Journal of Artificial Intelligence Research 71 (2021) 953-992 Submitted 07/2021; published 08/2021

Agent-Based Markov Modeling for Improved COVID-19 Mitigation Policies

Roberto Capobianco* ROBERTO.CAPOBIANCO@SONY.COM
Sony AI
Sapienza University of Rome, 00185 Rome, Italy

Varun Kompella* VARUN.KOMPELLA@SONY.COM
Sony AI

James Ault JAULT@TAMU.EDU
Guni Sharon GUNI@TAMU.EDU
Texas A&M University, College Station, TX 77843, USA

Stacy Jong SVJ284@UTEXAS.EDU
Spencer Fox FOX@UTEXAS.EDU
Lauren Meyers LAURENMEYERS@AUSTIN.UTEXAS.EDU
The University of Texas at Austin, Austin, TX 78712, USA

Peter R. Wurman PETER.WURMAN@SONY.COM
Sony AI

Peter Stone PSTONE@CS.UTEXAS.EDU
Sony AI
The University of Texas at Austin, Austin, TX 78712, USA

Abstract

The year 2020 saw the COVID-19 virus lead to one of the worst global pandemics in history. As a result, governments around the world have been faced with the challenge of protecting public health while keeping the economy running to the greatest extent possible. Epidemiological models provide insight into the spread of these types of diseases and predict the effects of possible intervention policies. However, to date, even the most data-driven intervention policies rely on heuristics. In this paper, we study how reinforcement learning (RL) and Bayesian inference can be used to optimize mitigation policies that minimize economic impact without overwhelming hospital capacity. Our main contributions are (1) a novel agent-based pandemic simulator which, unlike traditional models, is able to model fine-grained interactions among people at specific locations in a community; (2) an RL-based methodology for optimizing fine-grained mitigation policies within this simulator; and (3) a Hidden Markov Model for predicting infected individuals based on partial observations regarding test results, presence of symptoms, and past physical contacts.

1. Introduction

Motivated by the devastating COVID-19 pandemic, much of the scientific community, across numerous disciplines, focused on developing safe, quick, and effective methods to prevent the spread of biological viruses or otherwise mitigate the harm they cause. These methods include vaccines, treatments, public policy measures, economic stimuli, and hygiene

* Joint first authors

education campaigns. Governments around the world are faced with high-stakes decisions regarding which measures to enact at which times, often involving trade-offs between public health and economic resiliency. When combating an ongoing epidemic, many authorities attempt to hinder or even halt the disease spread across the community. Several tools are utilized towards achieving this objective which can, in general, be divided into two classes: disease surveillance (Halliday et al., 2017) and containment strategies (Walensky & del Rio, 2020). Disease surveillance tools commonly use observations such as test results, presence of symptoms, and physical contact patterns along with epidemic models for anticipating the disease progression. Based on the surveillance projection, appropriate containment strategies can be applied. Containment strategies may include individual or collective quarantine orders, enforcing social distancing, closing certain public facilities such as schools or shops, etc. See Walensky and Del-Rio (2020) for a survey of such strategies. Effective combinations of surveillance tools and containment strategies have been shown to lead to desirable outcomes with respect to epidemic progression (Cohen & Kupferschmidt, 2020; Salathé et al., 2020; Kaplan & Forman, 2020). However, the proposed approaches generally rely on simple inference rules regarding individual infection likelihood, for instance by ranking the infection likelihood based on a weighted combination of observed symptoms (Grushka-Cohen et al., 2020).

The premise of this paper is that the challenge of mitigating the spread of a pandemic while maximizing personal freedom and economic activity is fundamentally a sequential decision-making problem: the measures enacted on one day affect the challenges to be addressed on future days. As such, Markov models, which maintain a current state that captures all relevant historical observations, are appealing as a basis for both policy optimization (solved as a Markov decision process - MDP) or statistical inference (solved as a Hidden Markov Model - HMM). Limited ability to perform real-world experiments with human subjects means that validating the proposed models and solutions requires an epidemiological model that accurately captures the spread of the pandemic as well as the effects of government measures. To the best of our knowledge, no existing epidemiological simulator (Khadilkar et al., 2020a; Larremore et al., 2020; Hoertel et al., 2020; Aleta et al., 2020) has the resolution to allow reinforcement learning to explore the regulations that governments are currently struggling with.

Motivated by this observation, the main contributions of this article are:

1. The introduction of PandemicSimulator, a novel open-source[1] agent-based simulator that models the interactions between individuals at specific locations within a community. Developed in collaboration between AI researchers and epidemiologists (the co-authors of this paper), PandemicSimulator models realistic effects such as testing with false positive/negative rates, imperfect public adherence to social distancing measures, contact tracing, and variable spread rates among infected individuals. Crucially, PandemicSimulator models community interactions at a level of detail that allows the spread of the disease to be an emergent property of people's behaviors and the government's policies. An interface with OpenAI Gym (Brockman et al., 2016) is provided to enable support for standard RL libraries;

[1] `https://github.com/SonyAI/PandemicSimulator`

2. A demonstration that a reinforcement learning algorithm using only the aggregated infection summaries of the population through randomized testing, can indeed identify a policy that outperforms a range of reasonable baselines within this simulator;

3. An analysis of the resulting learned policy, which may provide insights regarding the relative efficacy of past and potential future COVID-19 mitigation policies;

4. A Hidden Markov Model that learns and updates individual infection probabilities over a given community, in which the infection probabilities are updated at every time step and evolve between time steps;

5. Theoretical justification for the proposed HMM under specific simplifying assumptions; and

6. A comprehensive empirical study using PandemicSimulator, in which the reported results demonstrate the effectiveness of the proposed model when considering more realistic scenarios (where the simplifying assumptions do not hold).

