HHAN: Comprehensive Infectious Disease Source Tracing via Heterogeneous Hypergraph Neural Network

Qiang He[1], Yunting Bao[1], Hui Fang[2*], Yuting Lin[1], Hao Sun[1]
[1]Northeastern University, Shenyang, China
[2]Research Institute for Interdisciplinary Sciences and Key Laboratory of Interdisciplinary Research of Computation and Economics, Shanghai University of Finance and Economics, China
heqiangcai@gmail.com, baoyunting706@gmail.com, fang.hui@mail.shufe.edu.cn,
yutinglin857@gmail.com,2371308@stu.neu.edu.cn

Abstract

Infectious diseases have profoundly impacted global health, economies, and social structures. Effective disease tracing is crucial for immediate public health responses and future prevention strategies. Traditional methods often focus on homogeneous networks, overlooking the diverse transmission dynamics in heterogeneous populations. This research addresses two critical challenges: the heterogeneity of transmission across various media and modes, and the significant yet underexplored influence of community structures on epidemic spread and tracing. We propose a Heterogeneous Hypergraph Attention Network (HHAN) model that accounts for multiple transmission pathways and patterns within heterogeneous networks. Specifically, HHAN integrates a heterogeneous graph neural network module to handle the complexity of communication among different populations, and an Agent-Based Modeling Module that combines agent-based ideas to model individual behaviors. This approach effectively captures complex interactions within community structures and addresses individual variability. Experimental results on three real-world datasets demonstrate that HHAN significantly outperforms other state-of-the-art methods in tackling the complex challenge of tracing infectious diseases in heterogeneous populations.

Introduction

Infectious diseases have profoundly impacted global health, economies, and social structures, from the 1918 Spanish flu to the 2020 COVID-19 pandemic (Ciotti et al. 2020). Effective tracing is crucial for managing public health crises and preventing future outbreaks. However, the COVID-19 pandemic has exposed the vulnerabilities of global health systems in handling highly transmissible diseases. Information spread is not limited to diseases alone; it also encompasses phenomena ranging from social media rumors to viruses in cyber-physical systems (Verma et al. 2023; Bobbio et al. 2023; Cheng et al. 2024), making source detection in fields such as fake news and malware defense equally critical.

Traditional tracing methods often prove inadequate for large-scale transmission. For instance, the shortest path algorithm in the Susceptible-Infectious-Recovered model (Chai, Wang, and Zhu 2021) identifies multiple source nodes, while Luo, Tay, and Leng (2017) proposed to use the Jordan center as a universal infection source estimator. However, these approaches impose strict constraints on information diffusion patterns, which are difficult to meet in real-world applications. To address this, various estimators (Yang et al. 2020; Altarelli et al. 2014; Kumar, Borkar, and Karamchandani 2017) have been introduced to relax the constraints from tree networks to general networks, while recent Graph Neural Network (GNN) algorithms have also been developed to identify source nodes in various diffusion models. However, these methods primarily focus on homogeneous networks, neglecting the complexities of heterogeneous networks and the variability in infection rates among individuals (illustrated in Figure 1). These general approaches fail to capture the need for precise, individualized modeling.

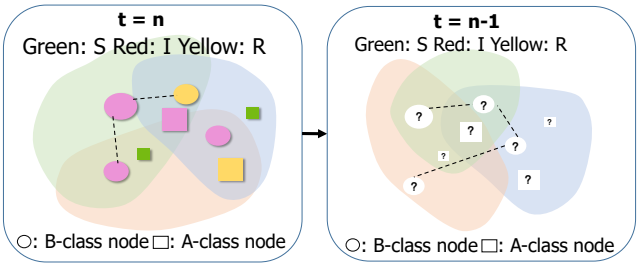

Figure 1: Tracking node states in a heterogeneous hypergraph. Squares and circles represent two node types, with green, pink and yellow indicating S, I and R states, respectively. Irregular circles denote hyperedges, while dotted lines represent ordinary edges. In the right figure, white nodes with question marks indicate unknown states.

The COVID-19 pandemic has underscored the challenges of tracing transmission across heterogeneous mediums. For instance, healthcare workers follow fixed activity patterns, while the general public exhibits more random behavior, resulting in varying infection probabilities and heterogeneous transmission groups. Although COVID-19 primarily spreads through droplets and aerosols, environmental factors like pollutants have also been linked to transmission in coronaviruses such as MERS-CoV and SARS-CoV-2 (Kim et al. 2017; Cai et al. 2020). Additionally, the virus's ability to survive on surfaces for days (Kasloff et al. 2021) compli-

*Corresponding author.

cates transmission control and contact tracing efforts. Consequently, modeling disease transmission in heterogeneous networks is essential for accurately simulating large-scale epidemics. Tracing early infections is equally critical to understanding transmission dynamics and shaping effective prevention strategies, enabling public health officials to optimize interventions for future epidemics.

Community structure plays a pivotal role in epidemic transmission. Highly connected networks increase transmission risk, with infection pressure rising as the number of infectious neighbors grows (Nunn et al. 2015). While studies have examined transmission within families, workplaces, and schools, comprehensive models must account for smaller units within communities where infection risk is nonlinear. Previous research has explored network dynamics with overlapping community structures and varying infection rates within groups (Bodó, Katona, and Simon 2016). A hypergraph-based epidemic tracing model provides a more accurate representation of infection pressure and nonlinear interactions within communities, offering deeper insight into the complexities of infectious disease spread.

