Precise Mobility Intervention for Epidemic Control Using Unobservable Information via Deep Reinforcement Learning

Tao Feng
Department of Electronic Engineering, Tsinghua University
Beijing, China

Tong Xia
Department of Computer Science and Technology, University of Cambridge
Cambridge, United Kingdom

Xiaochen Fan
Department of Electronic Engineering, Tsinghua University
Beijing, China

Huandong Wang
Department of Electronic Engineering, Tsinghua University
Beijing, China

Zefang Zong
Department of Electronic Engineering, Tsinghua University
Beijing, China

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

ABSTRACT

To control the outbreak of COVID-19, efficient individual mobility intervention for EPidemic Control (EPC) strategies are of great importance, which cut off the contact among people at epidemic risks and reduce infections by intervening the mobility of individuals. Reinforcement Learning (RL) is powerful for decision making, however, there are two major challenges in developing an RL-based EPC strategy: (1) the unobservable information about asymptomatic infections in the incubation period makes it difficult for RL's decision-making, and (2) the delayed rewards for RL causes the deficiency of RL learning. Since the results of EPC are reflected in both daily infections (including unobservable asymptomatic infections) and long-term cumulative cases of COVID-19, it is quite daunting to design an RL model for precise mobility intervention. In this paper, we propose a **V**ariational hi**E**rarc**HIC**al reinforcement **L**earning method for **E**pidemic control via individual-level mobility intervention, namely *Vehicle*. To tackle the above challenges, *Vehicle* first exploits an information rebuilding module that consists of a contact-risk bipartite graph neural network and a variational LSTM to restore the unobservable information. The contact-risk bipartite graph neural network estimates the possibility of an individual being an asymptomatic infection and the risk of this individual spreading the epidemic, as the current state of RL. Then, the Variational LSTM further encodes the state sequence to model the latency of epidemic spreading caused by unobservable asymptomatic infections. Finally, a Hierarchical Reinforcement Learning framework is employed to train *Vehicle*, which contains dual-level agents to solve the delayed reward problem. Extensive experimental results demonstrate that *Vehicle* can effectively control the spread of the epidemic. *Vehicle* outperforms the state-of-the-art baseline methods with remarkably high-precision mobility interventions on both symptomatic and asymptomatic infections.

*Yong Li is the corresponding author.

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3539195

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent planning**; **Planning and scheduling**; **Artificial intelligence**.

KEYWORDS

Epidemic Control Strategy; Hierarchical Reinforcement Learning; Variational LSTM; Contact-risk Bipartite Graph Neural Network

ACM Reference Format:
Tao Feng, Tong Xia, Xiaochen Fan, Huandong Wang, Zefang Zong, and Yong Li *. 2022. Precise Mobility Intervention for Epidemic Control Using Unobservable Information via Deep Reinforcement Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539195

1 INTRODUCTION

In the post-COVID-19 era, citywide lockdown and national mobility restrictions are not desirable EPidemic Control (EPC) strategies anymore. To effectively control the epidemic spreading while maintaining essential social functions and preserving major economic activities, precise prevention through individual-level mobility intervention, *i.e.,* isolating infections in time, has become a key solution. However, this is a challenging task due to the latency of epidemic spreading caused by unobservable asymptomatic infections [21]. Relevant studies [16, 18] have shown that unobservable asymptomatic infections are the main cause to the spread of the epidemic, and the key to control the epidemic is to isolate possible asymptomatic infections. Therefore, in this paper, we consider the latency of epidemic spreading and study to propose an effective individual-level EPC strategy for reducing infections, without significantly compromising people's daily lives.

Public health researchers have studied the problem of EPC for a long time. Traditional rule-based strategies [7] and heuristic methods [13, 19] can control the spread of the epidemic to a certain degree. Yet, they cannot handle the complex urban environment, dynamic mobility patterns, and uncertain pandemic situations. Recently, some deep reinforcement learning (DRL) methods have shown promising results on optimizations for lockdown and reopening policy [6, 15]. For instance, DRL models can efficiently extract information from the complex urban environment and give a flexible and dynamic control that considers the current pandemic

situation. Nevertheless, existing DRL-based methods still lack customized modules to model the latency of epidemic spreading to facilitate EPC decision making.

Inspired by the previous efforts of applying DRL to epidemic control, in this study, we plan to deploy this technique to achieve effective individual mobility intervention by modeling system dynamics and generating proactive strategies for COVID-19 epidemic control. Nevertheless, there are two major challenges in exploiting DRL in our investigated scenario. First, a DRL agent requires instant and holistic observations from the urban environment for action decision making [25]. Meanwhile, in the epidemic situation, the incubation period is uncertain and the asymptomatic infections can hardly be observed. Second, the effectiveness of mobility intervention is not only reflected by the daily COVID-19 infections (including unobservable asymptomatic infections), but also determined by the cumulative cases in a relatively long period. Therefore, for the RL agent, both daily infections and cumulative cases should be utilized as rewards to modify the current policy. However, daily infections include many asymptomatic infections that cannot be observed until the incubation period is over, which causes delayed rewards to the DRL agent. Moreover, cumulative cases take a long time to obtain, which causes sparse rewards to the DRL agent. The delayed and sparse rewards make it difficult to optimize the EPC policy on mobility intervention.

To tackle the above challenges, we propose ***Vehicle***, a **V**ariational hi**E**rarc**HIC**al reinforcement **L**earning method for **E**pidemic control via individual-level mobility intervention. Specifically, to address the first challenge, *Vehicle* employs an unobservable information rebuilding module that consists of a contact-risk bipartite graph neural network (CRBGN) and a variational LSTM (VLSTM). First, it extracts information about asymptomatic infections from current observations through the CRBGN, which models the relationships between individuals as the current state of RL and estimates the risk of each individual spreading the epidemic [17]. Second, it rebuilds intact information of epidemic spreading from the state sequence through the VLSTM, which encodes the state sequence as a basis for the RL's decision-making. In response to the second challenge, *Vehicle* exploits a Hierarchical Reinforcement Learning (HRL) training framework that includes dual-level agents, including an Upper-level Agent and a Lower-level Agent to train the model. The long-term reward given from the environment to the Upper-level Agent can measure cumulative cases in epidemic control. Such that, the overall reward will no longer be delayed, as most asymptomatic infections will transform into symptomatic infections and can be observed from the long-term reward. The Lower-level Agent directly interacts with the environment by inputting imperfect information and outputting actions to the environment. Furthermore, The Upper-level Agent module decomposes the final goal of EPC into multiple short-term goals, and it guides the Lower-level Agent to learn through these short-term goals for solving the delayed and sparse reward challenge. HRL employs a Proximal Policy Optimization (PPO) [20] algorithm for training both the Upper-level Agent and the Lower-level Agent.

In summary, we make the following contributions in this study.

- We investigate individual-level mobility intervention for epidemic control and propose *Vehicle*, a hierarchical reinforcement learning method that exploits unobservable information for more precise EPC strategy optimization.
- We propose an individual contact-risk bipartite graph neural network and a variational LSTM to model the latency of epidemic spreading, which supports the RL agent to rebuild the intact information for mobility intervention from the observation history.
- We exploit a Hierarchical Reinforcement Learning framework to train our DRL model and solve the delayed reward challenge with the goal of improving the training efficiency.
- We conduct extensive multi-scenario experiments and the simulation results show that *Vehicle* has outstanding performance in epidemic control and it significantly outperforms the state-of-the-art baseline methods.