While the true measure of any resulting policies would be how well they perform in the real world (which for obvious reasons is not something we are able to experiment with at this stage of the research), this article establishes the potential power of RL and Bayesian inference in an agent-based simulator, and may serve as an important first step towards real-world adoption.

The remainder of the paper is organized as follows. We first discuss related work in Section 2 and then introduce PandemicSimulator in Section 3. Section 4 presents our reinforcement learning setup, with results being reported in Section 5. In Section 6 we demonstrate how probabilistic inference can be applied online for proactively identifying infected individuals. Finally, Section 7 reports some conclusions and directions for future work.

2. Related Work

Epidemiological models differ based on the level of granularity at which they track individuals and their disease states. “Compartmental” models group individuals of similar disease states together, assume all individuals within a specific compartment to be homogeneous, and only track the flow of individuals between compartments (Tolles & Luong, 2020). While relatively simplistic, these models have been used for decades and continue to be useful for both retrospective studies and forecasts as were seen during the emergence of recent diseases (Rivers & Scarpino, 2018; Metcalf & Lessler, 2017; Cobey, 2020) like influenza, Ebola (Rivers & Scarpino, 2018), Zika (Metcalf & Lessler, 2017), and now during the ongoing SARS-CoV-2 pandemic (Cobey, 2020).

However, the commonly used macroscopic (or mass-action) compartmental models are less predictive when outcomes depend on the characteristics of heterogeneous individuals. In such cases, network models (Bansal et al., 2007; Liu et al., 2018; Khadilkar et al., 2020a) and agent-based models (Grefenstette et al., 2013; Del Valle et al., 2013; Aleta et al., 2020) may be more useful predictors. Network models encode the relationships between individuals as static connections in a contact graph along which the disease can propagate. For example, Khadilkar et al. (2020a) propose a network model that accounts for population size and geography at India's district-level, but does not consider the movements of specific individuals. Conversely, agent-based simulations, such as the one introduced in this paper, explicitly track individuals, their current disease states, and their interactions with other agents over time. Agent-based models allow one to model as much complexity as desired—even to the level of simulating individual people and locations as we do—and thus enable one to model people's interactions at offices, stores, schools, etc. Because of their increased detail, they enable one to study the hyper-local interventions that governments consider when setting policy. For instance, Larremore et al. (2020) simulate the SARS-CoV-2 dynamics both through a fully-mixed mass-action model and an agent-based model representing the population and contact structure of New York City. Willem et al. (2021), instead, use an agent-based simulator to analyze the impact of contact tracing and location-specific re-openings. Finally, Kerr et al. (2020) propose an agent-based simulator with fine-grained intervention policies and data-driven contact networks.

PandemicSimulator has the level of details needed to allow us to apply RL to optimize dynamic government intervention policies – sometimes referred to as "trigger analysis" since it pertains to identifying triggers for policy changes, such as test positivity rate; e.g. (Duque et al., 2020). RL has been applied previously to several mass-action models (Libin et al., 2020; Song et al., 2020). Libin et al. (2020), for example, adopt a meta-population model that consists of several patches, where each patch represents one administrative region in Great Britain. More specifically, within each patch, they consider an age-structured SEIR model. They use this model to conduct an experiment to learn a joint policy to control a community of 11 tightly coupled districts. Likewise, Song et al. (2020) use RL to search mobility control policies that can simultaneously minimize infection spread and maximally retain mobility. To this end, they use a model based on the traditional SIR model, where they introduce the hospitalized condition, while they simulate a city's urban-mobility demand at any time step as a mobility matrix. These models, however, do not take into account individual behaviors or any complex interaction patterns. The work that is most closely related to our own includes the SARS-CoV-2 epidemic simulators from Hoertel et al. (2020), Kerr et al. (2020), and Aleta et al. (2020), which model individuals grouped into households who visit and interact in the community. While their approach builds accurate contact networks of real populations, it does not support modeling how the contact network would change as the government intervenes, as they directly modify the network connections to implement different strategies. Instead, interventions generally percolate based on actions taken on the level of individuals or venues. Moreover, although potentially supported, they do not perform any kind of learning in the simulator. A different line of work (Xiao et al., 2020) presents a detailed, pedestrian level simulation that simulates transmission indoors and studies relevant interventions such as social distancing and mandatory face masks in a specific closed area. Liu (2020) presents a microscopic approach to model and optimize epidemic spread based on self-imposed individual behaviours, specifically confinement, self-isolation, and two-meter distance. In this work, a simplified infection model is used where only two infection states are considered ('infected' and 'healthy'), interactions among agents are assumed to be uniform, and meeting with infected agents results in immediate infection. Other work (Khadilkar et al., 2020b) propose using RL for optimizing a policy enacting or lifting a full lockdown, but do not consider alternative intervention methods.