To address the challenges of heterogeneous networks and the need for precise, community-structure-based modeling, we propose a novel method to identify all early infected individuals. It provides a comprehensive understanding of transmission dynamics, offering insights into pathways and enabling more effective epidemic control. Experimental results on three real-world datasets demonstrate our effectiveness. The main contributions of this study are as follows: **(1) Heterogeneous Hypergraph Attention Network Model (HHAN)**: We introduce the novel HHAN model for precise tracing in complex infectious disease scenarios. By capturing heterogeneity in transmission mediums (e.g., waterborne vs. interpersonal) and groups (e.g., healthcare workers vs. the general public), this model accurately represents disease spread dynamics. **(2) Agent-Based Optimization for Tracing**: The model integrates agent-based approaches with hypergraph neural networks to improve tracing accuracy and efficiency using deep learning. By simulating individual behaviors and both direct and indirect infection pathways, the model offers a realistic, detailed representation of disease spread. **(3) Modeling Community Structure in Epidemic Spread**: By leveraging hypergraphs to represent community structures (e.g., families, workplaces), the model incorporates nonlinear relationships and infection pressure within these units. This realistic simulation framework supports more effective epidemic modeling and control strategies in complex real-world scenarios.

Related Work

Our work is related to two areas of research: agent-based modeling and source tracing in complex networks.

Agent-Based Modeling (ABM): It is a key tool for simulating complex systems through agent interactions, widely used in infectious disease research. Prajapati, Bhaumik, and Kumar (2023) combined the SIR model with hybrid CNN-LSTM networks to assess epidemic spread in urban areas and proposed mitigation strategies. Antelmi et al. (2020) introduced a time-varying hypergraph model to incorporate human mobility and multiple transmission pathways, enhancing the accuracy of traditional models. Saliba et al. (2023) demonstrated ABM's potential in source tracing by simulating disease transmission under various parameters and behaviors. These studies highlight ABM's effectiveness in source identification and transmission.

Source Tracing: Various methods have been developed for identifying information sources. Wang et al. (2015) showed that multiple observations enhance detection rates, whereas Zhu and Ying (2014) proposed a sample-path method within an SIR framework to identify Jordan-infection-centers (JICs). The types-center method has been used to estimate sources on tree networks (Kesavareddigari et al. 2019). Jiang et al. (2016) modeled time-varying networks through snapshots, though traditional Maximum Likelihood Estimation (MLE) methods struggle with multi-source scenarios. Label propagation was adapted for single-source detection (Fan et al. 2020) in static networks (Hu et al. 2017), but with limited application to time-varying contexts. DDMIX (Li et al. 2023), utilizing a Variational Autoencoder (VAE) to reconstruct dissemination paths, suffers from reduced accuracy over time due to the increasing complexity of the solution space. SLVAE (Ling et al. 2022) localizes sources but fails to reconstruct paths. Lastly, DDMSL (Yan, Fang, and He 2024) proposed a probabilistic model leveraging diffusion processes for source localization in complex networks. However, these methods mainly focus on homogeneous networks.

The HHAN Model

Here, we introduce the HHAN model and provide an explanation of the problem definition, propagation process, and model design. As illustrated in Figure 3, the model consists of two modules: Agent-Based Modeling Module and Heterogeneous Graph Neural Network Module.

Problem Definition

To adapt to the complex transmission context of infectious diseases, we divide the nodes into two categories. Type A nodes represent special groups or sources of transmission, such as infectious items, with higher transmission probabilities and more complex transmission characteristics. Type B nodes represent the general population, primarily transmitting through common modes such as contact transmission. The transmission probabilities and pathways differ between Type A and Type B nodes. We first define a hypergraph (Feng et al. 2019). A hypergraph consists of a set of nodes $V = V_A \cup V_B$, where $V_A = \{a_1, a_2, \ldots, a_{N_A}\}$ represents the set of A-type nodes, and $V_B = \{b_1, b_2, \ldots, b_{N_B}\}$ represents the set of B-type nodes, with N_A and N_B being the number of A-type and B-type nodes, respectively. The hypergraph also includes a set of hyperedges $E_H = \{e_1, e_2, \ldots, e_{M_H}\}$, where each hyperedge $e_i \subseteq V$ for all $i = 1, 2, \ldots, M_H$; and a set of normal edges $E_O = \{f_1, f_2, \ldots, f_{M_O}\}$, where each normal edge $f_j \subseteq V$ for all $j = 1, 2, \ldots, M_O$. The set (V, E_H, E_O) is referred to as a heterogeneous hypergraph.

Nodes represent individuals, and hyperedges represent community structures such as households, workplaces, or

restaurants. We consider hyperedges to be infectious. Source nodes undergo diffusion over the graph G governed by a diffusion model $g(\cdot)$ over T time steps, with the set of node states at time step t represented as $X_t = \{x_1^t, \ldots, x_N^t\}$, where $x_i^t \in \{0,1\}^M$. For example, if $g(\cdot)$ is the SIR model, then $M \in \{S, I, R\}$, where S, I, R denote susceptible, infected, and recovered states, respectively. We define the research problem as finding the sequence of latent states $\mathbf{z} = \{\mathbf{z}_0, \ldots, \mathbf{z}_{T-1}\}$ that maximizes the likelihood function:

$$\mathbf{z}^* = \arg\max P(\mathbf{z}_T \mid \mathbf{z}, G, P_{\text{infected}}(G)) \quad (1)$$

given the observed sequence of node states $\mathbf{z}_T$. Here, $P_{\text{infected}}(G)$ represents the infection probability, which considers multiple transmission pathways, such as direct, indirect, and hyperedge-related diffusion.

After incorporating the multi-path transmission probabilities, the expression for the likelihood function becomes:

$$P(\mathbf{z}_T \mid G, \theta) = \prod_{t=1}^{T} \prod_{i \in V} P(\mathbf{z}_i(t) \mid \mathbf{z}_i(t-1), G, P_{\text{infected}}(i,t), \theta) \quad (2)$$

where θ represents the model parameters, $\mathbf{z}_i(t)$ denotes the probability of node i being in a specific state at time t, and $P_{\text{infected}}(i,t)$ is the infection probability of node i at time t, considering all relevant transmission pathways.