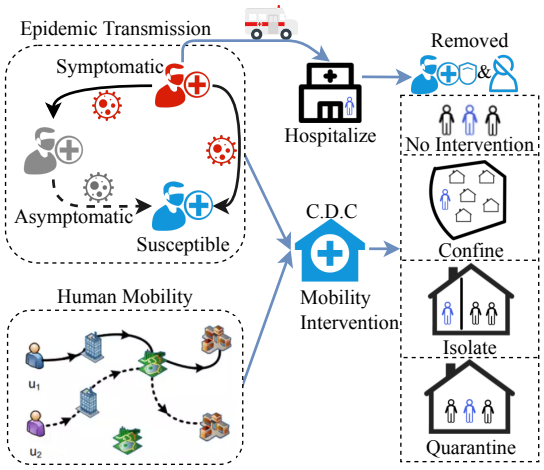

Figure 1: Mobility intervention for epidemic control by the CDC.

2 PROBLEM FORMULATION

We consider an epidemic control for $\mathcal{T}$ days within a city, which is composed of $\mathcal{N}$ areas and $\mathcal{M}$ people. The health status of each individual includes: **Susceptible**, **Asymptomatic**, **Symptomatic**, and **Removed**. Individuals in Asymptomatic or Symptomatic statuses are infected. Susceptible individuals may become Asymptomatic through contact with the infected. The Symptomatic individuals sent to the hospital or isolated at home will transit to Removed status (including recovered and dead) after a period of time. The above four health status transitions are based on the SEIR model [27].

As shown in Figure 1, the Center for Disease Control and Prevention (CDC) acts as the policy maker for epidemic control. The CDC models both the epidemic transmission and the human mobility to select a mobility intervention action for each individual per day. Meanwhile, the policy makers cannot: (1) distinguish between Susceptible individuals and Asymptomatic individuals via non-pharmaceutical methods, and (2) observe the asymptomatic infections. Specifically, we define four kinds of mobility intervention actions: **No Intervention**, **Confine** (no contact with people living outside his/her community), **Quarantine** (no contact with people living outside his/her home), and **Isolate** (no contact at all).

We aim to provide a DRL-based EPC strategy for policy makers to reduce infections and minimize the mobility intervention simultaneously. However, the above two goals are decoupled and even conflicting in strategy optimization, as fewer infections typically require a higher degree of mobility intervention. From the perspective of economic cost C^E [22], we define the Infection-spread-cost as C^I and the Mobility-intervention-cost as C^M to represent the impacts of infections and mobility interventions on economic cost, respectively. In addition, when the number of infections exceeds a certain threshold, the medical system will be infiltrated, resulting in a rapid increase in C^E. Similarly, when mobile intervention is greater than a certain threshold, the social production system will be paralyzed, leading to a sharp rise in C^E. To this end, we use exponential forms by referencing the work of [22] to express the above costs as follows.

$$Q = \lambda_h * n_h + \lambda_i * n_i + \lambda_q * n_q + \lambda_c * n_c,$$

$$C^I = exp\left\{I/\theta_I\right\}, C^M = exp\left\{Q/\theta_Q\right\},$$

$$C^E = C^I + C^M,$$

where $\lambda_h, \lambda_i, \lambda_q$ and λ_c denote scale factors, n_h, n_i, n_q and n_c denote accumulated numbers of hospitalized, isolated, quarantined, and confined people for T simulation days, I denotes the total number of infected people within T days, Q denotes the aggregate of mobility interventions, θ_I and θ_Q refer to the soft thresholds for the medical system's capacity and economic system's endurance, respectively.

EPC Problem Statement: Based on the above notations, we formulate the EPC problem via individual-level mobility intervention as follows.

Definition 1 (EPC problem via Individual-Level Mobility Intervention). *Given a city with $\mathcal{N}$ areas, $\mathcal{M}$ people, the people's historical mobility $\{L_m\}_{m\in\mathcal{M}}$ and the people's health status, the goal of this problem is to train an effective individual-based EPC agent to select a mobility intervention action for each individual per day, thus to minimize the economic cost C^E and make a desirable trade-off between reducing the infections and minimizing the mobility intervention.*

3 METHODOLOGY

3.1 System Overview

An overview of *Vehicle*'s architecture is presented in Figure 2. *Vehicle* first obtains the observation of individual features as well as the relationships from the environment, and then it selects a mobility intervention action for each individual per day to control the epidemic spreading in the environment. To learn a smart individual-based EPC strategy, *Vehicle* has two fundamental modules, *i.e.*, an unobservable information rebuilding module and a hierarchical reinforcement learning module.

Unobservable Information Rebuilding Module: This module aims to rebuild the information for RL's decision-making. Its inputs include individual features and relationships, which are imperfect information obtained from the environment due to the latency of epidemic spreading caused by unobservable asymptomatic infections. Through this module, the rebuilt information is then used as the basis for RL's decision-making. This module consists of an individual contact-risk bipartite Graph Neural Network (GNN) and Variational LSTM. The details will be discussed in the following sections.

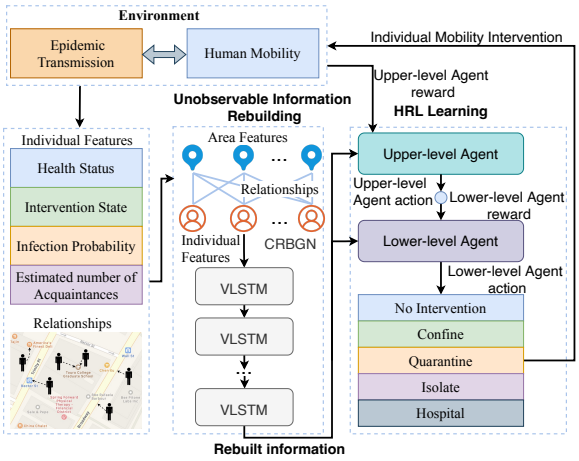

Figure 2: Overview of *Vehicle*. It first obtains the observation of individual features as well as relationships from the environment. Then, an unobservable information rebuilding module rebuilds the information of asymptomatic infections from the observation, thus to facilitate RL's decision-making.

HRL Module: This module is to solve the delayed and sparse reward challenge, which contains an Upper-level Agent module and a Lower-level Agent module. It aims to search for a smart individual-based EPC strategy based on the rebuilt information from the unobservable information rebuilding module.

3.2 Unobservable Information Rebuilding

The unobservable information rebuilding module of *Vehicle* includes individual contact-risk bipartite GNN and a variational LSTM, as illustrated in Figure 3. The Individual CRBGN inputs individual features as node information and the relationships as edges. Then, it outputs embedded individual features to VLSTM at each time step. The specific details will be introduced as below.