For any model to be accepted by real-world decision-makers they must be provided with a reason to trust that it accurately models the population and spread dynamics in their community. For both mass-action and agent-based models, this trust is typically best instilled via a model calibration process that ensures that the model accurately tracks past data. For example, Hoertel et al. (2020) perform a calibration using daily mortality data until 15 April with a loss function based on the Kolmogorov–Smirnov statistic. Similarly, Libin et al. (2020) calibrate their model based on the symptomatic cases reported by the British Health Protection Agency for the 2009 influenza pandemic. Aleta et al. (2020), instead, only calibrate the weights of intra-layer links by means of a rescaling factor, such that the mean number of daily effective contacts in that layer matches mean number of daily effective contacts in the corresponding social setting. Similarly, we demonstrate that our model can be calibrated to track real-world data in Section 3.6.

With respect to infection probability inference, previous work (Shoer et al., 2020; Gudbjartsson et al., 2020a) attempts to estimate COVID-19 infection probabilities based on different features, namely, age, gender, presence of prior medical conditions, general feeling, and the symptoms fever, cough, shortness of breath, sore throat and loss of taste or smell. The observed correlations between the mentioned attributes and positively testing for COVID-19 are reported in these studies, without explicitly considering physical contact tracing or inference over successive days. Grushka-Cohen et al. (2020) consider similar symptom attributes along with a binary attribute designating contact with a confirmed case. Full contact history is not considered based on the following explanation, “The purpose of data security is similar to the task of testing and contact tracing organizations. The sheer amount of daily user activity in IT database systems prevents testing and logging every action”. Grushka-Cohen et al. (2020) show that reported contact with a confirmed case is the dominant feature for determining the probability that an individual will test positive. Again, a Markov process is not assumed and probabilities are not updated over time. Another line of previous work (Li et al., 2018; Lefèvre & Simon, 2019; Marwa et al., 2018; Almaraz & Gómez-Corral, 2018; Britton & Pardoux, 2019) models epidemic progression as a Markov process. However, such models assume full observability regarding the susceptible, infected, and recovered sub-groups. The resulting statistical inference relates to the infection distribution for the entire population and not per individual. While Drakopoulos, Ozdaglar, and Tsitsiklis (2014) and Hoffmann and Caramanis (2018) do not assume full observability, their approach only considers susceptible–infectious–susceptible (SIS) processes. Differently, we assume a Hidden Markov Model that learns and updates individual infection probabilities over a given community, in which the infection probabilities are updated end evolve at every time step.

3. PandemicSimulator: A COVID-19 Simulator

The functional blocks of PandemicSimulator, shown in Figure 1, are:

- *locations*, with properties that define how people interact within them;
- *people*, who travel from one location to another according to individual daily schedules;
- an *infection model* that updates the infection state of each person;

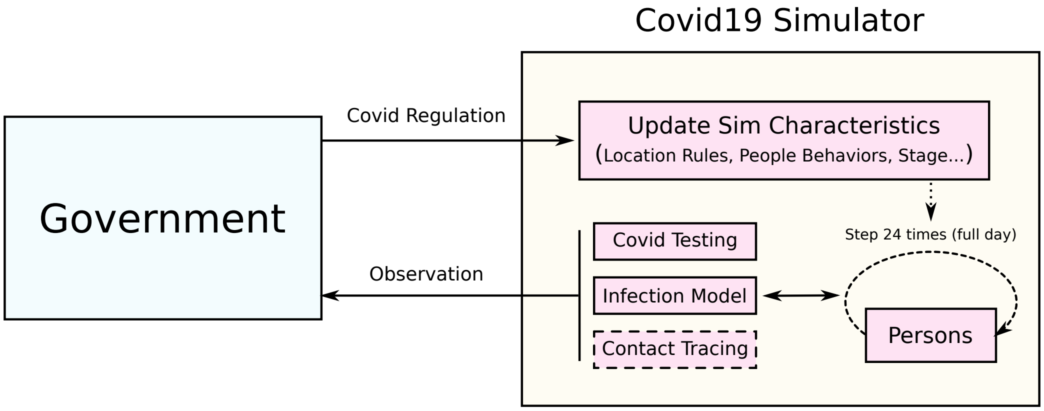


Figure 1: Block diagram of the simulator.

- an optional *testing strategy* that imperfectly exposes the infection state of the population;
- an optional *contact tracing* strategy that identifies an infected person's recent contacts;
- a *government* that makes policy decisions.

The simulator models a day as 24 discrete hours, with each person potentially changing locations each hour. At the end of a day, each person's infection state is updated. The government interacts with the environment by declaring *regulations*, which impose restrictions on the people and locations. If the government activates testing, the simulator identifies a set of people to be tested and (imperfectly) reports their infection state. If contact tracing is active, each person's contacts from the previous days are updated. The updated perceived infection state and other state variables are returned as an observation to the government. The process iterates as long as the infection remains active. The following subsections describe the functional blocks of the simulator in greater detail.

3.1 Locations

Each location has a set of attributes that specify when the location is open, what roles people play there (e.g. worker or visitor), and the maximum number of people of each role. These attributes can be adjusted by regulations, such as when the government determines that businesses should operate at half capacity. Non-essential locations can be completely closed by the government. The location types used in our experiments are *homes*, *hospitals*, *schools*, *grocery stores*, *retail stores*, *hair salons*, *bars* and *restaurants*. The simulator provides interfaces to make it easy to add new location types.

One of the advantages of an agent-based approach is that we can more accurately model variations in the way people interact in different types of locations based on their roles. The base location class supports workers and visitors, and defines a *contact rate*, b^{loc}, as a 3-tuple $(x, y, z) \in [0, 1]^3$, where x is the worker-worker rate, y is the worker-visitor rate, and z is the visitor-visitor rate. These rates are used to sample interactions every hour in each location to compute disease transmissions. For example, consider a location that has a contact rate of $(0.5, 0.3, 0.4)$ and 10 workers and 20 visitors. In expectation, a worker would make contact with 5 co-workers and 6 visitors in the given hour. Similarly, a visitor would be expected to make contact with 3 workers and 8 other visitors. Refer to Table 1 for a listing of the contact rates and other parameters for all location types used in our experiments.