The formulation aims to find the most probable sequence of latent states $\mathbf{z}$ given the observed data and the underlying graph structure, considering the complex diffusion dynamics driven by multiple infection transmission mechanisms.

Symbol	Meaning
V	Node set, divided into Type A nodes (V_A) and Type B nodes (V_B).
E	Edge set, including ordinary edges (E_O) and hyperedges (E_H).
X_t	Node state set at time step t, $X_t = \{x_1^t, x_2^t, \ldots, x_N^t\}$, where $x_i^t \in \{0,1\}^M$ represents the state of node i at time t.
z	Latent state sequence, the goal is to maximize the likelihood function to predict these states, $z = \{z_0, z_1, \ldots, z_{T-1}\}$.
τ	Propagation rate, including direct, hyperedge and indirect propagation rates τ_{direct}, τ_{hyper}, and $\tau_{indirect}$, respectively.
P	Infection probability, combined by direct infection probability ($P_{\text{infected}}^{\text{direct}}$), hyperedge infection probability ($P_{\text{infected}}^{\text{hyper}}$), and indirect infection probability ($P_{\text{infected}}^{\text{indirect}}$).
γ	Recovery rate, including the recovery rate for Type A nodes (γ_A) and Type B nodes (γ_B).
H	$H^{(1)}$ and $H^{(2)}$ are the input feature matrices for the first and second convolution layers, representatively, including feature matrices for different node types (e.g., $H_{\text{BB}}^{(2)}$, $H_{\text{AC}}^{(2)}$, $H_{\text{BC}}^{(2)}$), and H_{final} is the final feature matrix obtained after weighted concatenation.
X	Feature matrix after feature engineering, including node features, lagged features, and graph structure features.

Table 1: Symbol descriptions.

Agent-Based Modeling Module

The Susceptible-Infected-Recovered (SIR) and Susceptible-Infected (SI) models (Allen 1994) are widely recognized frameworks for simulating diffusion phenomena in natural systems (Wang et al. 2024; Jahanbakhsh-Nagadeh et al. 2023). These models are fundamental in epidemiology and information diffusion studies due to their simplicity and ability to capture key dynamics of spreading processes. Our model can be applied to both the SI model and the SIR model. In this paper, we employ the SIR model to demonstrate our methodology. The SIR model classifies node states into three distinct categories: Susceptible (S), Infected (I), and Recovered (R). Notably, transitions between these states are irreversible, reflecting the permanent immunity or removal from the susceptible population post-recovery. To enhance the realism and applicability of our model, we integrate an ABM (Antelmi et al. 2020) approach. The ABM framework allows us to simulate the behavior of individual agents (nodes) within the network, providing a more granular view of the diffusion process. In our approach, the tracing process is divided into several discrete steps. During each simulation step, every agent operates independently, executing its step function and updating its internal state. These updates take effect in the subsequent simulation phase at time $t+1$, ensuring that the system evolves in a stepwise and controlled manner.

Our model incorporates two distinct types of nodes and edges to represent different aspects of the diffusion process. Specifically, we define A-type nodes as contaminants, and B-type nodes as humans, representing the general population. Additionally, the model includes two types of edges: hyperedges and normal edges. Hyperedges are employed to represent indirect transmission through community structures, such as shared environments or networks that connect multiple nodes simultaneously. In contrast, normal edges represent direct contact between nodes, such as person-to-person interactions. At each time step t, B-type nodes can exist in one of the three possible states:

- **Susceptible (S) State:** A B-type node remains in the S state if it was also in the S state in the previous time step, indicating that it has not yet been infected.
- **Infected (I) State:** A B-type node may transition to the I state if it was either in the I state during the previous time step and has not yet recovered or if it was in the S state and became infected during the current time step. The infection could result from direct contact with another infected B-type node via a normal edge or from indirect transmission through a hyperedge.
- **Recovered (R) State:** A B-type node transitions to the R state if it was in the I state in the previous time step and has since recovered, or if it was already in the R state in the previous time step, indicating sustained immunity.

While A-type nodes also follow a similar infection process, there are key differences in their behavior compared to B-type nodes. A-type nodes do not transmit infection to other A-type nodes. Additionally, the infection rate through hyperedges is calculated based on the number of infected nodes connected by the hyperedges. However, once the number of infected nodes connected to a hyperedge reaches a certain threshold, the infection rate becomes constant, reflecting a saturation effect where additional infected nodes no longer significantly increase the transmission probability.

Figure 2 illustrates the state transitions and their causes during the transmission process between two types of nodes

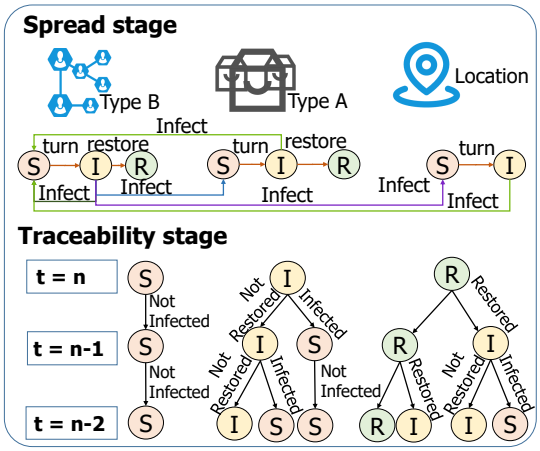

Figure 2: State transitions during the spread stage.

(A and B) and locations. When a B-type node is susceptible, it can be infected through self-propagation, A-type nodes, or locations. Once in the infected state, a B-type node can either recover or continue to spread. An A-type node in the susceptible state can only be infected by a B-type node, while in the infected state, it can recover. Locations only have susceptible and infected states, and they can only infect and be infected by B-type nodes. The tracing process is also shown, where we infer the possible previous states for nodes in the recovered state (R) and infected state (I) at the previous time step, and identify the possible infection sources for infected nodes. If a node is in the susceptible state (S), its previous state can only have been susceptible.