3.2.1 ***Individual Contact-Risk Bipartite GNN.*** The imperfect information obtained from the environment consists of raw individual features s_d and relationships A in each time slot, which is also the input of CRBGN. Since the *Vehicle* plays the role of a policy maker, it observes the features of all individuals and gives corresponding control measures for each individual. Here, the raw individual features s_d are the integration of each individual's information. For the k-th individual, the feature $s_d^{(k)}$ is defined as $s_d^{(k)} = (l^{(k)}, v^{(k)}, u^{(k)}, e^{(k)})$, where $l^{(k)}$ represents the health status, v^k represents the intervention state, $u^{(k)}$ represents the estimated number of acquaintances and $e^{(k)}$ represents the probability of infection. Correspondingly, the global raw individual features are denoted by $s_d = (s_d^{(1)}, s_d^{(2)} \dots s_d^{(M)})$. In particular, the probability of infection e measures the risk of epidemic transmission caused by either random movement or contact of asymptomatic infections. Besides, e is estimated based on the historical contacts between

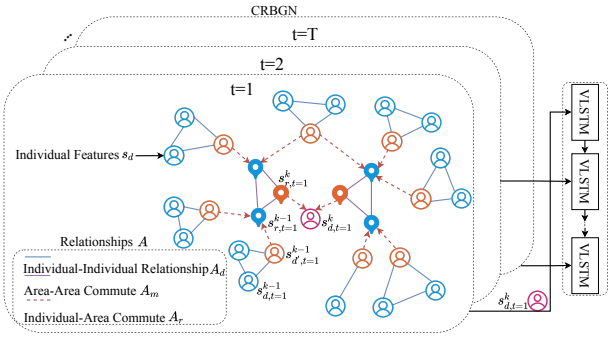

Figure 3: The Unobservable information rebuilding module is composed of an individual contact-risk bipartite GNN and a variational LSTM.

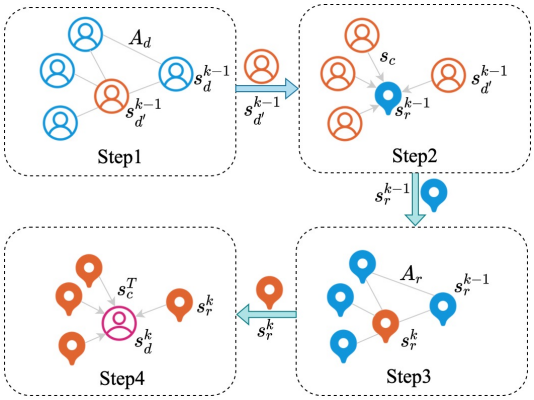

Figure 4: The construction process of CRBGN.

the susceptible and the infected (see Appendix A.2 for details). The estimated number of acquaintances u reflects the risk of the asymptomatic infections spreading the epidemic to others, which can be inferred from individual historical trajectories with the community detection and friendship prediction methods [26]. The unobservable asymptomatic infections and the latent spreading of epidemic together make the raw individual features obtained from the environment not absolutely accurate, and the imperfect information will impact the precision of RL's decision for EPC policy. In addition, the relationships A consists of individual-individual relationship A_d, area-area commute A_m and individual-area commute A_r. Note that A_d is an estimation relationship inferred from historical contacts between individuals (see Appendix A.5 for details).

The CRBGN of *Vehicle* extracts information about asymptomatic infections through individuals' regular commute and social relationships, which can help estimate each individual's infection risk [17]. Moreover, CRBGN regards individuals and city areas as two kinds of nodes, thus enabling us to model individual-individual contact risk by individual-area-individual relationship. In addition, individuals' regular commute and social relationships can also be modeled with the help of the CRBGN module.

Specifically, Individual Contact-Risk bipartite GNN is designed on the basis of GCN [14] and GraphsSAGE [9]. We use s_r^k, s_d^k to

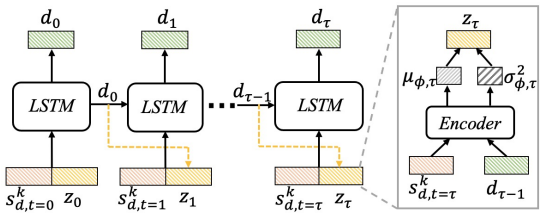

Figure 5: The structure of Variational LSTM.

denote the area-nodes' features and individual-nodes' features outputted by the k-th GNN layer, respectively. The detailed layer-calculation of Individual Contact-Risk GNN is as follows:

$$s_{d'}^{k-1} = \sigma(A_d s_d^{k-1} W_1^{k-1} + B_1^{k-1}), \tag{1}$$

$$s_c = softmax(A_m), \tag{2}$$

$$s_r^{k-1} = \sigma(W_2^{k-1}(s_c^T s_{d'}^k + B_2^{k-1}), \tag{3}$$

$$s_r^k = \sigma(A_r s_r^{k-1} W_3^{k-1} + B_3^{k-1}), \tag{4}$$

$$s_d^k = \sigma(W^k s_c^T s_r^k + B^k), \tag{5}$$

where $s_{d'}^{k-1}$ denotes graph embedding of individuals through the social relationships of the individual nodes, W_1^{k-1}, B_1^{k-1}, W_2^{k-1}, B_2^{k-1}, W_3^{k-1}, B_3^{k-1}, W^k and B^k are trainable parameters.

The construction process of GNN is shown in Figure 4. In step 1, individual-individual relationships are used as edge weights to calculate the individual-node features, as in Eq. (1). In step 2, related commuting individual characteristics are aggregated to calculate the area-node features, as in Eqs. (2-3). In step 3, area-area commute is used as edge weights to calculate the area-node embedded features, as in Eq. (4). In step 4, it aggregates the features of areas where an individual has commuted regularly to calculate the final individual-node embedded feature, as in Eq. (5).

3.2.2 ***Variational LSTM.*** Making decisions solely based on s_d^k is not enough and will cause biased results, as the asymptomatic infections information cannot be fully modeled with features in a single time slot. Indeed, the historical sequences of individual features also contain the information of susceptible people transiting into the symptomatic infections in the environment. Modeling such transition can assist *Vehicle* in estimating the possibility of the individual transforming into symptomatic in the future, so as to better estimate the possibility that the individual is an asymptomatic infection with the currently observed features. Therefore, we exploit LSTM-based method to encode historical sequences of individual features $(s_{d,t=0}^k, s_{d,t=1}^k, \ldots s_{d,t=\tau}^k)$ embedded by individual contact-risk GNN to model such transitions [25] and rebuild intact information for RL's decision-making. Since the performance of LSTM is not stable when dealing with a dynamic environment (*e.g.*, epidemic spreading) [10], we combine the Variational Autoencoder (VAE) with LSTM to enhance its robustness [1]. The integrated recurrent latent variable model contains an inference model and a generative model, and it can learn to encode complicated sequential features of $s_{d,t}$ with a stochastic latent variable z_t.

The inference model of VLSTM approximates the latent variable z_t given t-th embedded observation $s_{d,t}$ and d_{t-1}. Note that d_t models historical feature sequence information with asymptomatic

infections through LSTM. On the basis of d_{t-1} and $s_{d,t}$, the latent variable z_t further models the dynamic epidemic spreading into the feature sequence information with asymptomatic infections. The overall structure of VLSTM is shown in Figure 5. By denoting the inference model as ϕ, we present the detailed process of modeling z_t and d_{t-1} as follows.

$$[\mu_{\phi,t}, diag(\sigma^2_{\phi,t})] = \phi^{encoder}(s^k_{d,t}, d_{t-1}), \tag{6}$$

$$z_t \mid s^k_{d,t} \sim N(\mu_{\phi,t}, diag(\sigma^2_{\phi,t})). \tag{7}$$

Then, d_t can be obtained by

$$d_t = f^{LSTM}(d_{t-1}; z_t, s^k_{d,t}). \tag{8}$$

We denote the generative model of VLSTM by θ. We leverage θ to predict the next state variable distribution of VLSTM and calculate its loss by

$$[\mu_{\theta,t}, diag(\sigma^2_{\theta,t})] = \theta^{prior}(d_{t-1}), \tag{9}$$

where θ^{prior} and $\phi^{encoder}$ are parameterized mappings of neural networks and d_t is the state variable of VLSTM.