The base location type can be extended for more complex situations. For example, the *hospital* location type is extended to include an additional role (critically sick patients), a capacity representing ICU beds, and contact rates between workers and patients.

3.2 Population

A *person* in the simulator is an automaton that has a state and a person-specific behavior routine. These routines create person-to-person interactions throughout the simulated day and induce dynamic contact networks.

Individuals are assigned an age, drawn from the distribution of the U.S. age demographics, and are randomly assigned to be either high risk or of normal health. Based on their age, each person is categorized as either a *minor*, a *working adult*, or a *retiree*. Working adults are assigned to a work location, and minors to a school, which they attend 8 hours a day, five days a week. Adults and retirees are assigned favorite hair salons which they visit once a month, and grocery and retail stores which they visit once a week. Each person has a compliance parameter that determines the probability that the person flouts regulations each hour.

The simulator constructs households from this population such that 15% house only retirees, and the rest have at least one working adult and are filled by randomly assigning the remaining children, adults, and retirees. To simulate informal social interactions, households may attend social events twice a month, subject to limits on gathering sizes.

At the end of each simulated day, the person's infection state is updated through a stochastic model based on all of that individual's interactions during the day (see next section). Unless otherwise prescribed by the government, when a person becomes ill they follow their routine. However, even the most basic government interventions require sick people to stay home, and at-risk individuals to avoid large gatherings. If a person becomes critically ill, they are admitted to the hospital, assuming it has not reached capacity.

3.3 SEIR Infection Model

PandemicSimulator implements a modified SEIR (susceptible, exposed, infected, recovered) infection model, as shown in Figure 2. See Table 2 for specific parameter values and the transition probabilities of the SEIR model. Once exposed to the virus, an individual's path through the disease is governed by the transition probabilities. However, the transition from the susceptible state (S) to the exposed state (E) requires a more detailed explanation.

At the beginning of the simulation, a small, randomly selected set of individuals seed the pandemic in the latent non-infectious, exposed state (E). The rest of the population starts in S. The exposed individuals soon transition to one of the infectious states and start interacting with susceptible people. For each susceptible person i, the probability they become infected on a given day, $P_i^{S \to E}(day)$, is calculated based on their contacts with

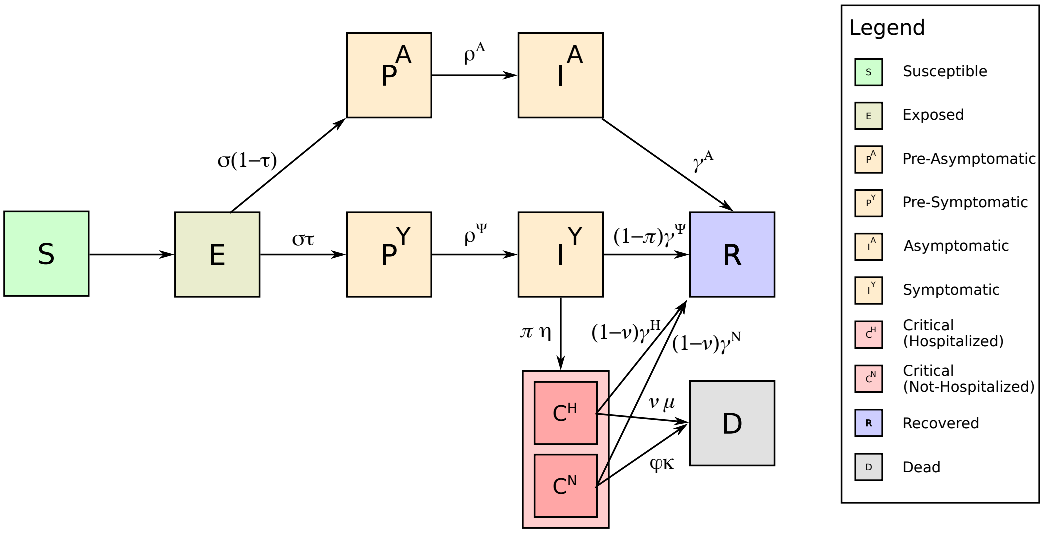


Figure 2: SEIR model used in PandemicSimulator

infectious people that day.

$$P_i^{S\to E}(day) = 1 - \prod_{t=0}^{23} \overline{P}_i^{S\to E}(t) \tag{1}$$

where $\overline{P}_i^{S\to E}(t)$ is the probability that person i is *not* infected at hour t. Whether a susceptible person becomes infected in a given hour depends on whom they come in contact with. Let $\mathcal{C}_i^j(t) = \{p \overset{b^j}{\sim} N_j(t) | p \in N_j^{\text{inf}}(t)\}$ be the set of infected contacts of person i in location j at hour t where $N_j^{\text{inf}}(t)$ is the set of infected persons in location j at time t, $N_j(t)$ is the set of all persons in j at time t, and b^j is a hand-set contact rate for j. To model the variations in how easily individuals spread the disease, each individual k has an infection spread rate, $a^k \sim \mathcal{N}^{\text{bounded}}(a, \sigma)$ sampled from a bounded Gaussian distribution. Accordingly,

$$\overline{P}_i^{S\to E}(t) = \prod_{k \in \mathcal{C}_i^j(t)} (1 - a^k). \tag{2}$$

3.4 Testing and Contact Tracing

PandemicSimulator enables enactment of arbitrary testing policies to identify positive cases of covid-19. In order to mimic real-world conditions, testing is based on a configurable false positive and false negative rate as well as a testing cap. A testing cap is relevant for simulating initial stages of an epidemic when testing resources are scarce. Asymptomatic and symptomatic individuals – and individuals that previously tested positive – get tested all at different configurable rates. A provided testing interface allows the user to define criteria for assigning tests. For instance, the user can specify that all untested symptomatic individuals are to be tested while extra available tests are to be randomly assigned.