The infection probability calculations are as follows (taking B-type nodes as an example): **(1)** The direct infection probability $P^{\text{direct}}_{\text{infected}}$ is calculated as:

$$P^{\text{direct}}_{\text{infected}} = 1 - \exp\left(-\tau_{\text{direct}} \sum_{j \in N_i^{\text{direct}}} J_{ij} f((x^t D)_j) \Delta t\right) \tag{3}$$

where τ_{direct} is the direct transmission rate, N_i^{direct} represents the set of directly neighboring infected nodes of node i, D_{ij} is the connection strength between nodes i and j, $f((x^t D)_j)$ is a function representing the state of node j at time t, and Δt is the time step. **(2)** The hypergraph infection probability $P^{\text{hyper}}_{\text{infected}}$ is calculated as:

$$P^{\text{hyper}}_{\text{infected}} = 1 - \exp\left(-\tau_{\text{hyper}} \sum_{j \in N_i^{\text{hyper}}} J_{ij} f((x^t D)_j) \Delta t\right) \tag{4}$$

where τ_{hyper} is the hypergraph transmission rate, and N_i^{hyper} represents the set of neighboring infected nodes of node i in the hypergraph. **(3)** The indirect infection probability $P^{\text{indirect}}_{\text{infected}}$ is calculated as:

$$P^{\text{indirect}}_{\text{infected}} = 1 - \exp\left(-\tau_{\text{indirect}} \sum_{j \in N_i^{\text{indirect}}} J_{ij} f((x^t D)_j) \Delta t\right) \tag{5}$$

where τ_{indirect} is the indirect transmission rate, and N_i^{indirect} is the set of indirectly neighboring infected nodes of node i.

(4) The total infection probability P_{infected} is calculated as:

$$P_{\text{infected}} = P^{\text{direct}}_{\text{infected}} + P^{\text{hyper}}_{\text{infected}} + P^{\text{indirect}}_{\text{infected}} - \left(P^{\text{direct}}_{\text{infected}} \cdot P^{\text{hyper}}_{\text{infected}} \cdot P^{\text{indirect}}_{\text{infected}}\right) \tag{6}$$

Equation 6 combines the effects of direct, hypergraph, and indirect infections, while subtracting the overlap to avoid double-counting. **(5)** The recovery probability for A-type nodes $P^{A}_{\text{recovered}}$ is calculated as:

$$P^{A}_{\text{recovered}} = 1 - \exp\left(-\gamma_A \Delta t\right) \tag{7}$$

where γ_A is the recovery rate for A-type nodes. Similarly, the recovery probability for B-type nodes $P^{B}_{\text{recovered}}$ is:

$$P^{B}_{\text{recovered}} = 1 - \exp\left(-\gamma_B \Delta t\right) \tag{8}$$

where γ_B is the recovery rate for B-type nodes.

Initially, the infection rate and recovery rate are calculated using the known set of initially infected nodes and the SIR model formula, with parameter selection guided by epidemic dynamics estimation methods. The evolution of infection transmission is then determined based on the transmission rules between nodes. To more accurately simulate the process, a nonlinear infection pressure formula adjusts the infection probability, allowing the infection rate $\beta(t)$ to vary with the nodes' infection pressure. As the number of infected individuals surpasses a threshold, $\beta(t)$ stabilizes at a constant value. The recovery rate γ is linked to the average recovery time of infected nodes, with shorter recovery times corresponding to higher γ values, and longer recovery times corresponding to lower γ values.

Heterogeneous Graph Neural Network Module

The model transforms the hypergraph, where hyperedges possess infectious properties, enabling them to both become infected and transimit infections to other nodes. Thus, hyperedges are treated as a new node type C, while nodes of types A and B within the same hyperedge are considered connected to the corresponding C-type node. This transformation simplifies the network structure, facilitating more representative interactions among different node types. Hyperedges Q are thus defined as follows:

$$Q = \{C_1, C_2, \ldots, C_m\} \tag{9}$$

where each C_i contains several nodes, represented as:

$$C_i = \{A_j, B_k \mid j \in J_i, k \in K_i\} \tag{10}$$

where J_i and K_i represent the index sets of A-type and B-type nodes associated with the hyperedge C_i, respectively. This transformation enhances the interpretability of the network and lays the foundation for subsequent feature extraction and model training.

Next, we process the features of the three types of nodes A, B, and C through feature engineering. We utilize both current and lagged features to capture the dynamic trends of features over time. Lagged features effectively reflect the autocorrelation of time series data, aiding the model in understanding the timeliness of infection transmission. By introducing a time window Δ_t, we construct the feature matrix at time t as:

$$X_t = \begin{bmatrix} X_{A,t}, & X_{A,t-\Delta t} \\ X_{B,t}, & X_{B,t-\Delta t} \\ X_{C,t}, & X_{C,t-\Delta t} \end{bmatrix} \tag{11}$$

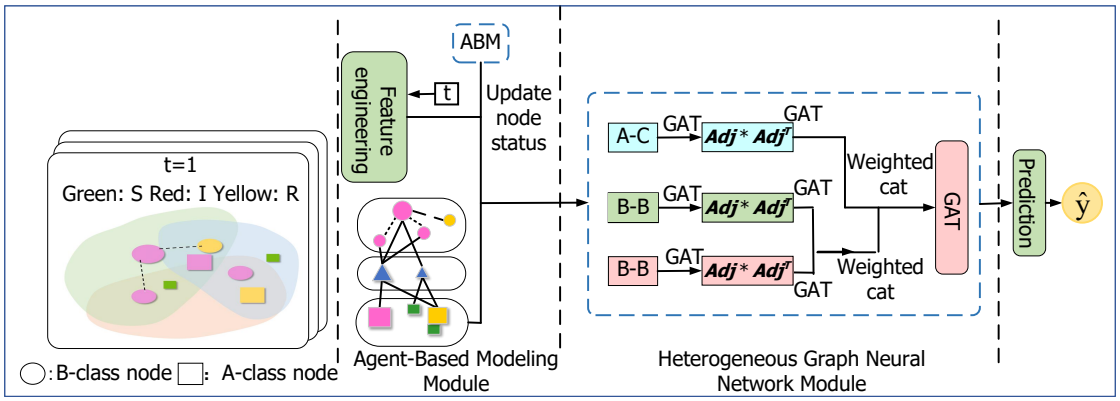

Figure 3: Overall framework of HHAN.

