2023]	0.559	0.511	0.485
IVGD [Wang *et al.*, 2022]	0.417	0.366	0.321
SL_VAE [Ling *et al.*, 2022]	0.344	0.352	0.340
GCSSI [Dong *et al.*, 2022]	0.208	0.225	0.265
MCGNN [Shu *et al.*, 2021]	0.226	0.271	0.188
GPT-4o	0.371	0.402	0.364
GPT-4	<u>0.586</u>	<u>0.602</u>	<u>0.573</u>
CRSLL	**0.951**	**0.946**	**0.889**

Table 2: Source identification performance based on the real-world dataset based on the F1-score metric. The bold values represent the best results, while underlined values denote the second-best.

cally adjusted and more complete information is considered for the source inference process.

5.3 Ablation Study for Datasets

To verify the effectiveness of the collected user profiles, comments, and comment analysis for the localization task, the ablation study for these features is conducted on CRSLL and the lower-cost LLM (i.e., GPT-4o). Due to the limited space, we only present the experiment results in the Twitter15 and Twitter16 datasets. We consider the combinations of topology (T), user profiles (U), comments (C), and comment analysis (A). As can be seen from Tab. 3, both CRSLL and GPT-4o are initially conducted using traditional information that solely includes structural topology for source inference. These models progressively incorporate additional propagation features, including user profiles and comments, to validate the performance on the source detection tasks. It can be observed that the lack of user profiles or comment information leads to a decrease in model performance for each method, underscoring the importance of these features for source localization. More importantly, we have discovered that LLMs are highly sensitive to textual information. The localization performance significantly drops (28.3% in Twitter15 and 32.5% in Twitter16) when comment information is lacking. This suggests that LLMs have limited capability in parsing structural topology information, they have a stronger ability for processing and analyzing textual content in source localization tasks.

5.4 Ablation Study for CSRLF

We further study the influence of designed components of CSRLF on the source detection performance to prove their contributions. The critical modules of CSRLF include the self-loop attention mechanism, contrastive learning, learnable feature masking, modal attention mechanism, and weighted binary classification loss. So five variant models of CSRLF are developed as follows.

- CSRLF_S- removes the self-loop attention in Eq. (8).
- CSRLF_C- removes the contrastive learning in Eq. (9).
- CSRLF_M- removes the learnable feature masking in Eqs. (10)-(12).
- CSRLF_A- removes modal attention in Eqs. (13)-(14).
- CSRLF-CE replaces the weighted binary classification loss with a standard cross-entropy loss.

Variants		*Twitter15*	*Twitter16*
GPT-4o	T	0.114 ($\downarrow$ 69.2%)	0.116 ($\downarrow$ 71.1%)
	T + U	0.266 ($\downarrow$ 28.3%)	0.271 ($\downarrow$ 32.5%)
	T + U + C	$\mathbf{0.371}^*$	$\mathbf{0.402}^*$
CRSLL	T	0.587 ($\downarrow$ 38.2%)	0.584 ($\downarrow$ 38.2%)
	T + U	0.913 ($\downarrow$ 3.9%)	0.906 ($\downarrow$ 4.2%)
	T + U + C	0.921 ($\downarrow$ 3.1%)	0.917 ($\downarrow$ 3.0%)
	T + U + A	$\mathbf{0.951}^*$	$\mathbf{0.946}^*$

Table 3: The detection performance of different combinations of available features in the Twitter15 and Twitter16 datasets. T is the topology information, U is the user profiles, C is the user comments, and A is the comment analysis. $*$ represents the optimal experimental settings for a method, which also can be seen in Tab. 2.

Variants	*Twitter15*	*Twitter16*
CSRLF_S-	0.871 ($\downarrow$ 8.4%)	0.882 ($\downarrow$ 6.7%)
CSRLF_C-	0.904 ($\downarrow$ 4.9%)	0.913 ($\downarrow$ 3.4%)
CSRLF_M-	0.917 ($\downarrow$ 3.5%)	0.920 ($\downarrow$ 2.7%)
CSRLF_A-	0.922 ($\downarrow$ 3.0%)	0.917 ($\downarrow$ 3.0%)
CSRLF-CE	0.741 ($\downarrow$ 22.0%)	0.707 ($\downarrow$ 25.2%)
CSRLF	**0.951**	**0.946**

Table 4: The detection performance of variants from CRSLL and GPT-4o in Twitter15 and Twitter16.

