

## **Distribution Agreement**

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Songyuan Liu

April 9, 2025

Higher-order Interaction Matters: Dynamic Hypergraph Neural Networks for  
Epidemic Modeling

By

Songyuan Liu

Wei Jin  
Advisor

Computer Science

Wei Jin  
Advisor

Carl Yang  
Committee Member

Jiaying Lu  
Committee Member

2025

Higher-order Interaction Matters: Dynamic Hypergraph Neural Networks for  
Epidemic Modeling

By

Songyuan Liu

Wei Jin  
Thesis Advisor

An abstract of  
A thesis submitted to the Faculty of the  
Emory College of Arts and Sciences  
in partial fulfillment of the requirements for the degree of  
Bachelor of Arts with Honors  
in Computer Science  
2025

## Abstract

Higher-order Interaction Matters: Dynamic Hypergraph Neural Networks for  
Epidemic Modeling  
By Songyuan Liu

The ongoing need for effective epidemic modeling has driven advancements in capturing the complex dynamics of infectious diseases. Traditional models, such as Susceptible-Infected-Recovered, and graph-based approaches often fail to account for higher-order interactions and the nuanced structure pattern inherent in human contact networks. This study introduces a novel Human Contact-Tracing Hypergraph Neural Network framework tailored for epidemic modeling called EpiDHGNN, leveraging the capabilities of hypergraphs to model intricate, higher-order relationships from both location and individual level. Both real-world and synthetic epidemic data are used to train and evaluate the model. Results demonstrate that EpiDHGNN consistently outperforms baseline models across various epidemic modeling tasks, such as source detection and forecast, by effectively capturing the higher-order interactions and preserving the complex structure of human interactions. This work underscores the potential of representing human contact data as hypergraphs and employing hypergraph-based methods to improve epidemic modeling, providing more reliable insights for public health decision-making.

Higher-order Interaction Matters: Dynamic Hypergraph Neural Networks for  
Epidemic Modeling

By

Songyuan Liu

Wei Jin  
Thesis Advisor

A thesis submitted to the Faculty of the  
Emory College of Arts and Sciences of Emory University  
in partial fulfillment of the requirements for the degree of  
Bachelor of Arts with Honors  
in Computer Science  
2025

## Acknowledgments

I'd like to thank my advisor, Dr. Wei Jin, for all the guidance and support throughout this project. Your advice and feedback have been really helpful every step of the way. Thanks also to my thesis committee — Dr. Jiaying Lu and Dr. Carl Yang — for taking the time to read my work and offer helpful suggestions. I'm also grateful to everyone who helped me during this process, whether it was discussing ideas, reading drafts, or just offering encouragement when things got tough. Your support made a big difference.

Lastly, thank you to my parents for always supporting my education. I really appreciate everything you've done to help me get here.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background</b>	<b>5</b>
2.1	Mechanistic Epidemic Modeling . . . . .	5
2.2	Graphs for Epidemic Modeling . . . . .	7
2.3	Hypergraphs for Epidemic Modeling . . . . .	8
<b>3</b>	<b>Formulation</b>	<b>9</b>
3.1	Hypergraph . . . . .	9
3.2	Dynamic Hypergraph . . . . .	9
3.3	Epidemic Modeling Tasks . . . . .	10
3.3.1	Source Detection . . . . .	11
3.3.2	Infection Forecasting . . . . .	11
<b>4</b>	<b>Method</b>	<b>13</b>
4.1	HGNN Module . . . . .	14
4.2	Temporal Convolution Module . . . . .	14
4.3	Contact Pattern Awareness Module . . . . .	15
4.4	EpiDHGNN . . . . .	16
<b>5</b>	<b>Experiments</b>	<b>18</b>
5.1	Data Description . . . . .	18

5.2	Setup . . . . .	20
5.3	RQ1 - Performance . . . . .	20
5.4	RQ2 - Ablation . . . . .	22
5.5	RQ3 - Module Effectiveness . . . . .	23
5.6	RQ4 - Generalizability . . . . .	24
<b>6</b>	<b>Conclusion</b>	<b>26</b>
6.1	Future Work . . . . .	26
6.1.1	Hypergraph SIR . . . . .	26
6.1.2	Alternative Model Selection . . . . .	27
6.1.3	Human Contact Data Simulation . . . . .	27
6.2	Summary . . . . .	28
<b>A</b>	<b>Appendix</b>	<b>29</b>
A.1	Hypergraph SIR Simulation . . . . .	29
	<b>Bibliography</b>	<b>32</b>

# List of Figures

1.1	Illustration of how various graph encoding methods can be employed to capture complex interactions. Hypergraphs, in particular, offer significant advantages over traditional graphs by retaining both individual-level and location-level information, while also capturing higher-order interactions. This enriched representation facilitates a more nuanced understanding of epidemic dynamics. . . . .	3
4.1	Model Architecture of proposed EpiDHGNN model. The arrows in the top left corner refers to the three time stamps defined in section 3.3, where $[0 : tsh]$ is the black interval, $[tsh : ks]$ is the orange interval, and $[ks + 1 : ps]$ is the green interval. All individual state is masked to 0 in $[0 : tsh]$ as shown in the top left black module. Corresponding inputs for source detection and forecast defined in section 3.3.1 and 3.3.2 is then feed to the model as input. The light blue HyperConv module in defined in section 4.1; the dark blue temporal convolution module is defined in 4.2; and the contact pattern awareness module is defined in section 4.3. . . . .	13
5.1	Visualization of Alpha’s impact on source detection performance . . .	24

5.2 Forecast generalizability analysis. PS stands for prediction steps. The models can successfully the future infection dynamics within various PS. We also provide the Mean Absolute Error (MAE) of Naive Model (A naive time series model forecasts future values by assuming they will be the same as the most recent observed value or follow a simple pattern) and our approach. . . . . 25

# List of Tables

5.1	Dataset Summary . . . . .	19
5.2	Experiment Result of Source Detection Task on UVA Dataset. Best performance under each setting is bolded and equal performance is underlined. . . . .	20
5.3	Experiment Result of Source Detection Task on EpiSim Dataset. Best performance under each setting is bolded and equal performance is underlined. . . . .	21
5.4	Experiment Result of Forecast Task. Best Performance under each setting is bolded. . . . .	21
5.5	Ablation study on contact pattern awareness module. . . . .	22
5.6	Contact Pattern Prediction. The quantiles are ranked by the total number of interactions made in a location. . . . .	23

# Chapter 1

## Introduction

Since the onset of the COVID-19 pandemic, there has been a growing interest in studying epidemiological models [25, 22, 8]. Understanding and managing infection outbreaks is crucial for public health. Traditional mechanistic models like Susceptible-Infected-Recovered (SIR), which mathematically describe the transmission mechanisms of infectious diseases, often suffer from limitations of oversimplified or fixed assumptions, leading to sub-optimal predictive power and inefficiency in capturing complex epidemic patterns [27, 10]. (Section 2.1)

Motivated by these limitations, sequential models such as GRU [7] and LSTM [31] are used to model temporal relations in a data-driven manner. Compared to mechanistic models, sequential models have demonstrated superior performance in forecasting infection counts. However, these models often struggle to incorporate spatial dependencies, such as human mobility patterns and geographical distributions, which play a crucial role in epidemiology modeling [22]. Mobility data captures how individuals move and interact across different locations, influencing disease transmission dynamics beyond simple temporal trends. To enhance the ability to capture both spatial and temporal information, graph-based approaches have emerged as a popular tool in epidemic research. Graph Neural Networks (GNNs) [30, 21] have become popular for their

ability to model human mobility patterns. They achieve this by representing nodes as regions and weighted edges as mobility volume, effectively capturing movement between locations. Through a message-passing mechanism, GNNs enable nodes to share information with their neighbors, allowing for a more comprehensive understanding of mobility patterns. Additionally, by leveraging dynamic graph modeling and dynamic GNNs, they can further represent changes in human movement over time, enhancing their ability to model relational dynamics within mobility networks. [32, 24, 10]

Despite the utility of GNN-based methods, they primarily focus on pair-wise interactions and therefore neglect the higher-order interactions that are inherent in actual human contact networks [28, 13, 3]. Specifically, higher-order interactions refer to interactions or contacts that involve more than two individuals simultaneously in the context of epidemic modeling [18]. For example, public transportation, workplaces, and schools are shared spaces where groups of people interact following higher-order transmission dynamics. As illustrated in Figure 1.1, while standard graphs can model these interactions by representing individuals as nodes and forming fully connected subgraphs for each group, this approach is often inefficient and obscures the true higher-order structure. In contrast, hypergraphs provide a more natural and explicit way to represent higher-order interactions through hyperedges, eliminating the artificial clique. Additionally, hypergraphs can model overlapping interactions by representing locations as hyperedges, encompassing multiple individuals simultaneously. These enhancements can lead to more accurate and interpretable modeling of epidemics than standard graphs [28, 13, 5].