where $X_{A,t}$, $X_{B,t}$, and $X_{C,t}$ represent the feature matrices of node types A, B, and C at time t, and $X_{A,t-\Delta t}$, $X_{B,t-\Delta t}$, and $X_{C,t-\Delta t}$ denote their corresponding lagged features.

In addition to the infection rate $I(t)$ and recovery rate $R(t)$, we incorporate graph features, such as node degree, clustering coefficient, and shortest path. These features describe the network position of nodes and their potential impact during the transmission process. By integrating graph features into the feature matrix, we represent the complete feature matrix for the network as:

$$X = \begin{bmatrix} X_{A,t}, & X_{A,t-\Delta t} & F_A \\ X_{B,t}, & X_{B,t-\Delta t} & F_B \\ X_{C,t}, & X_{C,t-\Delta t} & F_C \end{bmatrix} \tag{12}$$

where F_A, F_B, and F_C are the graph features related to node types A, B, and C, respectively.

To improve computational efficiency and mitigate the risk of overfitting, we standardize and reduce the dimensionality of the processed features. Assuming mean-variance standardization is used, the standardized features X' are:

$$X' = \frac{X - \mu}{\sigma} \tag{13}$$

where μ and σ represent the mean and standard deviation of each feature in the matrix X, respectively. The standardization process unifies the scales of different features, ensuring balanced learning across all features. After standardization, dimensionality reduction techniques, such as Principal Component Analysis (PCA), are employed to lower the feature dimensions. This approach retains most of the critical information while reducing computational complexity.

The processed features, along with the hypergraph network, are input into the model, where Graph Attention Network (GAT) convolution layers are used to assign varying weights to different connections. Specifically, the GAT layer updates node features by computing the attention weight α_{ij} for each pair of adjacent nodes using the following formula::

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T[Wh_i \parallel Wh_j]))}{\sum_{k \in N(i)} \exp(\text{LeakyReLU}(a^T[Wh_i \parallel Wh_k]))} \tag{14}$$

where W is the weight matrix applied to the node features, a is the weight vector for attention calculation, $\parallel$ denotes vector concatenation, and $N(i)$ is the set of adjacent nodes of node i.

In the first convolution layer, we primarily focus on the connections between type A and type B nodes to type C nodes, as well as the direct connections between type B nodes, without fully capturing the relationships of other neighboring nodes within the same hyperedge. In other words, the initial convolution emphasizes specific node connections while overlooking interactions among other nodes within the same hyperedge. Therefore, we design a second layer of convolution that learns the relationships of connected neighboring nodes within the same hyperedge by multiplying the adjacency matrix U with its transpose U^T. This process can be represented as:

$$H^{(2)} = \sigma(UU^T H^{(1)} W^{(2)}) \tag{15}$$

where U is the adjacency matrix, U^T is its transpose, σ is the activation function, $H^{(1)}$ is the input feature matrix at layer 2, and $W^{(2)}$ is the weight matrix at layer 2.

After completing the learning of different parts of the network, we perform weighted concatenation to integrate information from each part, thus providing a richer feature representation for the overall network. The weighted concatenation operation can be expressed as:

$$H_{\text{final}} = \text{GAT}\left([H_{\text{BB}}^{(2)}, H_{\text{AC}}^{(2)}, H_{\text{BC}}^{(2)}]\right) \tag{16}$$

H_{final} is the final feature matrix after the GAT operation. $H_{\text{BB}}^{(2)}$, $H_{\text{AC}}^{(2)}$, and $H_{\text{BC}}^{(2)}$ represent the feature matrices of type BB, AC, and BC nodes at layer 2, respectively. These are concatenated and passed through the GAT to obtain the final representation.

Additionally, our model incorporates a source tracing algorithm based on an agent-based model (ABM) to enhance model performance and practicality. The agent-based model provides a flexible framework for tracing infection sources, allowing us to simulate transmission behavior under various strategies. Simultaneously, the tracing algorithm optimizes the modeling of relationships between nodes through deep learning mechanisms, further improving the model's accuracy and reliability in practical applications.

In particular, the model leverages the Heterogeneous Graph Neural Network (HGNN) Module to update node states. The HGNN module is used to update the latent state of each node based on its previous state and the states of its neighboring nodes. The state update function is given by:

$$\mathbf{z}_i(t) = \text{HGNN}(\mathbf{z}_i(t-1), \{\mathbf{z}_j(t-1)\}_{j \in N(i)}, G) \tag{17}$$

The Heterogeneous Graph Neural Network Module updates the latent state $\mathbf{z}_i(t)$ of node i, combining its previous state $\mathbf{z}_i(t-1)$ with the previous states $\mathbf{z}_j(t-1)$ of its neighboring nodes $j \in N(i)$, aggregating and updating them accordingly. Through deep learning, we define the objective function L as:

$$\begin{aligned} L(\mathbf{Z}) = -\sum_{t=0}^{T-1} \sum_{i \in V} [y_i(t) \log \sigma(\mathbf{z}_i(t)) \\ +(1-y_i(t)) \log(1-\sigma(\mathbf{z}_i(t)))] \end{aligned} \tag{18}$$

where $y_i(t)$ is the true label of node i at time step t, $\sigma(\mathbf{z}_i(t))$ is the predicted probability of node i being infected (or in a specific state) at time t, and $\mathbf{z}_i(t)$ is the latent state of node i at time t. The objective function minimizes the difference between the predicted and true values while optimizing the state transitions based on the HHAN framework. This objective function aims to minimize the error between predicted and true labels while enhancing the model's generalization ability by controlling its complexity. The final output of the model is a binary classification prediction for each node.