Unlike the variational recurrent model [10] that uses LSTM as an independent prediction model, the proposed VLSTM acts as the encoder for the entire model, as it reveals in Figure 3. Correspondingly, the variational loss of VLSTM is calculated by

$$\begin{aligned} L_{VLSTM} &= D_{KL}[q_\phi(z_t)||p_{\theta\phi}(z_t)] \\ &= log\frac{\sigma_{\phi,t}}{\sigma_{\theta,t}} + \frac{(\mu_{\phi,t} - \mu_{\theta,t})^2 + \sigma^2_{\phi,t}}{2\sigma^2_{\theta,t}} - \frac{1}{2}. \end{aligned} \tag{10}$$

The VLSTM is trained together with the HRL framework, and the process is discussed in detail in the following subsection.

3.3 Hierarchical Reinforcement Learning

We devise a hierarchical reinforcement learning framework to train our model and solve the delayed and sparse reward challenge. The HRL framework contains dual-level agents, including an Upper-level Agent module and a Lower-level Agent module. The Upper-level Agent module decomposes the final goal of EPC into multiple short-term goals, and it guides the Lower-level Agent to learn through these short-term goals, thus solving the delayed and sparse reward challenge. Specifically, the HRL model is designed based on FeUdal Networks (FuN) [23], a novel architecture that formulates sub-goals of reinforcement learning as directions in latent state space. Compared with FuN, our HRL adds an Individual Contact-Risk GNN module and a VLSTM module as unobservable information rebuilding modules. A schematic illustration of HRL is shown in Figure 6, and we introduce its details as follows.

Upper-level Agent: Recall the architecture of *Vehicle* in Figure 2, the Upper-level Agent does not directly interact with the environment. The long-term reward given from the environment to the Upper-level Agent can measure cumulative cases of epidemic control. Such that, the overall reward will no longer be delayed, as most asymptomatic infections will transform into symptomatic infections and can be observed from the long-term reward. The settings of the Upper-level Agent are as follows.

- **State:** At each time step t, the state of Upper-level Agent $s_{d,t}$ is defined as the individual features $s_{d,t}$ as discussed in Sec 3.2.

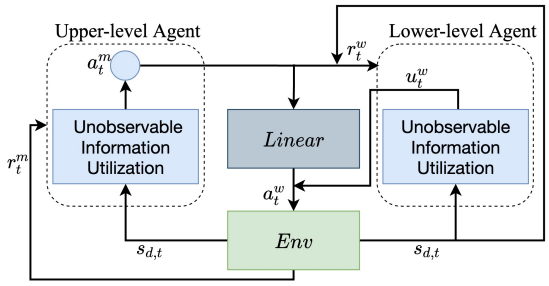

Figure 6: The schematic illustration of HRL.

- **Action:** In our HRL setting, the Upper-level Agent's action a^m_t is to decompose the long-term epidemic control into intrinsic short-term goals to guide the Lower-level Agent,

$$a^m_t = f_u(s_{d,t}, s_{d,t-1} \cdots s_{d,0}), \tag{11}$$

where f_u denotes unobservable information rebuilding module.
- **Reward:** The reward of the Upper-level Agent r^m_t is to evaluate the final result of the EPC strategy. Therefore, we set our reward as negative C^E as follows.

$$r^m_t = \begin{cases} -exp\left\{\frac{I}{\theta_I}\right\} - exp\left\{\frac{Q}{\theta_Q}\right\}, & t = T \\ 0, & else \end{cases} \tag{12}$$

where T denotes total number of EPC days.

Lower-level Agent: The Lower-level Agent directly interacts with the environment by inputting imperfect information and outputting actions to the environment. The intrinsic reward used to train the Lower-level Agent is given by the Upper-level Agent and it can be consecutively obtained at each time step. In this way, the dual-level agents together solve the sparse reward challenge. The settings of the Lower-level Agent are as follows.

- **State:** The state of Lower-level Agent $s_{d,t}$ is the same as the state of Upper-level Agent.
- **Action:** The action of the Lower-level Agent a^w_t is determined by u^w_t and a^m_t. As shown in Figure 6, the a^w_t can be calculated by

$$u^w_t = f_u(s_{d,t}, s_{d,t-1} \cdots s_{d,0}), \tag{13}$$

$$a^w_t = softmax(u^w_t Linear(a^m_t)), \tag{14}$$

where a^w_t denotes the probability of a joint intervention action for all individuals. For the k-th individual, the space for $a^w_{k,t}$ contains four actions, *e.g.*, no intervention, confine, isolate and quarantine.
- **Reward:** The goal given by the Upper-level Agent each day composes the intrinsic reward of the Lower-level Agent and encourages it to output intervention actions to the environment. As such, *Vehicle* can achieve desirable EPC results by making the reward no longer sparse. The detailed calculation is shown in Eq. (15).

$$r^w_t = \frac{1}{c}\sum_{i=1}^{c} d_{cos}(s^w_{d,t} - s^w_{d,t-i}, a^m_{t-i}), \tag{15}$$

where $d_{cos}(\alpha, \beta) = \alpha^T\beta(|\alpha| \cdot |\beta|)$ is the cosine similarity between two vectors and it guides the Lower Level agent's policy toward the goal. Here, c is hyperparameter.

Based on the HRL framework, we introduce how to utilize it to train *Vehicle*. HRL employs a Proximal Policy Optimization (PPO) [20] algorithm for both Upper-level Agent and Lower-level Agent modules. PPO is an actor-critic method, which includes two estimators: (1) an actor that plays the role of our RL agent and generates actions according to the current policy, and (2) a critic to estimate the value of the current state-action pair during training and assists with the training of actor. The loss contains the loss of RL and the variational loss of VLSTM. We denote La^m, Lc^m, La^w, Lc^w and h_t as the actor loss of the Upper-level Agent, the critic loss of the Upper-level Agent, the actor loss of the Lower-level Agent, the critic loss of the Lower-level Agent and the feature history, respectively. The actor and critic losses of Upper-level Agent are calculated by

$$La^m = A_t^m d^w min[q_t^m, clip(q_t^m, 1-\epsilon, 1+\epsilon)] + L_{VLSTM}^m, \tag{16}$$

$$q_t^m = \frac{p(g_t|h_t^m)}{p^{old}(g_t|h_t^m)}, \tag{17}$$

$$r_t^w = d_{cos}(s_{t+c}^m - s_t^m, g_t), \tag{18}$$

$$A_t^m = r_t^m + \gamma V_t^m(h_{t+1}^m) - V_t^m(h_t^m), \tag{19}$$

$$Lc^m = (r_t^m + \gamma V_t^m(h_{t+1}^m) - V_t^m(h_t^m))^2. \tag{20}$$

The actor and critic losses of Lower-level Agent are calculated by

$$La^w = A_t^w min[q_t^w, clip(q_t^w, 1-\epsilon, 1+\epsilon)] + L_{VLSTM}^w, \tag{21}$$

$$q_t^w = \frac{p(a_t|h_t^w)}{p^{old}(a_t|h_t^w)}, \tag{22}$$

$$A_t^w = r_t^w + \gamma V_t^w(h_{t+1}^w) - V_t^w(h_t^w), \tag{23}$$

$$Lc^w = (r_t^w + \gamma V_t^w(h_{t+1}^w) - V_t^w(h_t^w))^2, \tag{24}$$

where p is the probability value of the strategy output action, V is the value function of critic, γ is hyperparameter. We summarize the details of the training process of our proposed model in Algorithm 1. As we can observe, firstly, *Vehicle* interacts with the environment and collects a series of transitions by storing them in the set of transition D in preparation for training (lines 3-11). Then, it samples some transitions batches from D and trains them through the PPO method with batch gradient updating (lines 12-21). This process will be repeated for M episodes until both the Upper-level Agent and the Lower-level Agent converge.