Due to the limited space, we only present the experiment results in the Twitter15 and Twitter16 datasets. As can be seen from Tab. 4, it will lead to a performance decrease no matter removing or replacing any critical modules.

6 Conclusion

Comments in real-world propagation cascades contain rich information (such as background, emotions, etc.), which can provide a new perspective for locating the propagation source. However, models such as BERT have limitations in background knowledge when processing text information in social networks, resulting in limited performance. Therefore, we propose a contrastive rumor source localization via LLM, using the LLM to analyze whether a comment is the source of a rumor. On the one hand, we design contrastive learning to enhance the representational ability of comment analysis, and implement a differentiable masking technique to filter out invalid features. On the other hand, we introduce propagation dynamics and user profile features to construct propagation embeddings to jointly determine the propagation source. Experiments demonstrate the effectiveness of comments and user profiles in localization tasks.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. U22B2036, 62271411, 62471403), 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), and the XPLORER PRIZE.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Cheng *et al.*, 2024a] Le Cheng, Peican Zhu, Chao Gao, Zhen Wang, and Xuelong Li. A heuristic framework for sources detection in social networks via graph convolutional networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(11):7002–7014, 2024.

[Cheng *et al.*, 2024b] 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.

[Depoux *et al.*, 2020] Anneliese Depoux, Sam Martin, Emilie Karafillakis, Raman Preet, Annelies Wilder-Smith, and Heidi Larson. The pandemic of social media panic travels faster than the COVID-19 outbreak. *Journal of Travel Medicine*, 27(3):taaa031, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[Devlin, 2018] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[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*, 2022.

[Ellison, 2020] Glenn Ellison. Implications of heterogeneous sir models for analyses of covid-19. Technical report, National Bureau of Economic Research, 2020.

[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.

[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.

[Hu *et al.*, 2024] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113, 2024.

[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. Springer, 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.

[Jiang *et al.*, 2019] Jiaojiao Jiang, Sheng Wen, Bo Liu, Shui Yu, Yang Xiang, and Wanlei Zhou. *Malicious attack propagation and source identification*. Springer, Gewerbestrasse 11, 6330 Cham, Switzerland, 01 2019.

[Jin and Wu, 2021] Rong Jin and Weili Wu. Schemes of propagation models and source estimators for rumor source detection in online social networks: A short survey of a decade of research. *Discrete Mathematics, Algorithms and Applications*, 13(04):2130002, 2021.

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

[Lai *et al.*, 2024] Jianqiao Lai, Xinran Yang, Wenyue Luo, Linjiang Zhou, Langchen Li, Yongqi Wang, and Xiaochuan Shi. Rumorllm: A rumor large language model-based fake-news-detection data-augmentation approach. *Applied Sciences*, 14(8):3532, 2024.

[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. AAAI, 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.

[Meel and Vishwakarma, 2020] Priyanka Meel and Dinesh Kumar Vishwakarma. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153:112986, 2020.

[OpenAI, 2022] OpenAI. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2022. Accessed: 2023-08-13.

[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.

[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.

[Simmie *et al.*, 2013] D. Simmie, M.G. Vigliotti, and C. Hankin. Ranking twitter influence by combining network centrality and influence observables in an evolutionary model. In *2013 International Conference on Signal-Image Technology & Internet-Based Systems*, pages 491–498, 2013.

[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 *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, San Francisco, CA, 2017. PALO ALTO, CA 94303 USA.

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

[Wang *et al.*, 2023] Zhen Wang, Dongpeng Hou, Chao Gao, Xiaoyu Li, and Xuelong Li. Lightweight source localization for large-scale social networks. In *Proceedings of the ACM Web Conference 2023*, pages 286–294, 2023.

[Yang *et al.*, 2019] Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(01), pages 5644–5651, 2019.

[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.

[Zhang *et al.*, 2021] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476, 2021.

[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.