As illustrated in Figure 1.1, a fundamental limitation of prior graph-based approaches lies in their inability to simultaneously preserve both individual-level and location-level information, as well as their failure to capture higher-order interactions. These shortcomings significantly hinder the accurate modeling of real-world human contact patterns, which are essential for understanding and predicting the spread of

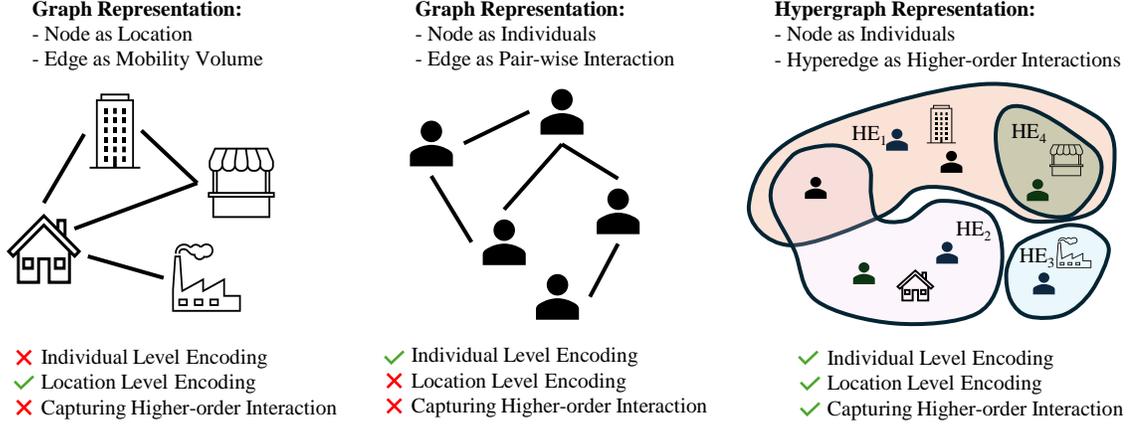


Figure 1.1: Illustration of how various graph encoding methods can be employed to capture complex interactions. Hypergraphs, in particular, offer significant advantages over traditional graphs by retaining both individual-level and location-level information, while also capturing higher-order interactions. This enriched representation facilitates a more nuanced understanding of epidemic dynamics.

infectious diseases. To address this, we propose **EpiDHGNN**, a novel framework that models human contact data as dynamic hypergraphs. This approach enables the encoding of complex, higher-order interactions and supports a richer, more granular representation of epidemic dynamics. The major contributions of this paper are threefold:

- (1) We propose a novel method to model human contact as dynamic hypergraphs, which encodes nodes as individuals and hyperedges as locations, leveraging both granular level information and higher-order interactions.
- (2) We develop EpiDHGNN, a model tailor-made for epidemic modeling with a self-supervised contact-pattern awareness module, capturing the higher-order interactions and contact patterns that are inherent in human contacts.
- (3) Extensive experiments are conducted to demonstrate the superiority of encoding human contact as hypergraphs, as well as the effectiveness of our proposed models in various epidemic tasks.

Our preliminary results indicate that EpiDHGNN significantly improve the performance of epidemic models by capturing higher-order interactions. This advancement holds the potential to enhance infection control strategies in healthcare environments,

ultimately contributing to better public health outcomes.

# Chapter 2

## Background

### 2.1 Mechanistic Epidemic Modeling

In the past, when data was not sufficiently recorded, scientists were unable to build empirical models that successfully captured the dynamics of epidemics. Empirical models rely heavily on accurate and comprehensive data to make predictions and understand patterns. In contrast, mechanistic models[20, 11] are designed to capture the underlying complexity of infections and the recovery processes of various diseases, even with limited data. Among mathematical models of infectious disease spread, *compartmental models* are among the most foundational and widely used. One of the most prominent examples is the *Susceptible-Infectious-Recovered (SIR)* model [15]. This model segments the total population  $N$  into three compartments:  $S(t)$  for the number of susceptible individuals,  $I(t)$  for infectious individuals, and  $R(t)$  for recovered (and immune) individuals, such that  $S(t) + I(t) + R(t) = N$ , assuming a closed population with no births or deaths.

The dynamics of the SIR model are governed by the following system of ordinary differential equations:

$$\frac{dS}{dt} = -\beta \frac{SI}{N}$$

$$\frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Here, the parameter  $\beta$  represents the transmission rate, capturing the likelihood of disease spread per contact between susceptible and infectious individuals, while  $\gamma$  denotes the recovery rate, corresponding to the inverse of the average infectious period. The ratio  $\mathcal{R}_0 = \frac{\beta}{\gamma}$  defines the basic reproduction number, a critical threshold quantity indicating whether an infection will spread ( $\mathcal{R}_0 > 1$ ) or die out ( $\mathcal{R}_0 < 1$ ). The SIR model’s ability to incorporate core epidemiological mechanisms in a mathematically tractable form makes it especially valuable for analyzing and forecasting epidemic progression. However, traditional compartmental models like SIR suffer from two major limitations. First, they rely on strong assumptions about the underlying infection dynamics. The SIR model, for example, simplifies disease progression into only two parameters: the transmission rate  $\beta$  and the recovery rate  $\gamma$ . While this simplification makes the model analytically tractable, it limits its flexibility and generalizability across different infectious diseases, especially those with more complex transmission mechanisms, latent periods, reinfections, or varying recovery trajectories. Such rigid parameterization can fail to capture the heterogeneity present in real-world epidemics. Second, the SIR model is not inherently data-driven. It typically requires predefined parameters that are either estimated from limited historical data or assumed based on prior knowledge. In contrast, modern data-driven approaches—such as machine learning models or neural networks—can automatically learn complex patterns directly from large-scale epidemic or clinical data. These models not only tend to achieve better predictive performance but also offer greater adaptability to diverse and dynamic epidemiological contexts. Moreover, recent advances in interpretable AI have enabled some data-driven models to provide insights comparable to, or even more nuanced than, traditional mechanistic frameworks.

## 2.2 Graphs for Epidemic Modeling

Recent advancements in dynamic graph modeling have underscored the utility of such models in epidemic source detection and spread prediction. Initially developed for traffic forecasting, dynamic graph models have been rapidly adapted to epidemiological contexts, where nodes represent geographical locations.[22, 10, 29] Furthermore, Lokhov et al. introduced a dynamic message-passing (DMP) inference algorithm tailored for the SIR (Susceptible-Infected-Recovered) model to estimate the origin of an epidemic outbreak.[23] This algorithm iteratively transmits messages along network edges, updating each node’s state probabilities based on the states of its neighbors. However, the algorithm operates on static graphs, thereby overlooking the inherent dynamics of contact networks in human societies.

Furthermore, due to the lack of publicly available data, synthetic graph construction allows epidemic models to be trained on data that closely approximates real-world conditions. The *Erdős-Rényi (ER) model* [12] generates a graph by connecting each pair of  $n$  nodes with a fixed probability  $p$ . The *Barabási-Albert (BA) model* [4] constructs scale-free networks where new nodes are more likely to connect to highly connected existing nodes. The *Stochastic Block Model (SBM)* [17] divides nodes into blocks (or communities), effectively capturing the community structures frequently present in real-world networks. More recently, to mimic real-life scenarios where people do not have equal probabilities to visit any place in a city, but rather that they will focus a small number of places, Higham etc. [16] proposes a *Gilbert Graph Model* that generates new graphs based on an established random graph. Each model offers unique features that can be tailored to the specific properties of the network being studied, ensuring that epidemic models are trained on data that closely resembles real-world conditions.