The default simulation setting assigns each individual with a test every set time interval. The set time interval is different for asymptomatic, symptomatic, and previously-positive individuals. Refer to Table 1 for a listing of the testing time intervals used for simulation calibration.

The government can also implement a contact tracing strategy that tracks, over the last N days, the duration of time each pair of individuals interacted.

PandemicSimulator assumes cell-phone-based contact tracing (Kleinman & Merkel, 2020) which requires active public participation. As a result, the level of public participation can be adjusted using the *participation rate*, *pr*, hyperparameter.

When *pr* is set to 0% no contact tracing information is available. However, home and work/school addresses are still assumed to be known. In the other extreme case, when *pr* is set to 100%, full knowledge regarding interaction times during the last N days is assumed. When an individual tests positive, the default setting in PandemicSimulator will test all recent 1$^{\text{st}}$-order contacts (for cases where $pr > 0\%$) as well as their households and colleague/classmates.

3.5 Government Regulations

As discussed earlier (see Figure 1), the government announces regulations to try to control the pandemic. The government can impose the following rules:

- Social distancing: a value $\beta \in [0, 1]$ that scales the contact rates of each location by $(1 - \beta)$. 0 corresponds to unrestricted interactions; 1 eliminates all interactions.

- Stay home if sick: a boolean. When set, people who have tested positive are requested to stay at home.

- Practice good hygiene: a boolean. When set, people are requested to practice better-than-usual hygiene.

- Wear facial coverings: a boolean. When set, people are instructed to wear facial coverings.

- Avoid gatherings: a value that indicates the maximum recommended size of gatherings. These values can differ for high risk individuals and those of normal health.

- Closed businesses: A list of non-essential business location types that are not permitted to open.

These types of regulations, modeled after government policies seen throughout the world, are often bundled into progressive *stages* to make them easier to communicate to the population. Refer to Tables 1-3 for details on the parameters, their sources and the values set for each stage.

3.6 Simulation Parameters and Calibration

PandemicSimulator is a very flexible tool with which we study the propagation of the disease and the effects of various government regulations on that propagation. As such, it

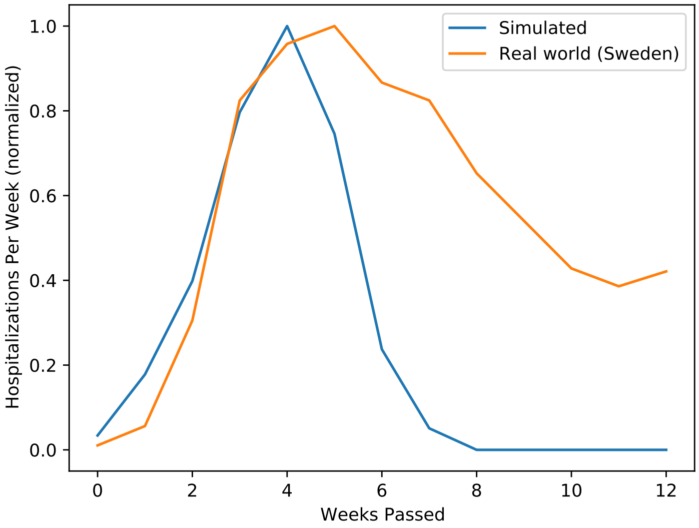


Figure 3: Hospitalization data of a simulator run with the final calibrated parameters graphed against Sweden's hospitalization data.

also has a lot of parameters that control its behaviour and that can be loosely grouped into three categories:

- Environmental: control the size of the population, the number and types of locations, and the ways people interact in those locations (Table 1).
- Epidemiological: control the progression of the disease in an individual (Table 2). These can be changed to model different pandemics.
- Regulatory: control the impacts and efficacy of the government regulations (Table 3). Regulations also determine the size of the action space for reinforcement learning experiments.

In order to build a trustworthy model, we calibrate our model to accurately track past data. We choose to calibrate the spread rate and social distancing parameter of the simulator based on our parameter sensitivity analysis (discussed in Section 5.1). Spread rate was found to have a clear linear relation with the number of hospitalizations (see Figure 7a). Similarly, modifying the contact rates at each location with a uniform multiplier also showed a linear relationship between the parameter and the spread of the virus, with the social distancing parameter representing said multiplier (see Figure 7b).

We use real-world data of Sweden as provided by the World Health Organization[2] to calibrate the simulator. Specifically, we use new hospitalizations data because it is least affected by imperfect testing strategies. We chose Sweden as our source of data as it was the nation where the least restrictions (Claeson & Hanson, 2021) were applied (at least) during

[2] `https://covid19.who.int/region/euro/country/se`

the rise of the first pandemic wave, and thus, where the dynamics of the virus is the most "natural", despite several factors influencing it (e.g, population density and mobility). The subsequent data following the peak is usually impacted by several factors that are either unknown to us or are not currently modeled in the simulator, such as, fine grained changes in people's behavior, exact regulations followed, significantly smaller simulated population sizes, etc.