Specially, in order to improve the exploration efficiency of the *Vehicle*, we use the individual infection probability e to constrain the agent's action space exploration. We assume that individuals with greater infection probability will be subject to more stringent control actions, which ensures individuals with high probability of infection to be identified as high risks. In addition, we summarize the commonly used notations in Table 5 (see Appendix A.1).

Algorithm 1 The training algorithm of *Vehicle*

1: Initialize critic and actor parameters for Upper-level Agent and Lower-level Agent with ϑ, ω, φ and ψ;
2: **for** $episodes = 0, 1, 2, ...M$ **do**
3: Initialize a set of transitions $D = \{\tau_i\}$
4: Receive an initial state s_0 (or history trajectory h_0)
5: **for** $t = 0, 1, 2, ...T$ **do**
6: Obtain goal a_t^m using Eq. (11)
7: Obtain action a_t^w using Eqs. (13) and (14)
8: Execute action a_t^w, obtain r_t^m and state s_t from the environment
9: Obtain reward r_t^w using Eq. (15)
10: Store transition $(s_{t-1}, h_{t-1}, a_t^m, a_t^w, r_t^m, r_t^w)$ into D
11: Update history trajectory $h_t = [h_{t-1}, s_t]$
12: Using set of trajectories D for batch gradient updating.
13: Update critic of Upper-level Agent by minimizing Eq. (20):
14: $\vartheta_{k+1} = \underset{\vartheta}{argmin} \frac{1}{|D|} Lc^m \sum_{\tau_i \in D} Lc_{\vartheta_k}^m$
15: Update actor of Upper-level Agent by maximizing Eq. (16):
16: $\omega_{k+1} = \underset{\omega}{argmax} \frac{1}{|D|} La^m \sum_{\tau_i \in D} La_{\omega_k}^m$
17: Update critic of Lower-level Agent by minimizing Eq. (24):
18: $\varphi_{k+1} = \underset{\varphi}{argmin} \frac{1}{|D|} Lc^w \sum_{\tau_i \in D} Lc_{\varphi_k}^w$
19: Update actor of Lower-level Agent by maximizing Eq. (21):
20: $\psi_{k+1} = \underset{\psi}{argmax} \frac{1}{|D|} La^w \sum_{\tau_i \in D} La_{\psi_k}^w$
21: Update critics and actors for Upper-level Agent and Lower-level Agent via the PPO method

4 PERFORMANCE EVALUATION

To fully evaluate the performance of *Vehicle*, in this section, we conduct extensive trace-driven experiments on mobility intervention for epidemic control by applying *Vehicle* and state-of-the-art baseline methods in multiple scenarios.

4.1 Experiment Setups

4.1.1 Simulation Environment. We build a simulation environment based on the Prescriptive Analytics for the Physical World (PAPW) Challenge [6] for pandemic mobile intervention competition. The simulator can accurately simulates EPC scenarios based on individual mobility and has been widely used by existing research studies [2, 6, 13, 19]. According to epidemiological research, the Basic Reproductive Rate R_0 denotes the average number of people infected by one person in a susceptible population. The R_0 of our simulated disease ranges from 2 to 4, which is consistent with COVID-19 (estimated between 2 and 4). The number of epidemic simulation days T is set as 60. Due to the page limit, details of the simulator and experimental settings are provided in Appendixes A.3 and A.4.

4.1.2 Experimental Scenarios. We consider a citywide epidemic control scenario composed of N areas with a population of 10,000. We set t_s as the days to start mobility intervention after discovering the first patient, I_n as the number of external daily contacted infections for the first 20 days, I_a as the number of initial infections and t_c as the back-to-work time of individuals. In addition, we set up five experimental scenarios, including **Basic**, **Larger**, **Changeable**, **Larger-Initial-Infections** and **Late**. Detailed introductions to the above scenarios can be referred from Table 6 in Appendix A.3. To make the scenarios more realistic, we follow a basic assumption that all symptomatic patients should be sent to the hospital.

4.1.3 Evaluation Metrics. We select three evaluation metrics from different perspectives as follows: **I**, the total number of infected

Table 1: Performance comparison over all baselines in five scenarios (best results in bold).

Scenario	Basic			Larger			Changeable			Larger-Initial-Infections			Late		
Method	I	Q	E	I	Q	E	I	Q	E	I	Q	E	I	Q	E
No Intervention	6254	82012	>10000	5822	74958	>10000	6198	78203	>10000	7787	116786	>10000	6179	70264	>10000
Lockdown [8]	**45**	300231	>10000	**45**	300225	>10000	**49**	300253	>10000	**461**	301475	>10000	**106**	275523	>10000
Expert(0.01)	255	6197	3.52	213	5423	3.25	239	6351	3.50	899	35728	41.65	245	6119	3.48
Expert(0.015)	246	6487	3.55	225	5993	3.39	283	7188	3.81	1197	29187	29.47	296	7492	3.93
Degree-Sample [24]	1489	103363	>10000	1164	103298	>10000	992	96564	>10000	3866	198892	>10000	1793	101058	>10000
Degree-Order [3]	1256	92397	>10000	1248	90546	>10000	1551	94128	>10000	3472	133198	>10000	1664	95247	>10000
GBM [19]	191	3887	2.94	206	4548	3.09	227	5033	3.23	529	26766	17.43	235	4671	3.19
EITL [13]	217	4239	3.10	261	5433	3.41	261	5315	3.39	801	22743	14.68	272	5428	3.44
HRLI [2]	207	4107	3.02	199	4455	3.05	210	4279	3.06	1200	24734	22.89	230	4775	3.20
IDRLECA [6]	188	3818	2.92	190	4288	2.99	188	3938	2.94	1017	20408	15.34	219	4396	3.1
Vehicle (Ours)	160	**3418**	**2.78**	163	**3486**	**2.80**	169	**3770**	**2.86**	898	**18170**	**12.18**	172	**3560**	**2.84**

Table 2: Performance comparison with perfect information (best results in bold).

Scenario	Basic			Larger			Changeable			Larger-Initial-Infections			Late		
Method	I	Q	E	I	Q	E	I	Q	E	I	Q	E	I	Q	E
IDRLECA	137	3749	2.77	170	4132	2.91	153	4069	2.86	983	20103	14.61	193	4178	2.99
Vehicle (Ours)	110	2566	2.54	144	3262	2.72	119	2719	2.58	840	17210	10.95	143	3144	2.70
Perfect	**88**	**1571**	**2.36**	**69**	**1197**	**2.28**	**93**	**1630**	**2.38**	**463**	**8404**	**4.84**	**124**	**2175**	**2.52**

people in all simulation days, which measures the effectiveness of EPC strategies in epidemic control; **Q**, the aggregated mobility interventions as defined in Section II, which measures the effectiveness of EPC strategies in minimizing mobility interventions; and **E**, the economic cost of the EPC strategy, which measures the impact of EPC strategies on economy. To make a fair comparison with baseline methods, the settings of Q are the same with the PAPW Challenge [6], where $\lambda_h = 1$, $\lambda_i = 0.5$, $\lambda_q = 0.3$ and $\lambda_c = 0.2$. Similarly, we set $\theta_I = 500$ and $\theta_Q = 10,000$ for E by referring to the PAPW Challenge [6].