## 2.3 Hypergraphs for Epidemic Modeling

Similar to graphs, hypergraphs can also be utilized in the aggregation stage of SIR models [16, 9, 33]. In the pathogen propagation function proposed by Hypergraph-HeterSIS [3], the infection state of each node is first aggregated to hyperedge, which is then followed by a nonlinear function  $f$  to remove linearity. The result is then mapped back to node level to provide the next step update. The method has been shown that hypergraph-based approaches are better at capturing the structural differences in contact networks and improving the accuracy of infection dynamics modeling. However, these approaches are based on variable calibration, therefore neglecting the higher-level representation generated through deep learning approaches [13, 6, 2].

# Chapter 3

## Formulation

### 3.1 Hypergraph

A hypergraph is a higher-order representation of a graph where an edge can connect any number of vertices. Formally, a hypergraph  $G = (V, \mathcal{E}, \mathbf{X})$  consists of a set of nodes  $V$ , a set of hyperedges  $\mathcal{E}$ , where each hyperedge is a subset of  $V$ , and a feature matrix  $\mathbf{X} \in \mathbb{R}^{|V| \times d}$ , where each row encodes the node feature. The hypergraph structure can be described by an incidence matrix  $\mathbf{H} \in \mathbb{R}^{|\mathcal{E}| \times N}$ , where  $\mathbf{H}_{i,j} = 1$  only when the node  $v_i$  is incident to the edge  $e_j$ .

### 3.2 Dynamic Hypergraph

A dynamic hypergraph is an extension of a hypergraph that evolves over time, consisting of a sequence of hypergraphs observed over  $T$  discrete time stamps. Formally, a dynamic hypergraph is represented as  $G^{(0:T)} = \{G^{(0)}, G^{(1)}, \dots, G^{(T)}\}$ , where each hypergraph  $G^{(t)} = (V^{(t)}, \mathcal{E}^{(t)}, \mathbf{X}^{(t)})$  denotes the hypergraph at time stamp  $t \in [0 : T]$ . Here,  $V^{(t)}$  is the set of nodes,  $\mathcal{E}^{(t)}$  is the set of hyperedges, and  $\mathbf{X}^{(t)}$  denotes the node features at time  $t$ . It is worth noting that **both** the graph structure and node features are dynamic, since in some works, dynamic graphs have static features.

### 3.3 Epidemic Modeling Tasks

Consider an input of a dynamic hypergraph  $G^{(0:T)} = \{G^{(0)}, \dots, G^{(T)}\}$ , where each node represents an individual and each hyperedge represents a location. At an arbitrary time stamp  $t$ , the nodes in a hyperedge  $e^{(t)} \in \mathcal{E}^{(t)}$  represent a single contact between these entities. Each hypergraph  $G^{(t)}$  is associated with an individual state matrix  $\mathbf{X}^{(t)} \in \mathbb{R}^{N \times d}$ , where  $d$  is the feature dimension of the individual. For example, in the SIR setting,  $d$  can consist of three dimensions, which correspond to the {Suspected, Infected, Recovered} status of a specific individual.

Additionally, we define three time stamps to clarify the time interval of our downstream epidemic tasks.  $[0 : \text{tsh}]$  where tsh stands for Time Stamp Hidden;  $[\text{tsh} : \text{ks}]$  where ks stands for Known Time Stamp;  $[\text{ks} : \text{ps}]$  where ps stands for Prediction Time Stamp. The three time stamps are ordered such that  $0 \leq \text{tsh} \leq \text{ks} \leq \text{ps} \leq T$ . Note that for a time stamp  $t \in T$ , when  $t < \text{tsh}$ , only contact hypergraph can be observed. When  $\text{tsh} \leq t \leq \text{ks}$ , both contact hypergraph and individual state can be observed. When  $t > \text{ps}$ , neither contact hypergraph nor individual state can be observed. An illustration of the three time stamps is shown in Figure 4.1.

This setup aligns closely with real-life epidemic monitoring and response workflows. In practice, detailed individual-level data such as infection states often become available only after some delay due to testing and reporting lags, which corresponds to the interval from tsh to ks where such information is accessible. Earlier periods ( $t < \text{tsh}$ ) typically rely on structural data like contact patterns from mobile sensing or location tracking, while individual health statuses remain unobserved. As the outbreak unfolds beyond ps, data collection often lags behind real-time events due to limitations in surveillance infrastructure or data privacy concerns, making future observations inaccessible for immediate analysis. This temporal partitioning, therefore, reflects the partial and delayed nature of real-world epidemic data, making it a practical framework for modeling and forecasting tasks.

### 3.3.1 Source Detection

The source detection task focuses on identifying the initial node responsible for the spread of an epidemic, often referred to as "patient zero." Given the dynamic hypergraph  $G^{(0:ks)}$ , or its incidence matrix  $\mathbf{H}^{(0:ks)}$ , and the corresponding state matrix  $\mathbf{X}^{(tsh:ks)}$ , we aim to infer the likelihood distribution over all nodes at the initial time stamp  $T = 0$  using a model  $f$  parametrized by weight  $\theta$ . Mathematically, we are interested in using  $f_\theta$  to estimate the distribution:

$$f_\theta(\mathbf{H}^{(0:ks)}, \mathbf{X}^{(tsh:ks)}) \approx p(\mathbf{X}^{(0)} | \mathbf{H}^{(0:ks)}, \mathbf{X}^{(tsh:ks)}).$$

This task leverages both the structural properties of the hypergraph and the temporal evolution of the feature maps to backtrack the probable origin of the epidemic. The node labels  $y_{\text{detect}}$  are extracted from specific columns of  $\mathbf{X}^{(0)}$  that represent the infection state — for example, **the "infected" column** in the case of the SIR model. To optimize the model, we'll use weighted binary cross-entropy loss between the predictions and node labels  $y_{\text{detect}}$ , where  $w_1 = \frac{|V|}{|y_{\text{detect}}=1|}$  and  $w_0 = \frac{|V|}{|y_{\text{detect}}=0|}$

$$\mathcal{L}_{\text{detect}}(\theta) = -\frac{1}{|V|} \sum_{v \in V} [w_1 y_{\text{detect}} \log(f_\theta) + w_0 (1 - y_{\text{detect}}) \log(1 - f_\theta)],$$

### 3.3.2 Infection Forecasting

Forecasting tasks in epidemics are usually defined as finding the total number of infections and recoveries in a range of future time stamps. This is because previous approaches encode nodes as areas, neglecting the individual level information. On the other hand, when using human contact hypergraphs, we can deduce a more fine-grained forecasting on an individual level. Therefore, we treat our forecast task as a binary node classification task, where we are interested in using a model  $g$  parametrized by  $\theta$

to estimate the distribution:

$$g_{\theta}(\mathbf{H}^{(0:ks)}, \mathbf{X}^{(tsh:ks)}) \approx p(\mathbf{X}^{(ks+1:ps)} | \mathbf{H}^{(0:ks)}, \mathbf{X}^{(tsh:ks)})$$

Similar to source detection, we'll use the binary cross-entropy loss between the predictions and node labels  $y_{\text{forecast}}$ . The labels are extracted from specific columns of  $\mathbf{X}^{(ks+1:ps)}$  that represent the infection state, similar to source detection label extraction.

$$\mathcal{L}_{\text{forecast}}(\theta) = -\frac{1}{|V|} \sum_{v \in V} [y_{\text{forecast}} \log(g_{\theta}) + (1 - y_{\text{forecast}}) \log(1 - g_{\theta})],$$

# Chapter 4

## Method

In this section, we will formulate our proposed model EpiDHGNN, which serves as  $f_\theta$  and  $g_\theta$  defined in section 3.3. Here we define  $t \in T_{\text{interest}}$  where  $T_{\text{interest}}$  is the corresponding input interval for  $\mathbf{H}$  defined in section 3.3.1 and 3.3.2.