Experiments

In this section, we conduct experiments on three widely used real-world datasets to evaluate HHAN's effectiveness.

Experimental Settings

Datasets: During the data preprocessing stage, we performed community detection on each dataset and represented the connections between nodes within a community using hyperedges. ACM Hypertext **Conference** Dataset: it was collected during the 2009 ACM Hypertext Conference, where the SocioPatterns project deployed the Live Social Semantics application. Attendees voluntarily wore radio badges to monitor their face-to-face distances. The published version represents a dynamic network of approximately 110 attendees over a period of about 2.5 days (Isella et al. 2011). **School** Dataset: it corresponds to the contact and friendship relationships among students at a high school in Marseille, France, measured using various techniques in December 2013 (Mastrandrea, Fournet, and Barrat 2015). **Hospital** Dataset: it contains the contact network between patients and healthcare workers (HCWs) within a hospital ward in Lyon, France, from 1:00 PM on December 6, 2010, to 2:00 PM on December 10, 2010, and 46 HCWs and 29 patients are included (Vanhems et al. 2013).

We conduct SIR diffusion simulations on each dataset by randomly selecting 5% of the nodes as source nodes at the initial time, and stopping the simulation when approximately 30% of the nodes are infected. Each generated dataset is randomly divided into training, validation, and test sets in an 8:1:1 ratio. Selecting 5% of the nodes as sources models a small-scale initial infection, consistent with common assumptions in epidemic spread models, while the proportion can vary depending on the specific transmission scenario. The simulation is halted at a 30% infection rate to balance computational complexity with simulating an extended transmission process. This threshold is generally recognized as indicative of large-scale spread, enabling the evaluation of the model's accuracy in identifying source nodes while ensuring timely completion of the experiment.

Evaluation Metrics: Our objective is to trace the early states of nodes, framed as a binary classification problem. To evaluate the model, we use five metrics: Accuracy (ACC), which is the proportion of correctly classified nodes; Precision (PR), representing the proportion of nodes correctly identified as true sources; Recall (RE), indicating the proportion of actual source nodes that are correctly predicted; F1 Score, the harmonic mean of PR and RE, which balances precision and recall; and ROC-AUC (AUC), which measures the model's ability to classify accurately and reflects its performance across different thresholds. These metrics assess not only the model's performance at the initial moment but also its ability to correctly identify source nodes throughout the entire transmission process as time progresses.

Baselines and Experimental Settings: We compare four source tracing algorithms with HHAN: DDMIX (Li et al. 2023), which uses a Variational Autoencoder (VAE) to reconstruct the epidemic transmission process; LPSI (Wang et al. 2017), which learns source detection by aggregating information via label propagation; SLVAE (Ling et al. 2022), which applies a generative model to learn the distribution of source nodes; and CULT (Rozenshtein et al. 2016), which formulates the problem as a temporal Steiner-tree computation and designs a fast algorithm leveraging the specific problem structure. The model is trained with a dynamically adjusted learning rate using a learning rate scheduler. A dropout rate of 0.4 is applied after each GAT layer to prevent overfitting, and the AdamW optimizer is used with a learning rate of 0.005 and a weight decay of 1×10^{-4}. The learning rate scheduler reduces the learning rate by half if validation performance plateaus, with a minimum learning rate of 1×10^{-6}. The batch size is set to 20, and training is conducted over 300 epochs to ensure sufficient learning and convergence of the model.

Experimental Results

Comprehensive Comparative Results: In our study, we conduct a comprehensive comparative analysis of the HHAN algorithm against various diffusion models, placing particular emphasis on the widely recognized SIR model.

First, HHANS's performance is presented in Tables 3 and 4, which provide a comprehensive overview of our performance across different datasets, with results averaged over multiple experimental rounds. Table 3 specifically highlights HHANS's performance across the three distinct datasets. The results clearly indicate that HHAN consistently achieves high levels of all metrics across these datasets, demonstrating its robustness and effectiveness in diverse environments. Notably, HHAN performs exception-

	Hospital					Conference					School				
Model	**Acc**	**Pre**	**Recall**	**F1**	**AUC**	**Acc**	**Pre**	**Recall**	**F1**	**AUC**	**Acc**	**Pre**	**Recall**	**F1**	**AUC**
SLVAE	0.209	0.209	**1.000**	0.345	0.515	0.111	0.111	**1.000**	0.200	0.500	0.111	0.111	**1.000**	0.200	0.500
DDMIX	0.209	0.209	**1.000**	0.345	0.499	0.105	0.105	**1.000**	0.190	0.500	0.041	0.041	**1.000**	0.079	0.500
LPSI	0.068	**1.000**	0.048	0.091	0.524	0.012	**1.000**	0.007	0.014	0.504	0.190	**1.000**	0.190	0.314	0.644
CULT	0.9087	0.460	0.036	0.068	0.516	0.887	0.430	0.034	0.063	0.514	**0.940**	0.430	0.022	0.042	0.510
HHAN	**0.955**	0.765	0.636	**0.694**	**0.809**	**0.981**	0.977	0.806	**0.883**	**0.902**	0.765	0.530	0.919	**0.672**	**0.814**

Table 2: Performance comparisons across different datasets.

Dataset	Accuracy	Precision	Recall	F1 Score	AUC
Conference	**0.981**	**0.977**	**0.806**	**0.883**	**0.902**
Hospital	0.955	0.765	0.636	0.694	0.809
School	0.765	0.530	0.919	0.672	0.814

Table 3: HHAN's performance across the three datasets.