4.1.4 Baseline Methods. To make a comprehensive comparison, nine baseline methods are adopted in performance evaluation. First, we set up three rule-based baseline methods, including: (1) **No Intervention**, an EPC strategy with no intervention at all; (2) **Lockdown** [8], an EPC strategy that conducts lockdown in a city during the mobility intervention period; (3) **Expert (0.01)** and **Expert (0.015)**, the same-type EPC strategies that isolate individuals based on an infection probability model and a given probability threshold. Second, we adopt two empirical epidemic control methods, including: (4) **Degree-Sample** [24], which sets the number of each individual's acquaintances as n and isolates them by an empirical probability of $(n-4)/n$ ($n \geq 5$); and (5) **Degree-Order** [3], which counts the number of contacts by an individual in the last five days and selects individuals with the top 30% of contacts for isolation. Third, we utilize two winning solutions in the PAPW challenge for comparison, including (6) **GBM** [19], a baseline for epidemic intervention by predicting each individual's health status and strikes a balance between precision and recall; (7) **EITL** [13], which adjusts the epidemic strategy through a heuristic algorithm. At last, two representative reinforcement learning baselines are adopted, including: (8) **HRLI** [2], which combines individual prevention with regional control in the EPC strategy; and (9) **IDRLECA** [6], a state-of-the-art baseline method that exploits GNN as the encoder to extract features from individuals for EPC strategy optimization. In particular, we do not compare the methods of the 1st [11] and 2nd [28] winners in the PAPW challenge, because they assume that the asymptomatics are observable to policy makers. However, this is not feasible in practical applications.

4.2 Results Analysis

Overall Epidemic Control Results. To verify that *Vehicle* can effectively make a trade-off between metrics I and Q while simultaneously minimize E, we conduct epidemic control experiments across five scenarios. Each experimental result is derived from the average of ten simulations under the same setting. Table 1 summarizes the overall performance comparison, where *Vehicle* significantly outperforms all baseline methods in terms of I, Q and E. In comparison with expert baselines, *Vehicle* can trade off between reducing the infections and minimizing the economic costs. In comparison with the empirical epidemic control methods, *Vehicle* exploits deep neural networks to extract high-level representations from each individual's raw features to make more accurate decisions. Compared to the baselines of PAPW winners, *Vehicle* achieves significantly lower economic costs and successfully reduces the number of infections with fewer mobility interventions. Compared to the state-of-the-art RL-based methods, *Vehicle* outperforms them in all metrics by exploiting VLSTM and HRL to solve the partially-observable problems in epidemic observations. The above results also validate that daily reward would effectively enhance *Vehicle*'s decision-making with imperfect information.

Spatio-temporal Comparison with the IDRLECA [6]. We draw the spatial distribution and histogram of infections (the symptomatic and asymptomatic) given by *Vehicle* and IDRLECA for two periods in Figure 7. It can observed that *Vehicle* has fewer infections in most residential areas than IDRLECA in the first twenty days with imported infections. In the next forty days, our method can still play a more effective role in epidemic control.

Performance Comparison with the Optimum. To demonstrate that *Vehicle* can compensate for the decision bias caused by unobservable information, we set up an optimal epidemic control method called **Perfect**. Perfect has full knowledge of asymptomatic infections and is able to isolate all types of infections immediately.

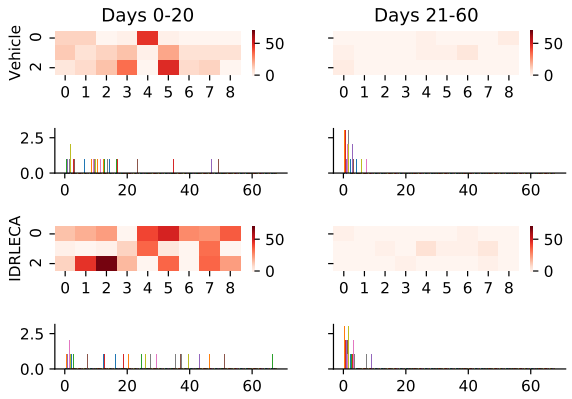

Figure 7: A visualized spatio-temporal comparison of epidemic control results by *Vehicle* and IDRLECA [6] during days 0-20 and days 21-60. The 27 grids represent the spatial infections of symptomatic and asymptomatic in the most significant regions out of 98 areas in scenario-Larger. The histogram shows the temporal distributions of infections during the same period.

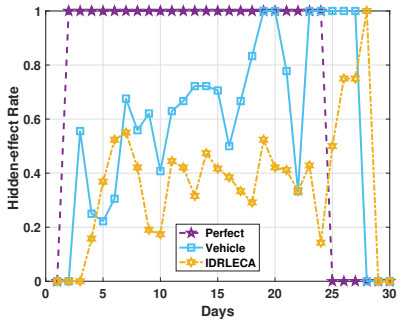

Figure 8: Performance of isolating asymptomatic infections.

We compare the best performances of *Vehicle* and IDRLECA (state-of-the-art baseline) with Perfect in Table 2. The results verify that in most scenarios, *Vehicle* can almost approach the optimal epidemic control results by exploiting the unobservable information via hierarchical reinforcement learning.

Isolation Effectiveness. To verify the effectiveness of *Vehicle* in isolating asymptomatic infections, we further compare it with Perfect and IDRLECA on the asymptomatic-effect rate. The asymptomatic-effect rate indicates the ratio of asymptomatic infections being quarantined to asymptomatic infections along with time. From Figure 8, we can find that the line curve of *Vehicle* in asymptomatic-effect rate is more and more close to that of Perfect, showing that our method can effectively isolate a larger proportion of asymptomatic infections than IDRLECA.

4.3 Generalization Ability

To verify the generalization ability of *Vehicle*, we conduct experiments under different infection settings. Specifically, we change the situation of asymptomatic infections by varying the values of I_n (the number of external daily contacted infections) and I_a (the number of initial infections). The results in Table 3 demonstrate

Table 3: Performance comparison with different I_n and I_a.

		Vehicle			Best Baseline		
I_n	I_a	I	Q	E	I	Q	E
7	0	160	3418	2.78	207	4107	3.02
10	0	190	4286	2.99	260	5672	3.45
13	0	323	7273	3.96	368	7578	4.22
0	300	898	18170	12.18	801	22743	14.68
0	400	1190	23783	21.59	1044	27627	23.92
0	500	1234	28780	29.58	1210	29931	31.19

that under different settings of I_n and I_a, *Vehicle* can always strike a balance between reducing infections and minimizing mobility intervention, thus achieving a better result than the best baseline on all evaluation metrics. The results also demonstrate the scalability of *Vehicle*, as it finds out effective policies to response variations in infection settings.

4.4 Ablation Study

To evaluate the effectiveness of each component of *Vehicle*, including Individual Contact-Risk GNN, VLSTM and HRL, we further conduct an ablation study. Under Scenario-Basic, we evaluate 4 variant models of *Vehicle* as follows. First, the RL-No-GNN denotes a variant of *Vehicle* without the GNN module, which is designed to verify that the relationship between individuals and the regular commuting of individuals are effective on estimating asymptomatic infections' impact. Second, the RL-No-HRL denotes a variant of *Vehicle* without the HRL module, which is designed to verify the importance of dense and fully observable rewards to epidemic control. Third, the RL-No-VLSTM denotes a variant of *Vehicle* without the VLSTM module, which is designed to verify the effectiveness of using historical observation information with imperfect information. At last, the RL-LSTM denotes a variant of *Vehicle* employing ordinary LSTM, which is designed to verify the robustness of the VAE structure to dynamic environment noises.