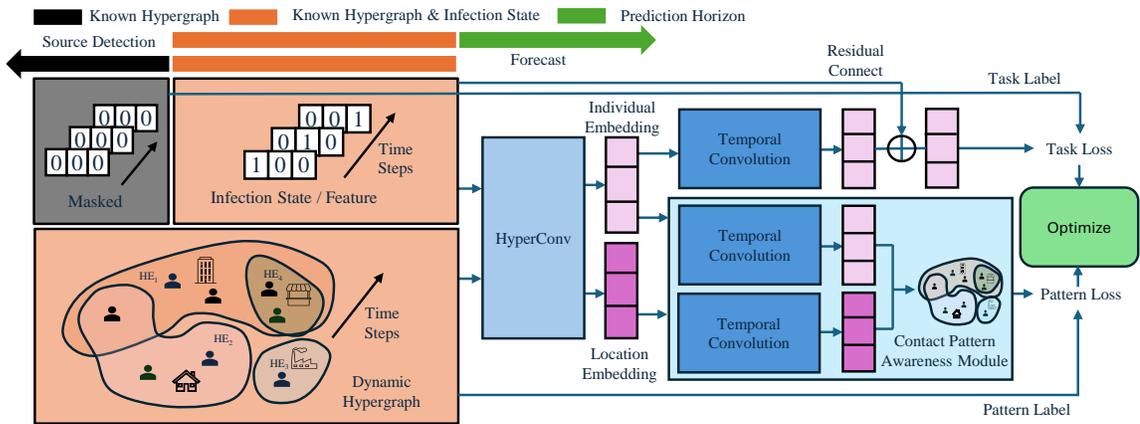


Figure 4.1: Model Architecture of proposed EpiDHGNN model. The arrows in the top left corner refers to the three time stamps defined in section 3.3, where  $[0 : tsh]$  is the black interval,  $[tsh : ks]$  is the orange interval, and  $[ks + 1 : ps]$  is the green interval. All individual state is masked to 0 in  $[0 : tsh]$  as shown in the top left black module. Corresponding inputs for source detection and forecast defined in section 3.3.1 and 3.3.2 is then feed to the model as input. The light blue HyperConv module in defined in section 4.1; the dark blue temporal convolution module is defined in 4.2; and the contact pattern awareness module is defined in section 4.3.

## 4.1 HGNN Module

Hypergraph Neural Network [13] (HGNN) is designed to map the original node features to a more refined feature space that captures higher-order neighbor information. Different from traditional GNN, HGNN is performing node-edge-node convolution. Specifically, one convolution layer is defined as:

$$\mathbf{X}_{\text{edge}}^{l+1,t} = \mathbf{D}_e^{-1} \mathbf{H}^t \mathbf{D}_v^{-1/2} \mathbf{X}_{\text{node}}^{l,t} \Theta \quad (4.1)$$

$$\mathbf{X}_{\text{node}}^{l+1,t} = \mathbf{D}_v^{-1/2} \mathbf{H}^t \mathbf{W} \mathbf{D}_e^{-1} \mathbf{H}^t \mathbf{D}_v^{-1/2} \mathbf{X}_{\text{node}}^{l,t} \Theta \quad (4.2)$$

where  $\mathbf{X}_{\text{node}}^{0,t}$  is the individual state matrix,  $\mathbf{X}^{l,t}$  is the feature matrix after  $l^{\text{th}}$  convolution,  $\Theta$  are the learnable parameters,  $\mathbf{H}^t$  is the incidence matrix,  $D$  are degree matrices for normalization, and  $W$  is the optional diagonal hyperedge weight matrix at time stamp  $t \in$  the input interval of specific task. Through HGNN, the embeddings used for final outputs not only contain higher-order neighbor information, but also location information. Stacking  $L$  layers of convolution layers will allow aggregation to capture  $L$  distance away neighbors. After the convolution has been applied to all time steps, we will concatenate all features along the temporal dimension:

$$\mathcal{X}_{\text{node}}^L = [\mathbf{X}_{\text{node}}^{L,t_0} | \mathbf{X}_{\text{node}}^{L,t_1} | \dots], \quad \mathcal{X}_{\text{edge}}^L = [\mathbf{X}_{\text{edge}}^{L,t_0} | \mathbf{X}_{\text{edge}}^{L,t_1} | \dots], \quad \forall t \in T_{\text{interest}} \quad (4.3)$$

## 4.2 Temporal Convolution Module

The temporal convolution, proposed by Lea etc. [19] and applied in dynamic graphs by Guo etc. [14] It operates by performing standard convolution on the temporal dimension. In analogy with image convolution, where spatial information is processed across width and height dimensions, the temporal convolution in our framework treats time as the width dimension, individuals as the length dimension, and the feature

channels obtained from the hypergraph convolution layer as the depth (or channels). Formally, the temporal convolution operation is defined as:

$$\hat{\mathbf{X}}_{\text{node}} = \sigma \left( \Phi_{\text{temporal}}^k \otimes \mathcal{X}_{\text{node}}^L \right) \quad (4.4)$$

where  $\otimes$  denotes the convolution operation,  $\Phi_{\text{temporal}}$  is the convolution kernel with size  $k$  refining the learned representations across time steps,  $C$  is the features' dimension,  $N$  is the number of individuals, and  $T$  is the number of time steps, and  $\sigma$  is the non-linear activation function. This temporal convolution module enables the model to efficiently capture multi-channel temporal dependencies in epidemic progression by considering the dynamic evolution of infection states. As a result, the model progressively learns representations of disease spread patterns.

### 4.3 Contact Pattern Awareness Module

In a societal setting, human interactions occur with varying probabilities based on social structures and daily routines. For instance, individuals are highly likely to engage in frequent interactions with family members or colleagues at home or in the workplace, while social encounters with friends or individuals sharing similar interests may occur less frequently, such as on a weekly or monthly basis in clubs or shopping centers. Given this structured nature of human interactions, incorporating contact pattern prediction can enhance the accuracy of epidemic modeling by providing insights into both predictions and back-tracing.

To leverage this information, we propose a self-supervised *Contact Pattern Awareness Module*, designed to predict human interactions within the epidemic framework. Given a sequence of  $k$  hypergraphs  $\{G_t\}_{t=t_0}^{t_0+k-1}$  starting from a randomly selected time step  $t_0$ , the module aims to reconstruct the hypergraph at the final time step,  $G_{t_0+k}$ , using information from the preceding  $k - 1$  hypergraphs. Successful reconstruction

contributes additional structural knowledge about human contact patterns to the learned node embeddings, thereby improving performance in subsequent epidemic forecasting tasks.

To effectively capture these patterns, we utilize both the individual embeddings and location embeddings obtained from Section 4.1. These embeddings are then processed using the temporal convolution framework introduced in Section 4.2, enabling the model to extract temporal dependencies. Finally, the refined embeddings are passed through a fully connected layer to produce a confidence score for contact prediction. Mathematically, this operation is formulated as:

$$\mathbf{s} = \sigma(MLP(\Phi_{\text{pattern}}^k \otimes (\mathcal{X}_{\text{node}}^L) * \Phi_{\text{pattern}}^k \otimes (\mathcal{X}_{\text{edge}}^L))) \quad (4.5)$$

where  $\mathbf{s}$  is the output confidence score of the individual-location contact,  $*$  represents element-wise multiplication,  $MLP$  stands for multilayer perceptron, and  $\sigma$  represents the sigmoid activation function. To ensure class balance, the module will be optimized based on binary cross-entropy loss between positive contacts and randomly selected negative contacts of equal size:

$$\mathcal{L}_{\text{pattern}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log s_i + (1 - y_i) \log(1 - s_i)], \quad (4.6)$$

where  $y_i \in \{0, 1\}$  is the ground-truth label indicating whether a contact exists, and  $s_i$  is the predicted confidence score.

## 4.4 EpiDHGNN

The proposed EpiDHGNN follows an encoder-decoder paradigm, designed to model epidemic dynamics by integrating spatial, temporal, and contact pattern information. As illustrated in Figure 4.1, EpiDHGNN effectively captures the evolution of infection

states through hypergraph structure and temporal dependencies.