Dataset	Accuracy	Precision	Recall	F1 Score	AUC
Conference	**0.972**	**0.907**	**0.876**	**0.891**	**0.931**
Hospital	0.948	0.825	0.604	0.697	0.795
School	0.955	0.793	0.592	0.678	0.789

Table 4: HHAN's average performance over time series.

ally well on Conference, achieving an accuracy of 98.10% and an AUC of 90.20%, highlighting its ability to accurately and reliably handle complex and diverse data. Table 4 further shows the results for node state tracing over the entire time series.

Further, Table 2 illustrates HHAN's superiority over other state-of-the-art algorithms, including SLVAE, DDMIX, LPSI, and CULT, on the three datasets. The comparative results reveal that HHAN outperforms these alternative models by a significant margin, particularly in terms of accuracy and F1 score. Models such as SLVAE, DDMIX, and LPSI often struggle to achieve balanced performance, typically exhibiting either high recall but low precision, or vice versa. In contrast, HHAN maintains a strong balance across all evaluation metrics. These findings collectively highlight HHAN's robustness, confirming its potential as a leading approach for tasks involving heterogeneous data and complex relationships within networked systems. Moreover, Figure 4 shows HHAN's performance after 300 epochs under the same experimental conditions. It illustrates the stability and convergence of our model over time, offering valuable insights into its training dynamics and overall effectiveness in source tracing tasks. This further emphasizes HHAN's potential as a reliable tool for managing complex diffusion processes in various real-world applications.

Ablation Study:

To rigorously assess the importance of each model component, we conduct ablation studies on Conference, holding other parameters constant. Specifically, we consider three variants: **HHAN1**, where feature engineering is removed and manual feature selection is used; **HHCN**, where the Graph Attention Network is replaced by a Graph Convolutional Network; and **HHAN2** which omits the ABM component. Experimental results in Table 5 shows that HHAN consistently performs better than the three variant, demonstrating that feature engineering, attention mechanisms, and ABM each contribute significantly to the model's performance. Their combined use is essential for achieving optimal results in complex networks, affirming that their synergy enhances the model's ability to handle real-world data.

Model	Accuracy	Precision	Recall	F1 Score	AUC
HHAN1	0.871	0.609	0.609	0.609	0.766
HHCN	0.921	0.731	0.826	0.776	0.883
HHAN2	0.935	0.792	0.826	0.809	0.892
HHAN	**0.981**	**0.977**	**0.806**	**0.883**	**0.902**

Table 5: Ablation study.

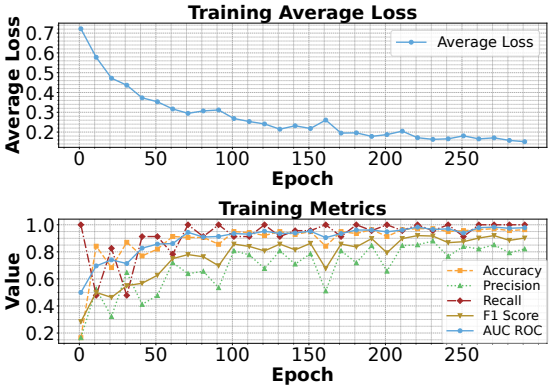

Figure 4: Loss of HHAN under SIR model with epochs.

Conclusions

In this paper, we introduced a novel heterogeneous hypergraph neural network for tracing infection sources, achieving significant performance improvements on three real-world datasets. Our model enhances accuracy through feature engineering, attention mechanisms, and ABM. Our model shows significant advantages. Unlike traditional methods that focus only on data from the initial day of an epidemic, our model captures and analyzes relationships across all days, offering a comprehensive temporal view of disease transmission dynamics. Future work can consider to explore more dynamic network structures, more advanced feature selection, and more comprehensive comparisons with existing methods.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62202089, U22A2004, 72371148 and 72192832), in part by the Shanghai Rising-Star Program (Grant No. 23QA1403100), in part by the Fundamental Research Funds for the Central Universities under Grant N2319004.

References

Allen, L. J. 1994. Some discrete-time SI, SIR, and SIS epidemic models. *Mathematical biosciences*, 124(1): 83–105.

Altarelli, F.; Braunstein, A.; Dall'Asta, L.; Lage-Castellanos, A.; and Zecchina, R. 2014. Bayesian inference of epidemics on networks via belief propagation. *Physical review letters*, 112(11): 118701.

Antelmi, A.; Cordasco, G.; Spagnuolo, C.; Scarano, V.; et al. 2020. A design-methodology for epidemic dynamics via time-varying hypergraphs. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*, volume 2020, 61–69. International Foundation for Autonomous Agents and Multiagent Systems.

Bobbio, A.; Campanile, L.; Gribaudo, M.; Iacono, M.; Marulli, F.; and Mastroianni, M. 2023. A cyber warfare perspective on risks related to health IoT devices and contact tracing. *Neural Computing and Applications*, 35(19): 13823–13837.

Bodó, Á.; Katona, G. Y.; and Simon, P. L. 2016. SIS epidemic propagation on hypergraphs. *Bulletin of mathematical biology*, 78: 713–735.

Cai, J.; Sun, W.; Huang, J.; Gamber, M.; Wu, J.; and He, G. 2020. Indirect virus transmission in cluster of COVID-19 cases, Wenzhou, China, 2020. *Emerging infectious diseases*, 26(6): 1343.

Chai, Y.; Wang, Y.; and Zhu, L. 2021. Information sources estimation in time-varying networks. *IEEE Transactions on Information Forensics and Security*, 16: 2621–2636.

Cheng, H.; Xiao, M.; Yu, W.; Rutkowski, L.; and Cao, J. 2024. How to regulate pattern formations for malware propagation in cyber-physical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 34(3).