The comparison results between *Vehicle* and the above variants are shown in Table 4. Particularly, removing the GNN network structure makes it difficult for RL-No-GNN to capture the connections and contacts among individuals. As a result, the cost of infection prevention and epidemic control both increase to a large extent. Removing VLSTM or HRL will make it hard for RL to estimate the impact of asymptomatic infections and isolate them, thus leading to decision bias. By comparing RL-LSTM with *Vehicle*, we can find that the VAE can make the model more robust to noises and help it reduce unnecessary mobility interventions in a dynamic environment. Compared with all the variants, *Vehicle* can find asymptomatic infections more precisely with the help of VLSTM and HRL. By measuring the risk of contact among individuals with the individual contact-risk GNN module, *Vehicle* can achieve more desirable epidemic control results with the fewest infections and lowest economic cost, under a minimum level of mobility intervention.

5 RELATED WORKS

Individual-level Epidemic Intervention. Conventional epidemic control methods model people as graph nodes and use the graph's connectivity to determine the intervention measurement for each individual [2, 6, 12]. Methods based on epidemic prediction further

Table 4: Results of ablation study.

Scenario	Basic		
Method	I	Q	E
RL-No-GNN	209	4625	3.11
RL-LSTM	185	4388	2.98
RL-No-HRL	184	3859	2.92
RL-No-VLSTM	165	3697	2.84
Vehicle	**160**	**3418**	**2.78**

consider the impact of short-term intervention based on the individual's state, yielding better performance than traditional heuristics but still insufficient [13, 19]. Recently, RL-based methods that can capture long-term intervention effects have been explored to achieve more effective epidemic control with flexibility [2, 6, 12, 15]. However, there still lacks a framework to address the observable asymptomatic spread of virus between individuals. To bridge this gap, we propose an information rebuilding module to model the latency of epidemic spreading and capture the observable contact for RL's decision-making.

Asymptomatic Modeling. Existing works have developed SIR and SEIR models to study the potential spread of epidemiological disease [5]. Some researchers have also considered possible hidden infections to improve the prediction accuracy on epidemics [16, 18, 21]. Inspired by the existing achievements in asymptomatic modeling, in this study, we have included a plural of features related to the hidden infections in designing the state space of *Vehicle*'s RL agent, with the goal to achieve more effective decisions on epidemic control.

6 CONCLUSION

In this paper, we study the problem of individual mobility intervention for epidemic control and propose an effective framework named ***Vehicle***. In this framework, we devise an observable information rebuilding module to model the latency of epidemic spreading. We also exploit hierarchical reinforcement learning to train *Vehicle*, which can tackle the delayed and sparse reward challenge. Extensive experiments are conducted on different scenarios, demonstrating that our method can effectively estimate the impact of the asymptomatic and achieve superior epidemic control compared with other baselines. Our in-depth analysis provides valuable intuitions for practical applications, including mobility intervention and POI reopening, and the proposed *Vehicle* can support decision-making in epidemic control.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under 2020AAA0106000, National Natural Science Foundation of China under U21B2036, U20B2060, 62171260 and Young Elite Scientists Sponsorship Program by CIC (Grant No. 2021QNRC001).

REFERENCES

[1] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. *arXiv preprint arXiv:1506.02216* (2015).
[2] Yinzhao Dong, Chao Yu, and Lijun Xia. 2020. Hierarchical Reinforcement Learning for Epidemics Intervention. (2020).
[3] Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429, 6988 (2004), 180–184.
[4] Jie Feng, Can Rong, Funing Sun, Diansheng Guo, and Yong Li. 2020. PMF: A privacy-preserving human mobility prediction framework via federated learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–21.
[5] Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. 2020. Learning to simulate human mobility. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3426–3433.
[6] Tao Feng, Sirui Song, Tong Xia, and Yong Li. 2021. Reinforced Contact Tracing and Epidemic Intervention. (2021). https://www.researchgate.net/publication/349118652_Reinforced_Contact_Tracing_and_Epidemic_Intervention
[7] Abhirup Ghosh and Tong Xia. 2021. Mobility-based Individual POI Recommendation to Control the COVID-19 Spread. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 4356–4364.
[8] Thomas Hale, Anna Petherick, Toby Phillips, and Samuel Webster. 2020. Variation in government responses to COVID-19. *Blavatnik school of government working paper* 31 (2020).
[9] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in neural information processing systems*. 1024–1034.
[10] Dongqi Han, Kenji Doya, and Jun Tani. 2019. Variational recurrent models for solving partially observable control tasks. *arXiv preprint arXiv:1912.10703* (2019).
[11] Yuanshuang Jiang, Linfang Hou, Yuxiang Liu, Zhuoye Ding, Yong Zhang, and Shengzhong Feng. 2020. Epidemic Control Based on Reinforcement Learning Approaches. (2020).
[12] Harshad Khadilkar, Tanuja Ganu, and Deva P Seetharam. 2020. Optimising lockdown policies for epidemic control using reinforcement learning. *Transactions of the Indian National Academy of Engineering* 5, 2 (2020), 129–132.
[13] Joon-Seok Kim, Hyunjee Jin, and Andreas Züfle. 2020. Expert-in-the-Loop Prescriptive Analytics using Mobility Intervention for Epidemics. (2020).
[14] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
[15] Pieter Libin, Arno Moonens, Timothy Verstraeten, Fabian Perez-Sanjines, Niel Hens, Philippe Lemey, and Ann Nowé. 2020. Deep reinforcement learning for large-scale epidemic control. *arXiv preprint arXiv:2003.13676* (2020).
[16] Seyed M Moghadas, Meagan C Fitzpatrick, Pratha Sah, Abhishek Pandey, Affan Shoukat, Burton H Singer, and Alison P Galvani. 2020. The implications of silent transmission for the control of COVID-19 outbreaks. *Proceedings of the National Academy of Sciences* 117, 30 (2020), 17513–17515.
[17] Akihiro Nishi, George Dewey, Akira Endo, Sophia Neman, Sage K Iwamoto, Michael Y Ni, Yusuke Tsugawa, Georgios Iosifidis, Justin D Smith, and Sean D Young. 2020. Network interventions for managing the COVID-19 pandemic and sustaining economy. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30285–30294.
[18] Allyson M Pollock and James Lancaster. 2020. Asymptomatic transmission of covid-19.
[19] Rizzo. 2020. Balancing Precision and Recall for Cost-effective Epidemic Containment. (2020). https://prescriptive-analytics.github.io/file/3-strizzo.pdf
[20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
[21] Rahul Singh, Fang Liu, and Ness B Shroff. 2020. A POMDP based "Learning" Approach for tackling COVID-19. (2020).
[22] Sirui Song, Zefang Zong, Yong Li, Xue Liu, and Yang Yu. 2020. Reinforced Epidemic Control: Saving Both Lives and Economy. arXiv:2008.01257 [cs.AI]
[23] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. 2017. Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*. PMLR, 3540–3549.
[24] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *nature* 393, 6684 (1998), 440–442.
[25] Daan Wierstra, Alexander Foerster, Jan Peters, and Juergen Schmidhuber. 2007. Solving deep memory POMDPs with recurrent policy gradients. In *International conference on artificial neural networks*. Springer, 697–706.
[26] Ke Xu, Keju Zou, Yan Huang, Xiaoyang Yu, and Xinfang Zhang. 2016. Mining community and inferring friendship in mobile social networks. *Neurocomputing* 174 (2016), 605–616.
[27] Zifeng Yang, Zhiqi Zeng, Ke Wang, Sook-San Wong, Wenhua Liang, Mark Zanin, Peng Liu, Xudong Cao, Zhongqiang Gao, Zhitong Mai, et al. 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of Thoracic Disease* 12, 3 (2020), 165.
[28] Rong Zhou, Kaidong Zhao, and Li Ning. 2020. Solving Mobility Intervention of Epidemics using Genetic Algorithm. (2020).