In order to match real data (time-to-peak hospitalizations ≈ 10 weeks) we run a Bayesian Optimization algorithm on the spread rate mean in the range $[0.005, 0.03]$ and the social distancing rate mean in the range $[0.0, 0.8]$, resulting in a final parameter settings of spread rate $= 0.02056$ and social distancing $= 0.00198$ (see Figure 3).

4. RL for Optimization of Regulations

An ideal solution to minimize the spread of a new disease like COVID-19 is to eliminate all non-essential interactions and quarantine infected people until the last infected person has recovered. However, the window to execute this policy with minimal economic impact is very small. Once the disease spreads widely this policy becomes impractical and the potential negative impact on the economy becomes enormous. In practice, around the world we have seen a strict lockdown followed by a gradual reopening that attempts to minimize the growth of the infection while allowing partial economic activity. Because COVID-19 is highly contagious, has a long incubation period, and large portions of the infected population are asymptomatic, managing the reopening without overwhelming healthcare resources is challenging. In this section, we tackle this sequential decision making problem using reinforcement learning (RL; (Sutton & Barto, 2018)) to optimize the reopening policy.

To define an RL problem we need to specify the environment, observations, actions, and rewards.

Environment: The agent-based pandemic simulator PandemicSimulator is the environment.[3]

Actions: The government is the learning agent. Its goal is to maximize its reward over the horizon of the pandemic. Its action set is constrained to a pool of escalating stages, which it can either increase, decrease, or keep the same when it takes an action. Refer to Table 3 for detailed descriptions of the stages.

Observations: At the end of each simulated day, the government observes the environment. For the sake of realism, the infection status of the population is partially observable, accessible only via statistics reflecting aggregate (imperfect) test results and number of hospitalizations.[4]

Rewards: We designed our reward function to encourage the agent to keep the number of persons in critical condition (n^c) below the hospital's capacity ($C^{\max}$), while keeping the

[3] For the purpose of our experiments, we assume no vaccine is available and that survival rates remain constant. In practice, one may want to model the effect of improving survival rates as the medical community gains experience treating the virus.

[4] The simulator tracks ground truth data, like the number of people in each infection state, for evaluation and reporting.

Table 1: Environment Parameters (vetted by epidemiologists) used in PandemicSimulator

Parameter	**Value**	**Source**
age	Based on us population age distribution	`https://www.populationpyramid.net/united-states-of-america/2018/`
1k population parameters (number, worker capacity, visitor capacity)	Homes: (300, -, -) Grocery stores: (4, 5, 30) Offices: (5, 150, 0) Schools: (1, 40, 300) & the school has 10 classes Hospitals: (1, 30, 0) & patient capacity of 10 Retail stores: (4, 5, 30) Hair salons: (4, 3, 5) Restaurant: (2, 6, 30) Bar: (2, 5, 30) Cemetery: (1, -, -)	Hospital capacity is set based on the prescriptions from the French Red Cross on hospital building, which suggest to have between 8 and 11 hospital beds every 1000 people. Rest of the values are based on our best guess reflecting a small colony.
10k population parameters (number, worker capacity, visitor capacity)	Homes: (3000, -, -) Grocery stores: (40, 5, 30) Offices: (50, 150, 0) Schools: (10, 40, 300) & each school has 10 classes Hospitals: (10, 30, 0) & patient capacity of 100 Retail stores: (40, 5, 30) Hair salons: (40, 3, 5) Restaurant: (20, 6, 30) Bar: (20, 5, 30) Cemetery: (1, -, -)	1k population hospital capacity is scaled by 10 and also to match with 2020 data from the American Hospital Association (924,107 total beds distributed among 6,146 hospitals in the US ≈ 150 beds per hospital on an average). Rest of the values are based on our best guess reflecting a small town.
Infection spread rates (mean, standard deviation)	(0.0206, 0.01)	calibrated to fit Sweden's COVID-19 hospitalizations
Location contact rates $(b^{\text{loc}}, b^{\text{loc}}_{\text{min}})$	Homes: (0.5, 0.3, 0.3), (0, 1, 0) Grocery stores: (0.2, 0.25, 0.3), (0, 1, 0) Offices: (0.1, 0.01, 0.01), (2, 1, 0) Schools: (0.1, 0., 0.1), (5, 1, 0) Hospitals: (0.1, 0., 0.), (0, 3, 1) Retail stores: (0.2, 0.25, 0.3), (0, 1, 0) Hair salons: (0.5, 0.3, 0.1), (1, 1, 0) Restaurant: (0.3, 0.35, 0.1), (1, 1, 0) Bar: (0.7, 0.2, 0.1), (1, 1, 0) Cemetery: (0., 0., 0.05), (0, 0, 0)	verified through calibration to fit Sweden's COVID-19 hospitalizations `https://bit.ly/2X3a5jB`
Testing parameters	Random testing rate: 0.02 Symptomatic testing rate: 0.3 Critical testing rate: 1.0 False positive rate: 0.001 False negative rate: 0.01 Re-test previous-positive rate: 0.033	Based on our best guess.
Wear facial coverings (spread rate multiplier)	0.6	`https://bit.ly/2JCwSjc`
Practice good hygiene (spread rate multiplier)	0.8	Based on our best guess.
Rule compliance hour probability	0.99	
Social gatherings	House parties (duration of 5 hours) once every month in each house on random dates.	All house parties are open-invite events. This is done to represent all other gatherings like concerts, sporting events, etc.