Given an input sequence of dynamic hypergraphs  $\mathbf{H}^{0:ks}$  and the corresponding patient state features  $\mathbf{X}^{tsh:ks}$ , the model first encodes node representations using hypergraph convolution (Section 4.1) to extract higher-order spatial dependencies. These embeddings are then refined through temporal convolution (Section 4.2) to capture progression patterns over time. After the temporal convolution, we added a layer of residual connection of the initial individual state without the spatial processing. The encoded embeddings are then used for task-specific prediction, such as epidemic forecasting or source detection, with the loss function defined accordingly in Section 3.3.

In addition to task-specific learning, EpiDHGNN incorporates the self-supervised *Contact Pattern Awareness Module* (Section 4.3) to enhance embedding quality by reconstructing human interaction patterns. This module optimizes the pattern loss  $\mathcal{L}_{\text{pattern}}$ , enforcing structural consistency in the learned representations. To balance predictive accuracy and structural awareness, we introduce a weighting hyperparameter  $\alpha$  that controls the trade-off between the task-specific loss and the pattern loss. The final objective function is:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{task}} + (1 - \alpha) \mathcal{L}_{\text{pattern}}. \quad (4.7)$$

By jointly optimizing for both epidemic prediction and contact pattern consistency, EpiDHGNN ensures robust and generalizable representations, leading to more reliable epidemic modeling. This synergy between hypergraph structure, temporal dependencies, and self-supervised contact prediction enables EpiDHGNN to outperform traditional GNN-based epidemic models, particularly in capturing complex and dynamic transmission pathways.

# Chapter 5

## Experiments

In this section, we perform analysis on the datasets and conduct experiments to evaluate the proposed model. We will focus on the following research questions:

- **RQ1:** Does EpiDHGNN outperforms baseline dynamic graph models in various epidemic tasks?
- **RQ2:** Does the contact pattern awareness module facilitate the overall performance of EpiDHGNN?
- **RQ3:** Is contact patterns successfully captured? To what aspect of the task does the module helps the most?
- **RQ4:** Beyond individual-level prediction, can EpiDHGNN capture population-level infection dynamics over time?

### 5.1 Data Description

We assess the performance of baseline models and our proposed model on both graphs and hypergraphs settings. Because of the privacy nature of human contact data, we used both real-world and synthetic data. The University of Virginia **UVA** dataset

includes an extensive collection of clinical metadata sourced from the Epic-based SQL database at the UVA hospital. The interactions are derived from Electronic Healthcare Records (EHRs), which document the timing and locations of encounters between patients and healthcare workers (HCWs). We utilized the real-world infection case calibrated pathogen parameters provided by Anand et al. [3] to retrieve the patient infection states through simulations.

Table 5.1: Dataset Summary

<b>Metric</b>	<b>UVA</b>	<b>EpiSim</b>
# Individuals	2,500	10,000
# Locations	500	11
# Time Steps	169	47
# Contacts	94,134	664,177

The **EpiSim** dataset is derived on the Mobility Intervention of Epidemic’s Simulator[1]. The simulator is composed of two components. The *Human Mobility Model* simulates individual movement from 8 A.M. to 10 P.M., with each simulation step representing one hour. On weekdays, an individual moves from a residential area to a working area at time  $T_d \sim U(a, b)$  and stays there for  $T_w \sim U(c, d)$  hours. After work, they may visit a nearby commercial area before returning home. On weekends, individuals may visit a commercial area at  $T_e \sim U(g, h)$  with probability  $P_e$ , staying for  $T_m \sim U(i, j)$  hours before returning home. The *Disease Transmission Model* considers two types of contacts: *acquaintance* and *stranger* contacts. Each individual has a fixed number of acquaintance contacts in both their residential and working areas, with the sizes drawn from uniform distributions  $K_r \sim U(m, n)$  and  $K_w \sim U(o, p)$ . At each timestep, there is a probability  $P_a$  of infection from an infected acquaintance. Additionally, individuals encounter strangers in the same location, with a probability  $P_s$  of infection per timestep. The parameters of the simulator have been calibrated with the Covid-19  $R_0$  provided by WHO. The statistics for both datasets are shown in Table 5.1.

## 5.2 Setup

In our experiment, we utilize a 2-layer HGNN to capture neighborhood information. We perform a grid search over key hyperparameters, including hidden dimensions, learning rate, weight decay, kernel size, and  $\alpha$ . During training, we employ the ADAM optimizer with weight decay and gradient clipping activated to stabilize gradient updates and prevent exploding gradients. Models are trained for up to 100 epochs, with early stopping activated if the validation loss does not improve for 10 consecutive epochs. The experiments are conducted on a single NVIDIA Tesla V100 GPU with 16 GB of memory. Training time per epoch averages around 5 seconds. To enhance reproducibility, random seeds are fixed for data splitting, model initialization, and optimization processes.

## 5.3 RQ1 - Performance

Table 5.2: Experiment Result of Source Detection Task on UVA Dataset. Best performance under each setting is bolded and equal performance is underlined.

	TSH	MRR	Hit@1	Hit@3
STGCN	5	0.491 $\pm$ 0.056	0.300 $\pm$ 0.036	0.650 $\pm$ 0.057
	10	0.462 $\pm$ 0.064	0.315 $\pm$ 0.074	0.633 $\pm$ 0.024
	20	0.427 $\pm$ 0.033	0.175 $\pm$ 0.023	0.596 $\pm$ 0.078
ASTGCN	5	0.501 $\pm$ 0.026	0.300 $\pm$ 0.078	0.650 $\pm$ 0.052
	10	0.486 $\pm$ 0.046	0.250 $\pm$ 0.082	0.667 $\pm$ 0.022
	20	0.416 $\pm$ 0.029	0.205 $\pm$ 0.058	0.650 $\pm$ 0.032
MSTGCN	5	0.618 $\pm$ 0.026	0.417 $\pm$ 0.029	0.767 $\pm$ 0.076
	10	0.561 $\pm$ 0.026	0.350 $\pm$ 0.050	0.733 $\pm$ 0.058
	20	0.442 $\pm$ 0.029	0.150 $\pm$ 0.058	0.700 $\pm$ 0.052
<b>EpiDHGNN</b>	5	<b>0.704 <math>\pm</math> 0.033</b>	<b>0.517 <math>\pm</math> 0.076</b>	<b>0.917 <math>\pm</math> 0.029</b>
	10	<b>0.662 <math>\pm</math> 0.005</b>	<b>0.500 <math>\pm</math> 0.000</b>	<b>0.783 <math>\pm</math> 0.029</b>
	20	<b>0.582 <math>\pm</math> 0.031</b>	<b>0.350 <math>\pm</math> 0.000</b>	<b>0.765 <math>\pm</math> 0.050</b>

Our experimental results for both source detection and forecasting are presented in Table 5.2, Table 5.3, and Table 5.4, respectively, with the best performance under

Table 5.3: Experiment Result of Source Detection Task on EpiSim Dataset. Best performance under each setting is bolded and equal performance is underlined.

	TSH	MRR	Hit@1	Hit@3
STGCN	5	0.242 ± 0.075	0.145 ± 0.052	0.282 ± 0.035
	10	0.129 ± 0.039	0.100 ± 0.103	0.195 ± 0.052
	20	0.111 ± 0.047	0.089 ± 0.000	0.163 ± 0.017
ASTGCN	5	0.226 ± 0.036	0.167 ± 0.033	0.333 ± 0.042
	10	0.141 ± 0.067	0.100 ± 0.027	0.133 ± 0.014
	20	0.118 ± 0.087	0.076 ± 0.031	0.100 ± 0.087
MSTGCN	5	0.333 ± 0.029	0.167 ± 0.000	0.400 ± 0.058
	10	0.213 ± 0.058	0.100 ± 0.026	<b>0.200 ± 0.000</b>
	20	0.192 ± 0.016	0.089 ± 0.100	0.193 ± 0.029
EpiDHGNN	5	<b>0.401 ± 0.074</b>	<b>0.200 ± 0.100</b>	<b>0.500 ± 0.100</b>
	10	<b>0.218 ± 0.037</b>	<b>0.133 ± 0.058</b>	0.167 ± 0.058
	20	<b>0.219 ± 0.061</b>	<b>0.100 ± 0.100</b>	<b>0.200 ± 0.100</b>

Table 5.4: Experiment Result of Forecast Task. Best Performance under each setting is bolded.