A APPENDIX

A.1 Notation Summary

Table 5: A list of commonly used notations.

Notation	Description
λ_h	Scale factor of accumulated numbers of hospitalized people.
λ_i	Scale factor of accumulated numbers of isolated people.
λ_q	Scale factor of accumulated numbers of quarantined people.
λ_c	Scale factor of accumulated numbers of confined people.
n_h	Accumulated numbers of hospitalized people.
n_i	Accumulated numbers of isolated people.
n_q	Accumulated numbers of quarantined people.
n_c	Accumulated numbers of confined people.
I	Total number of infections.
Q	The aggregate of mobility interventions.
C^E	Economic cost of EPC strategy.
C^I	Infection- spread-cost of EPC strategy .
C^M	Mobility-intervention-cost of EPC strategy.
$s_d^{(k)}$	The feature of the k-th individual.
$l^{(k)}$	The health status of the k-th individual.
$v^{(k)}$	The intervention state of the k-th individual.
$u^{(k)}$	The estimated number of acquaintances of the k-th individual.
$e^{(k)}$	The probability of infection of the k-th individual.
s_d	The global raw individual features.
$s_{d'}^{k-1}$	Graph embedding of individuals through the social relationships.
s_c	The weights of area-area commute.
s_r	Area-node features.
s_r	Area-node embedded features.
s_d^k	Final individual-node embedded features.
A_d	Individual-individual relationship.
A_m	Area-area commute.
A_r	Individual-area commute.
θ_I	Medical system's capacity.
θ_Q	Economic system's endurance.
$W.B$	Learnable parameters.

A.2 Infection Probability Model

The difficulty of epidemic prevention and control lies in finding asymptomatic infections and taking effective measures in time. To help ***Vehicle*** efficiently make use of effective information, we design an infection probability model to estimate the probability of an individual being infected. We define the probability of infection and health of the i-th person as p_i^{infe} and p_i^{hel}, respectively. The simulation environment also calculates the infection probabilities of contact with strangers and acquaintances, which are denoted as p_s and p_c, respectively. The infection probability model works as follows:

Step 1: Trace back all individuals' area-visit history in the past T time steps.

Step 2: For individual $i, i = 1, 2, ..., M$, define his/her probability of being healthy as $p_{i,t}^{hel}$ at time step t. $p_{i,0}^{hel}$ is initialized to be 1 if individual i is not infected. we have the following equation to update $p_{i,t}^{hel}$:

$$p_{i,t}^{hel} = p_{i,t-1}^{hel} * (1 - p_s \frac{N_{t-1}^{infe}}{N_{t-1}^{area}}), t = 1, 2, ..., T,$$

where N_{t-1}^{infe} and N_{area} refer to the number of discovered infections and total number of visitors to the same area as individual i, respectively.

Step 3: Update $p_{i,T}^{hel}$ for acquaintances' contacts:

$$\hat{p}_{i,T}^{hel} = p_{i,T}^{hel} * (1 - p_c).$$

Step 4: Acquire infection probability:

$$p_i^{infe} = 1 - \hat{p}_{i,T}^{hel}.$$

After the above steps, we can obtain the estimated probability of an individual being infected. We will use it as auxiliary information and add it to each individual's state. Also, the estimated probabilities are used as prior knowledge for each agent's action space exploration.

Specially, regarding the personal privacy issues that exist in the use of individual trajectory information for epidemic prevention and control, we tend to solve them from two aspects. On the one hand, the implementers of the strategy are normally credible government and social institutions (*e.g.*, CDC), which reduces the privacy risk to a certain extent. On the other hand, the problem can be solved by some encryption methods such as differential privacy or federated learning [4], thus to reduce the possibility of privacy leakage.

A.3 Introduction of Simulator

Here we present the details of the simulator[1] and experiment settings to help reproduce the results. The simulator is mainly composed of two parts: human movement and epidemic spread. Our simulator simulates the individual mobility in a city of N areas with M people. We consider three kinds of POI: working, residential, and commercial. Each individual's mobility is determined by pre-defined rules. An individual will move from the residential area to working area. After working, the individual may visit a nearby commercial area and then will return to the residential area. The time individuals start to depart or stay at each place meets a certain distribution. The mobility of individuals between the two areas satisfies a certain probability distribution. We have also designated corresponding acquaintances for each individual. The infection probabilities of contact with acquaintances and strangers are P_c and P_s, respectively. P_c is much higher than P_s because the contact distance between acquaintances is much closer. We divided the city into grids and consider that individuals in the same grid at the same time have contact possibilities. The process that contact possibilities lead to new infections is characterized by a SEIR model [27].

As shown in Table 6, we setup five experimental scenarios as follows. (1) **Basic** scenario is to verify the EPC performance of *Vehicle* in an ordinary epidemic scenario, (2) **Larger** scenario is to verify whether *Vehicle* is adaptive for scenarios with larger areas, (3) **Changeable** scenario is to verify whether *Vehicle* is applicable when there are greater differences in individuals' commuting features, (4) **Larger-Initial-Infections** scenario is to verify whether *Vehicle* is effective when there are much more initial asymptomatic infections, and (5) **Late** scenario is to verify the EPC performance of *Vehicle* with a late intervention.

[1]PAPW 2020: https://prescriptive-analytics.github.io/. Simulator: https://hzw77-demo.readthedocs.io/en/round2/.

Table 6: Configurations for multiple experimental scenarios (U denotes the uniform distribution).

Scenario	N	t_s	I_n	I_a	t_c
Basic	11	1	7	0	$U(1,5)$
Larger	98	1	7	0	$U(1,5)$
Changeable	11	1	7	0	$U(1,8)$
Larger-Initial-Infections	11	1	0	300	$U(1,5)$
Late	11	5	7	0	$U(1,5)$

A.4 Experiment Setting

The estimated R_0 is between 2 to 4. For the aggregated mobility interventions, we set $\lambda_h = 1$, $\lambda_i = 0.5$, $\lambda_q = 0.3$ and $\lambda_c = 0.2$, which are the same with the setting in the PAPW Challenge. For the reward and r, we set $\theta_I = 500$ and $\theta_Q = 10000$. For r_t^w, we set c=60 and γ=0.99.

The initial state of an episode is random every time in the training process. We train *Vehicle* for 40,000 steps and the optimizer for training is Adam with a learning rate of 0.00001. To cope with the randomness of the simulator, we take the average results of ten runs. Due to the commercial use of the simulator, we only release the source code of the model: at: https://github.com/tsinghua-fib-lab.

A.5 Estimation Relationship Inferred from Historical Contacts

In existing research studies, the estimation and inference of acquaintance relationships are often inferred from the travel trajectory of individuals. In this work, we exploit an existing method [26] to infer acquaintance relationships via modeling the similarity of individual trajectories. Then, we apply the obtained acquaintance relationship to the Infection Probability Model and Contact-Risk GNN of *Vehicle*.