Table 2: Epidemiological Parameters used in our SEIR model. Values given as five-element vectors are age-stratified with values corresponding to 0-4, 5-17, 18-49, 50-64, 65+ year age groups, respectively.

Parameter	**Value**	**Source**
σ: exposed rate	$\frac{1}{\sigma} \sim Tr(1.9, 2.9, 3.9)$	(Zhang et al., 2020)
τ: symptomatic proportion (%)	57	(Gudbjartsson et al., 2020b)
ρ^Y: pre-symptomatic rate	$\frac{1}{\rho^Y} = 2.3$	(He et al., 2020)
ρ^A: pre-asymptomatic rate	$\frac{1}{\rho^A} = 2.3$	
γ^Y: recovery rate in symptomatic non-treated compartment	$\frac{1}{\gamma^Y} \sim Tr(3.0, 4.0, 5.0)$	(He et al., 2020)
γ^A: recovery rate in asymptomatic compartment	$\frac{1}{\gamma^A} \sim Tr(3.0, 4.0, 5.0)$	
γ^H: recovery rate in hospitalized compartment	$\frac{1}{\gamma^H} \sim Tr(9.4, 10.7, 12.8)$	Fit to Austin admissions & discharge data (Avg=10.96. 95% CI = 9.37 to 12.76)
γ^N: recovery rate in hospitalization needed compartment	0.0214	
YHR: symptomatic case hospitalization rate (%), age and risk specific	Overall: $[0.07018, 0.07018, 4.735, 16.33, 25.54]$, Low risk: $[0.04021, 0.03091, 1.903, 4.114, 4.879]$, High risk: $[0.4021, 0.3091, 19.03, 41.14, 48.79]$	Adjusted from (Verity et al., 2020)
HFR: hospitalized fatality ratio, age specific (%)	$[4, 12.365, 3.122, 10.745, 23.158]$	Computed from the infected fatality ratio in (Verity et al., 2020)
π: rate of symptomatic individuals go to hospital	$\pi = \frac{\gamma^Y \mathrm{YHR}}{\eta + (\gamma^Y - \eta)\mathrm{YHR}}$	
η: rate from symptom onset to hospitalized	0.1695	(Tindale et al., 2020)
μ: rate from hospitalized to death	$\frac{1}{\mu} \sim Tr(5.2, 8.1, 10.1)$	Fit to Austin admissions & discharge data (Avg=7.8, 95% CI = 5.21 to 10.09)
ν: death rate on hospitalized individuals	$\nu = \frac{\gamma^H \mathrm{HFR}}{\mu + (\gamma^H - \mu)\mathrm{HFR}}$	
ϕ: death rate on individuals that need hospitalization	$[0.239, 0.3208, 0.2304, 0.3049, 0.4269]$	
κ: rate from hospitalization needed to death	0.3	

Table 3: Five stage Covid regulations

Stages	Stay home if sick, Practice good hygiene	Wear facial coverings	Social distancing	Avoid gathering size (Risk: number)	Locked locations
Stage 0	False	False	None	None	None
Stage 1	True	False	None	Low: 50, High: 25	None
Stage 2	True	True	0.3	Low: 25, High: 10	School, Hair Salon
Stage 3	True	True	0.5	Low: 0, High: 0	School, Hair Salon, Bar and Restaurant
Stage 4	True	True	0.7	Low: 0, High: 0	School, Hair Salon, Office, Retail Store, Bar and Restaurant

economy as unrestricted as possible. To this end, we use a reward that is a weighted sum of two objectives:

$$r = a \ \max\left(\frac{n^c - C^{\max}}{C^{\max}},\ 0\right) + b\ \frac{\text{stage}^p}{\max_j \text{stage}_j^p} \tag{3}$$

where stage $\in [0, 4]$ denotes one of the 5 stages with stage_4 being the most restrictive. a, b and p are set to -0.4, -0.1 and 1.5, respectively, in our experiments. To discourage frequently changing restrictions, we also use a small shaping reward (with -0.02 coefficient) proportional to $|stage(t-1) - stage(t)|$. This linear mapping of stages into a $[0, 1]$ reward space is arbitrary; if PandemicSimulator were being used to make real policy decisions, policy makers ought to use values that represent the real (estimated) economic costs of the different stages.

Table 4: Learning Parameters used in our experiments

Parameter	Value	Comment
RL critic inputs	Global infection summary, stage	Critic is only used during training.
RL actor inputs	Global testing summary, stage	To keep it realistic.
RL Actions	[-1, 0, 1]	Stage change
Critic and actor networks	2 hidden layers of 256 ReLU units each	
Simulator steps per action	24	A new action at the start of each day
Learning rates	Critic: 1e-3, actor: 1e-4	
SAC entropy coefficient α	0.01	
Stale network refresh rate	0.005	
RL discount factor	0.99	