	PS	UVA		EpiSim	
		F1	AUROC	F1	AUROC
STGCN	5	0.526 ± 0.022	0.714 ± 0.035	0.632 ± 0.036	0.816 ± 0.029
	10	0.343 ± 0.028	0.688 ± 0.010	0.473 ± 0.025	0.692 ± 0.043
	20	0.398 ± 0.031	0.655 ± 0.031	0.195 ± 0.073	0.593 ± 0.010
ASTGCN	5	0.544 ± 0.038	0.731 ± 0.060	0.624 ± 0.030	0.801 ± 0.005
	10	0.376 ± 0.013	0.692 ± 0.012	0.489 ± 0.014	0.712 ± 0.009
	20	0.367 ± 0.009	0.652 ± 0.011	0.154 ± 0.045	0.612 ± 0.025
MSTGCN	5	<b>0.721 ± 0.063</b>	<b>0.846 ± 0.013</b>	0.869 ± 0.084	0.895 ± 0.010
	10	0.401 ± 0.041	0.647 ± 0.012	0.502 ± 0.042	0.729 ± 0.020
	20	0.358 ± 0.024	0.617 ± 0.065	0.223 ± 0.035	0.658 ± 0.056
EpiDHGNN	5	0.712 ± 0.023	0.837 ± 0.019	<b>0.918 ± 0.042</b>	<b>0.957 ± 0.065</b>
	10	<b>0.576 ± 0.012</b>	<b>0.750 ± 0.008</b>	<b>0.612 ± 0.001</b>	<b>0.874 ± 0.017</b>
	20	<b>0.454 ± 0.007</b>	<b>0.685 ± 0.008</b>	<b>0.298 ± 0.080</b>	<b>0.779 ± 0.071</b>

each setting highlighted in bold. We evaluated the models under diverse conditions to assess their robustness. For source detection, we masked timesteps of varying lengths (5, 10, and 20) to examine the models’ ability to backtrack across different scenarios. Similarly, we tested forecasting performance using prediction horizons of 5, 10, and 20 timesteps. In most settings, EpiDHGNN outperforms the majority of baseline

graph-based models, underscoring the advantages of hypergraph-based approaches in epidemic modeling through capturing the high-order contact interaction.

## 5.4 RQ2 - Ablation

We conducted an ablation study on the contact pattern awareness module to investigate Question 2. As shown in Table 5.5, removing this module led to a noticeable decline in performance, indicating its crucial role in capturing individual contact patterns. The results suggest that incorporating individual contact behaviors enhances the model’s ability to encode social interactions more effectively, aligning with societal norms. This highlights the importance of modeling personalized contact dynamics in improving the overall predictive capability of our approach.

Table 5.5: Ablation study on contact pattern awareness module.

		UVA			EpiSim	
		Setting	MRR	Hit@1	MRR	Hit@1
Detection	w/o CT module	tsh-5	0.692 ± 0.050	0.483 ± 0.104	0.381 ± 0.027	0.167 ± 0.058
		tsh-10	0.644 ± 0.032	0.467 ± 0.058	0.204 ± 0.020	0.100 ± 0.000
		tsh-20	0.558 ± 0.014	0.323 ± 0.029	0.197 ± 0.006	0.100 ± 0.100
	w/ CT module	tsh-5	<b>0.704 ± 0.033</b>	<b>0.517 ± 0.076</b>	<b>0.401 ± 0.074</b>	<b>0.200 ± 0.100</b>
		tsh-10	<b>0.662 ± 0.005</b>	<b>0.500 ± 0.005</b>	<b>0.218 ± 0.037</b>	<b>0.133 ± 0.058</b>
		tsh-20	<b>0.582 ± 0.031</b>	<b>0.350 ± 0.058</b>	<b>0.219 ± 0.061</b>	<u>0.100 ± 0.092</u>
Forecast	w/o CT module		F1	AUROC	F1	AUROC
		ps-5	0.709 ± 0.004	0.830 ± 0.003	0.891 ± 0.003	0.912 ± 0.007
		ps-10	0.571 ± 0.008	0.747 ± 0.006	0.513 ± 0.007	0.824 ± 0.005
	w/ CT module	ps-5	<b>0.712 ± 0.023</b>	<b>0.837 ± 0.019</b>	<b>0.918 ± 0.042</b>	<b>0.957 ± 0.006</b>
		ps-10	<b>0.576 ± 0.012</b>	<b>0.750 ± 0.008</b>	<b>0.612 ± 0.001</b>	<b>0.874 ± 0.004</b>
		ps-20	<b>0.454 ± 0.007</b>	<b>0.685 ± 0.008</b>	<b>0.298 ± 0.080</b>	<b>0.779 ± 0.071</b>

## 5.5 RQ3 - Module Effectiveness

To investigate whether the contact pattern is successfully captured, we evaluate the module’s performance in predicting contact existence at the location level. Specifically, locations are divided into four quantiles based on their contact intensity. For example, in the EpiSim dataset, households exhibit lower contact intensity compared to recreational locations. For each quantile, we report the prediction accuracy along with the overall accuracy in Table 5.6. The results suggest that the overall contact pattern is successfully reconstructed. While the UVA dataset shows little correlation between contact intensity and accuracy, the EpiSim dataset exhibits a strong negative correlation. This observation aligns with the underlying assumptions of our dataset. The UVA dataset includes hospital contacts, which may fluctuate due to patient movement, whereas in the EpiSim dataset, locations with low contact intensity likely correspond to households, where visits occur with high frequency and regularity.

Table 5.6: Contact Pattern Prediction. The quantiles are ranked by the total number of interactions made in a location.

Quantile	UVA		EpiSim	
	Range	F1	Range	F1
1	[ :6]	$0.795 \pm 0.009$	[ :616]	$0.997 \pm 0.002$
2	[6:11]	$0.773 \pm 0.012$	[616:1788]	$0.852 \pm 0.004$
3	[11:19]	$0.809 \pm 0.006$	[1788:1847]	$0.82 \pm 0.002$
4	[19:]	$0.841 \pm 0.018$	[1847:]	$0.639 \pm 0.005$
Overall	—	$0.804 \pm 0.011$	—	$0.827 \pm 0.003$

We further investigate the influence of the hyperparameter  $\alpha$ , selecting values from 0.3, 0.5, 0.7, 0.9, 1.0, to assess its impact on overall model performance. Lower values of  $\alpha$  were not considered, as  $\alpha = 0.3$  already exhibited significantly diminished performance, failing to effectively capture the model’s main task. As shown in Figure 5.1, our results indicate that  $\alpha$  has little correlation with the final performance, suggesting that it can be treated as a tunable hyperparameter for future studies.

Additionally, we observe that the model with a high timestep hidden state (TSH20)

consistently outperforms other configurations when incorporating  $\alpha$ . This suggests that integrating contact pattern information is particularly beneficial for tasks requiring a longer temporal memory, as it helps the model better capture long-term dependencies in contact patterns. These findings highlight the importance of tuning  $\alpha$  based on specific task requirements while reinforcing the advantage of incorporating contact-aware representations for long-horizon forecasting.

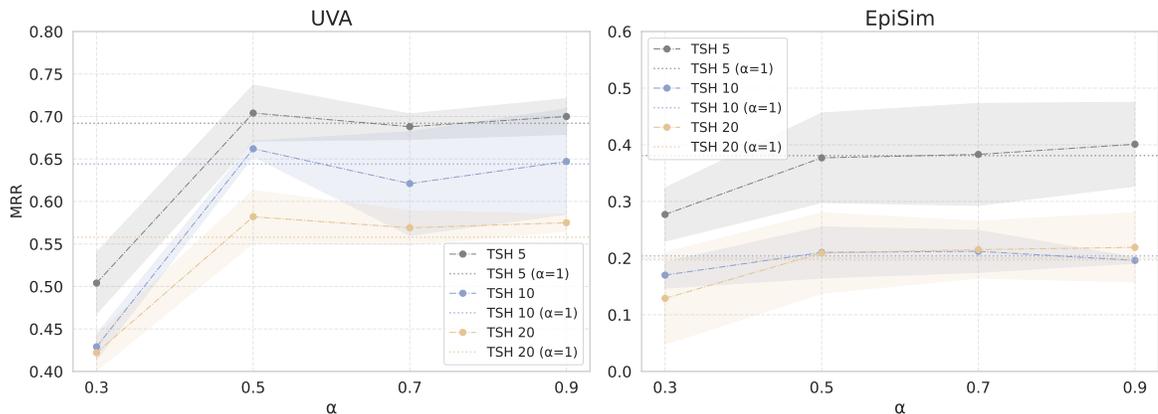


Figure 5.1: Visualization of Alpha’s impact on source detection performance

## 5.6 RQ4 - Generalizability

While we have demonstrated EpiDHGNN’s ability to forecast an individual’s probability of infection in Question 1, its effectiveness at the population level—a key focus of prior work—has yet to be established. To address this, we aggregated the daily sum of infected individuals to generate population-level data across various prediction horizons. As shown in Figure 5.2, EpiDHGNN accurately captures short-term infection dynamics and effectively tracks broader fluctuations at longer time steps, albeit with reduced precision.