Ciotti, M.; Ciccozzi, M.; Terrinoni, A.; Jiang, W.-C.; Wang, C.-B.; and Bernardini, S. 2020. The COVID-19 pandemic. *Critical reviews in clinical laboratory sciences*, 57(6): 365–388.

Fan, L.; Li, B.; Liu, D.; Dai, H.; and Ru, Y. 2020. Identifying propagation source in temporal networks based on label propagation. In *Data Science: 6th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2020, Taiyuan, China, September 18-21, 2020, Proceedings, Part I 6*, 72–88. Springer.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3558–3565.

Hu, Z.-L.; Han, X.; Lai, Y.-C.; and Wang, W.-X. 2017. Optimal localization of diffusion sources in complex networks. *Royal Society open science*, 4(4): 170091.

Isella, L.; Stehlé, J.; Barrat, A.; Cattuto, C.; Pinton, J.-F.; and Van den Broeck, W. 2011. What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1): 166–180.

Jahanbakhsh-Nagadeh, Z.; Feizi-Derakhshi, M.-R.; Ramezani, M.; Akan, T.; Asgari-Chenaghlu, M.; Nikzad-Khasmakhi, N.; Feizi-Derakhshi, A.-R.; Ranjbar-Khadivi, M.; Zafarani-Moattar, E.; and Balafar, M.-A. 2023. A model to measure the spread power of rumors. *Journal of Ambient Intelligence and Humanized Computing*, 14(10): 13787–13811.

Jiang, J.; Wen, S.; Yu, S.; Xiang, Y.; and Zhou, W. 2016. Rumor source identification in social networks with time-varying topology. *IEEE Transactions on Dependable and Secure Computing*, 15(1): 166–179.

Kasloff, S. B.; Leung, A.; Strong, J. E.; Funk, D.; and Cutts, T. 2021. Stability of SARS-CoV-2 on critical personal protective equipment. *Scientific reports*, 11(1): 984.

Kesavareddigari, H.; Spencer, S.; Eryilmaz, A.; and Srikant, R. 2019. Identification and asymptotic localization of rumor sources using the method of types. *IEEE transactions on network science and engineering*, 7(3): 1145–1157.

Kim, Y.; Yang, M.; Goyal, S. M.; Cheeran, M. C.; and Torremorell, M. 2017. Evaluation of biosecurity measures to prevent indirect transmission of porcine epidemic diarrhea virus. *BMC veterinary research*, 13: 1–9.

Kumar, A.; Borkar, V. S.; and Karamchandani, N. 2017. Temporally agnostic rumor-source detection. *IEEE Transactions on Signal and Information Processing over Networks*, 3(2): 316–329.

Li, B.; Čutura, G.; Swami, A.; and Segarra, S. 2023. Deep Demixing: Reconstructing the Evolution of Network Epidemics. *arXiv preprint arXiv:2306.07938*.

Ling, C.; Jiang, J.; Wang, J.; and Liang, Z. 2022. 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*, 1010–1020.

Luo, W.; Tay, W. P.; and Leng, M. 2017. On the universality of Jordan centers for estimating infection sources in tree networks. *IEEE Transactions on Information Theory*, 63(7): 4634–4657.

Mastrandrea, R.; Fournet, J.; and Barrat, A. 2015. Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS one*, 10(9): e0136497.

Nunn, C. L.; Jordán, F.; McCabe, C. M.; Verdolin, J. L.; and Fewell, J. H. 2015. Infectious disease and group size: more than just a numbers game. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1669): 20140111.

Prajapati, S. P.; Bhaumik, R.; and Kumar, T. 2023. An intelligent ABM-based framework for developing pandemic-resilient urban spaces in post-COVID smart cities. *Procedia Computer Science*, 218: 2299–2308.

Rozenshtein, P.; Gionis, A.; Prakash, B. A.; and Vreeken, J. 2016. Reconstructing an epidemic over time. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1835–1844.

Saliba, S.; Dadgostari, F.; Hoops, S.; Mortveit, H. S.; and Swarup, S. 2023. Active Sensing for Epidemic State Estimation Using ABM-Guided Machine Learning. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, 30–45. Springer.

Vanhems, P.; Barrat, A.; Cattuto, C.; Pinton, J.-F.; Khanafer, N.; Régis, C.; Kim, B.-a.; Comte, B.; and Voirin, N. 2013. Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PloS one*, 8(9): e73970.

Verma, P.; Gupta, A.; Kumar, M.; and Gill, S. S. 2023. FCMCPS-COVID: AI propelled fog–cloud inspired scalable medical cyber-physical system, specific to coronavirus disease. *Internet of Things*, 23: 100828.

Wang, W.; Nie, Y.; Li, W.; Lin, T.; Shang, M.-S.; Su, S.; Tang, Y.; Zhang, Y.-C.; and Sun, G.-Q. 2024. Epidemic spreading on higher-order networks. *Physics Reports*, 1056: 1–70.

Wang, Z.; Dong, W.; Zhang, W.; and Tan, C. W. 2015. Rooting our rumor sources in online social networks: The value of diversity from multiple observations. *IEEE Journal of Selected Topics in Signal Processing*, 9(4): 663–677.

Wang, Z.; Wang, C.; Pei, J.; and Ye, X. 2017. Multiple source detection without knowing the underlying propagation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Yan, X.; Fang, H.; and He, Q. 2024. Diffusion model for graph inverse problems: Towards effective source localization on complex networks. *Advances in Neural Information Processing Systems*, 36.

Yang, F.; Yang, S.; Peng, Y.; Yao, Y.; Wang, Z.; Li, H.; Liu, J.; Zhang, R.; and Li, C. 2020. Locating the propagation source in complex networks with a direction-induced search based Gaussian estimator. *Knowledge-Based Systems*, 195: 105674.

Zhu, K.; and Ying, L. 2014. Information source detection in the SIR model: A sample-path-based approach. *IEEE/ACM Transactions on Networking*, 24(1): 408–421.