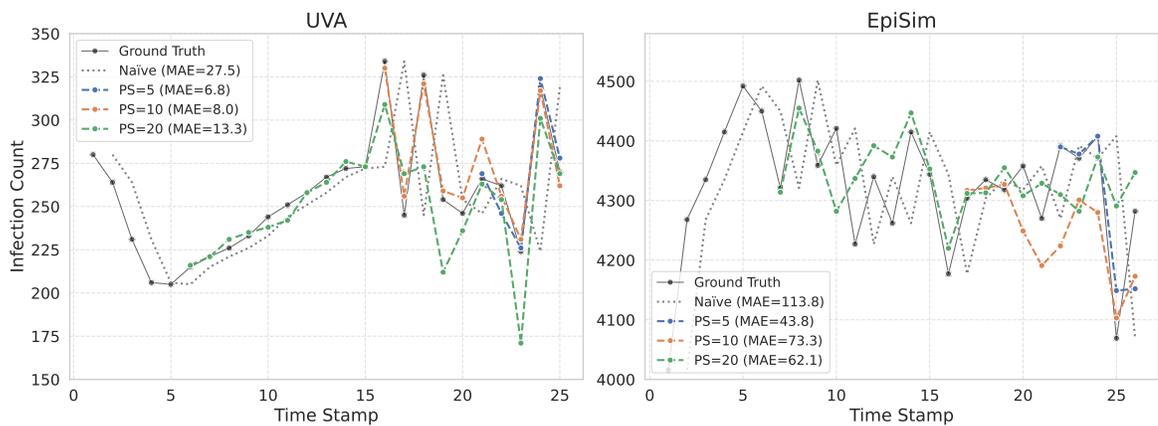


Figure 5.2: Forecast generalizability analysis. PS stands for prediction steps. The models can successfully the future infection dynamics within various PS. We also provide the Mean Absolute Error (MAE) of Naive Model (A naive time series model forecasts future values by assuming they will be the same as the most recent observed value or follow a simple pattern) and our approach.

# Chapter 6

## Conclusion

### 6.1 Future Work

#### 6.1.1 Hypergraph SIR

Although the hypergraph SIR simulation leverages higher-order node interactions, the key operation  $H^T\beta SH$  can inherently be approximated by a weighted graph. This implies that the simulation does not fully exploit the unique information encoded in hypergraphs that cannot be captured by graphs. Specifically, the linearity of the  $H^T H$  operation limits the model’s ability to represent complex higher-order interactions. In the work of H<sup>2</sup>ABM [5], a non-linear activation function  $f$  is introduced during training, effectively breaking the linear constraints of  $H^T H$  and enabling more expressive modeling of hypergraph structures. However, this approach is incompatible with our current simulation methodology due to the constraints of our SIR framework, which relies on the linear structure of the  $H^T\beta SH$  operation for efficient epidemic propagation modeling. A promising alternative could involve developing a hybrid simulation approach that integrates non-linear transformations with the existing linear operation.

### 6.1.2 Alternative Model Selection

The current EpiDHGNN’s temporal and spatial layers are implemented using HGNN and temporal convolution layers, respectively. However, the model’s structure is designed to allow flexible substitutions of these layers, providing opportunities to explore and integrate alternative architectures. Future work will investigate other promising candidates, such as the All-Set model[6], which has reported state-of-the-art performance in various hypergraph-based tasks, will be considered for the spatial layer. These substitutions aim to enhance the model’s capability and performance further. By experimenting with these advanced alternatives, future studies can seek to identify the most effective configurations for improving the accuracy and robustness of the THGNN in capturing complex temporal and spatial dependencies in epidemic modeling and source detection tasks.

### 6.1.3 Human Contact Data Simulation

It has been shown in the paper that human contact data plays a crucial role in research, particularly in fields such as epidemiology, where understanding interaction patterns is vital for accurate modeling and prediction. Due to the highly private nature of human contact data and its scarcity, developing algorithms that effectively capture realistic human contact patterns is essential. Such algorithms must balance complexity and privacy, ensuring simulations are representative while safeguarding individual identities. Additionally, since modeling human contact on a global scale is impractical, simulations over smaller, well-defined groups—such as universities, companies, or communities—are preferred. These simulations should incorporate factors like social clustering (e.g., friend zones), geographic considerations (e.g., different locations), and temporal dynamics to better mirror real-world interaction patterns. By generating accurate synthetic datasets that replicate these nuances, researchers can conduct experiments and validate models, advancing our understanding of human contact and

their broader implications.

## 6.2 Summary

In summary, this study introduced the EpiDHGNN framework, demonstrating its ability to effectively capture the complex dynamics of epidemic spread through higher-order interactions in human contact networks. By leveraging hypergraphs and deep learning, our model significantly improved performance in both epidemic forecasting and source detection tasks. Through rigorous experimentation on both real-world and synthetic datasets, the study validated the advantages of modeling human contact as a dynamic hypergraph, highlighting the importance of higher-order relationships in disease transmission modeling. By continuing to refine EpiDHGNN and improving data-driven epidemic simulations, the study aims to contribute to more effective infectious disease forecasting and mitigation strategies, ultimately informing public health interventions.

# Appendix A

## Appendix

### A.1 Hypergraph SIR Simulation

Due to the highly private nature of epidemic data, which is often not accessible for public research, existing works rely on simulated data generated by models such as SIR[26] and TimeGeo[22]. Inspired by the work of [3], hypergraphs serve as a more accurate approach to predict future infections in epidemic settings. Therefore, we will adapt the hypergraph further to a "discrete stochastic hypergraph SIR simulation". Specifically, for all nodes, we define a state matrix  $S \in \mathbb{Z}^{T*N*3}$  and a pathogen matrix  $P \in \mathbb{R}^{T*N*3}$ , where the three columns of both matrices respectively represent:

- Susceptible individuals and probability at time  $t$
- Infected individuals and probability at time  $t$
- Recovered individuals and probability at time  $t$

#### Pathogen Probability Updates

At each timestep  $t$ ,  $P_{t+1}$  is updated for each node based on the incidence matrix  $\mathbf{H}_t$  of the hypergraph  $G^{(t)}$  and the current state  $S_t$ . The new infections for node  $i$  at time

$t$  is calculated as:

$$\text{new\_infections}_{i,t+1} = \sum_{j=1}^N (\mathbf{H}_t^\top \beta \cdot S_{t,i,1} \mathbf{H}_t)_{ij}$$

which aggregates the contributions of infected nodes  $j$  to susceptible node  $i$  through the hypergraph structure.  $\beta \in \mathbf{R}^{|E|}$  is the infection hyperparameter assigned to each hyperedge at random.

The probability of new recoveries for node  $i$  is given by:

$$\text{new\_recoveries}_{i,t+1} = \gamma \cdot P_{i,t}$$

where  $\gamma$  is a real number hyperparameter defining the recovery rate of the virus. The state probabilities for each node are updated as follows:

$$P_{t,i,0} = P_{t-1,i,0} - \text{new\_cases}_{i,t} \tag{A.1}$$

$$P_{t,i,1} = P_{t-1,i,1} + \text{new\_cases}_{i,t} - \text{new\_recoveries}_{i,t} \tag{A.2}$$

$$P_{t,i,2} = P_{t-1,i,2} + \text{new\_recoveries}_{i,t} \tag{A.3}$$

The discrete state matrix  $S$  is updated stochastically based on the pathogen matrix  $P$

**Susceptible Nodes** If node  $i$  is susceptible at time  $t - 1$ :

$$P(\text{Node } i \text{ becomes infected at } t) = I_i(t)$$

The state update is:

$$S_i(t) = \begin{cases} [0, 1, 0], & \text{if } U < P_{i,1}(t) \\ [1, 0, 0], & \text{otherwise} \end{cases}$$

where  $U \sim \mathcal{U}(0, 1)$  is a uniform random variable.

**Infected Nodes** If node  $i$  is infected at time  $t - 1$ :

$$P(\text{Node } i \text{ recovers at } t) = P_{i,2}(t) \cdot (1 - \delta)$$

$$P(\text{Node } i \text{ becomes susceptible at } t) = P_{i,2}(t) \cdot \delta$$

where  $\delta$  is the probability of re-infection. The state update depends on two independent random variables  $U, V \sim \mathcal{U}(0, 1)$ :

- If  $U < P_{i,2}(t)$ :
  - If  $V < 1 - \delta$ ,  $S_i(t) = [0, 0, 1]$  (recovered).
  - Else,  $S_i(t) = [1, 0, 0]$  (susceptible).
- Else,  $S_i(t) = [0, 1, 0]$  (remains infected).

**Recovered Nodes** If node  $i$  is recovered at time  $t - 1$ :

$$S_i(t) = [0, 0, 1]$$

The "discrete stochastic hypergraph SIR simulation" process captures the probabilistic nature of disease transmission and recovery while leveraging the hypergraph structure for higher-order interactions, providing a realistic simulation of epidemic dynamics.

# Bibliography

- [1] Mobility intervention of epidemics, 2020. URL <https://hzw77-demo.readthedocs.io/en/latest/>. Accessed: 2025-02-13.
- [2] Shivam Agarwal, Ramit Sawhney, Megh Thakkar, Preslav Nakov, Jiawei Han, and Tyler Derr. THINK: Temporal hypergraph hyperbolic network. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 849–854. IEEE, 2022. doi: 10.1109/ICDM54844.2022.00096.
- [3] Vivek Anand, Jiaming Cui, Jack Heavey, Anil Vullikanti, and B. Aditya Prakash. *Hj sup 2j / sup ABM: Heterogeneous Agent-based Model on Hypergraphs to Capture Group Interactions*, pages 280–288. 2024. doi: 10.1137/1.9781611978032.32. URL <https://epubs.siam.org/doi/abs/10.1137/1.9781611978032.32>.
- [4] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [5] Qi Cao, Renhe Jiang, Chuang Yang, Zipei Fan, Xuan Song, and Ryosuke Shibasaki. Mepognn: Metapopulation epidemic forecasting with graph neural networks. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 453–468, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-26422-1.

- [6] Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are allset: A multiset function framework for hypergraph neural networks, 2022.
- [7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.
- [8] J. Christopher Clement, VijayaKumar Ponnusamy, K.C. Sriharipriya, and R. Nandakumar. A survey on mathematical, machine learning and deep learning models for covid-19 transmission and diagnosis. *IEEE Reviews in Biomedical Engineering*, 15:325–340, 2022. doi: 10.1109/RBME.2021.3069213.
- [9] Guilherme Ferraz de Arruda, Giovanni Petri, and Yamir Moreno. Social contagion models on hypergraphs. *Phys. Rev. Res.*, 2:023032, Apr 2020. doi: 10.1103/PhysRevResearch.2.023032. URL <https://link.aps.org/doi/10.1103/PhysRevResearch.2.023032>.
- [10] Songgaojun Deng, Shusen Wang, Huzefa Rangwala, Lijing Wang, and Yue Ning. Graph message passing with cross-location attentions for long-term ili prediction, 2019. URL <https://arxiv.org/abs/1912.10202>.
- [11] Lucy D’Agostino McGowan, Kyra H Grantz, and Eleanor Murray. Quantifying uncertainty in mechanistic models of infectious disease. *American Journal of Epidemiology*, 190(7):1377–1385, January 2021. ISSN 1476-6256. doi: 10.1093/aje/kwab013. URL <http://dx.doi.org/10.1093/aje/kwab013>.
- [12] Paul Erdős and Alfréd Rényi. On random graphs i. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.
- [13] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks, 2019.

- [14] Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 922–929, 2019.
- [15] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.
- [16] Desmond John Higham and Henry Louis De Kergorlay. Disease extinction for susceptible-infected-susceptible models on dynamic graphs and hypergraphs. *Chaos*, 32, 8 2022. ISSN 10897682. doi: 10.1063/5.0093776.
- [17] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.
- [18] Sunwoo Kim, Soo Yong Lee, Yue Gao, Alessia Antelmi, Mirko Polato, and Kijung Shin. A survey on hypergraph neural networks: An in-depth and step-by-step guide. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6534–6544, 2024.
- [19] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation, 2016. URL <https://arxiv.org/abs/1608.08242>.
- [20] Justin Lessler and Derek A. Cummings. Mechanistic models of infectious disease and their impact on public health. *American Journal of Epidemiology*, 183(5): 415–422, Mar 2016. doi: 10.1093/aje/kww021. Epub 2016 Feb 17.
- [21] Haiyang Liu, Chunjiang Zhu, Detian Zhang, and Qing Li. Attention-based spatial-temporal graph convolutional recurrent networks for traffic forecasting, 2023. URL <https://arxiv.org/abs/2302.12973>.

- [22] Zewen Liu, Guancheng Wan, B. Aditya Prakash, Max S. Y. Lau, and Wei Jin. A review of graph neural networks in epidemic modeling, 2024.
- [23] Andrey Y. Lokhov, Marc Mézard, Hiroki Ohta, and Lenka Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Phys. Rev. E*, 90(1):012801, 2014. doi: 10.1103/PhysRevE.90.012801.
- [24] Mingjie Qiu, Zhiyi Tan, and Bing-Kun Bao. Msgnn: Multi-scale spatio-temporal graph neural network for epidemic forecasting. *Data Mining and Knowledge Discovery*, 38(4):2348–2376, May 2024. ISSN 1573-756X. doi: 10.1007/s10618-024-01035-w. URL <http://dx.doi.org/10.1007/s10618-024-01035-w>.
- [25] Alexander Rodríguez, Harshavardhan Kamarthi, Pulak Agarwal, Javen Ho, Mira Patel, Suchet Sapre, and B. Aditya Prakash. Data-centric epidemic forecasting: A survey, 2022. URL <https://arxiv.org/abs/2207.09370>.
- [26] Hao Sha, Mohammad Al Hasan, and George Mohler. Source detection on networks using spatial temporal graph convolutional networks. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–11, 2021. doi: 10.1109/DSAA53316.2021.9564188.
- [27] Sirui Song, Zefang Zong, Yong Li, Xue Liu, and Yang Yu. Reinforced epidemic control: Saving both lives and economy, 2020.
- [28] Chee Wei Tan and Pei-Duo Yu. Contagion source detection in epidemic and infodemic outbreaks: Mathematical analysis and network algorithms. *Foundations and Trends® in Networking*, 13(2–3):107–251, 2023. ISSN 1554-0588. doi: 10.1561/13000000068. URL <http://dx.doi.org/10.1561/13000000068>.
- [29] Feng Xie, Zhong Zhang, Liang Li, Bin Zhou, and Yusong Tan. Epignn: Exploring

- spatial transmission with graph neural network for regional epidemic forecasting, 2022. URL <https://arxiv.org/abs/2208.11517>.
- [30] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3634–3640, 2018.
- [31] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.
- [32] Yufan Zheng, Wei Jiang, Alexander Zhou, Nguyen Quoc Viet Hung, Choujun Zhan, and Tong Chen. Epidemiology-informed graph neural network for heterogeneity-aware epidemic forecasting, 2024. URL <https://arxiv.org/abs/2411.17372>.
- [33] Ágnes Bodó, Gyula Y. Katona, and Péter L. Simon. Sis epidemic propagation on hypergraphs, 2015.