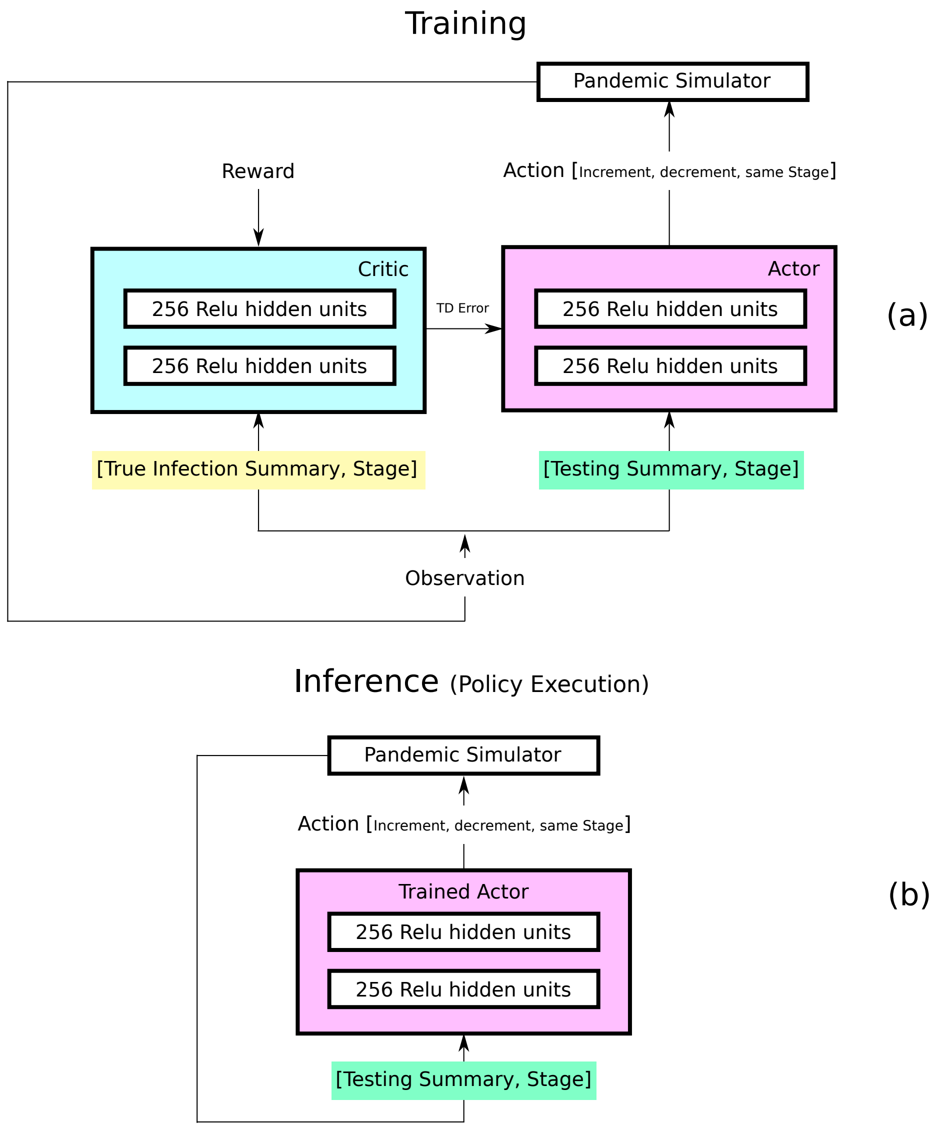


Figure 4: Control flow during training and policy execution stages while using an actor-critic RL framework. The critic receives true infection summary as inputs and is used only during training to guide learning a policy. However, the actor receives only the observable testing summary as inputs to decide whether to increment, decrement or keep the same stage for the next step.

Training

We use the discrete-action Soft Actor Critic (SAC; (Haarnoja et al., 2018)) off-policy RL algorithm to optimize a reopening policy, where the actor and critic networks are two hidden-layer deep multi-layer perceptrons with 256 hidden units (Figure 4). One motivation behind using SAC over deep Q-learning approaches such as DQN (Mnih et al., 2015) is that we can provide the true infection summary as input to the critic (Figure 4(a)) while letting

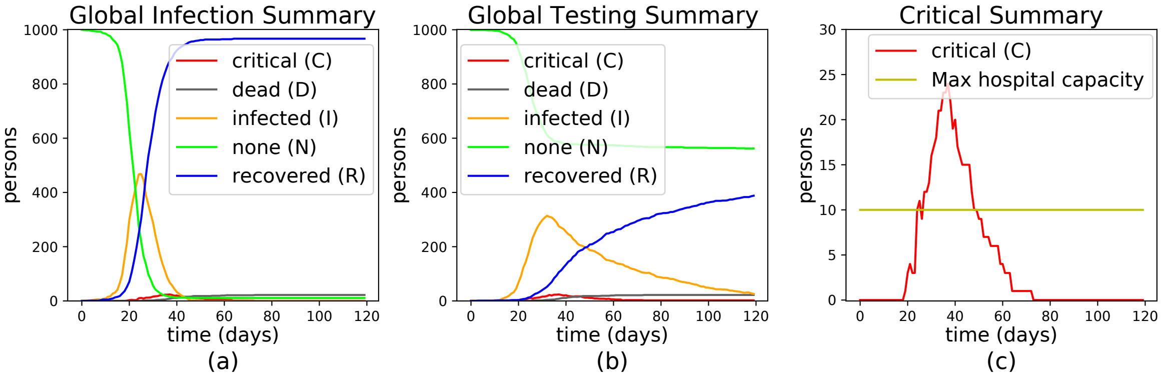


Figure 5: A single run of the simulator with no government restrictions, showing (a) the true global infection summary (b) the perceived infection state, and (c) the number of people in critical condition over time.

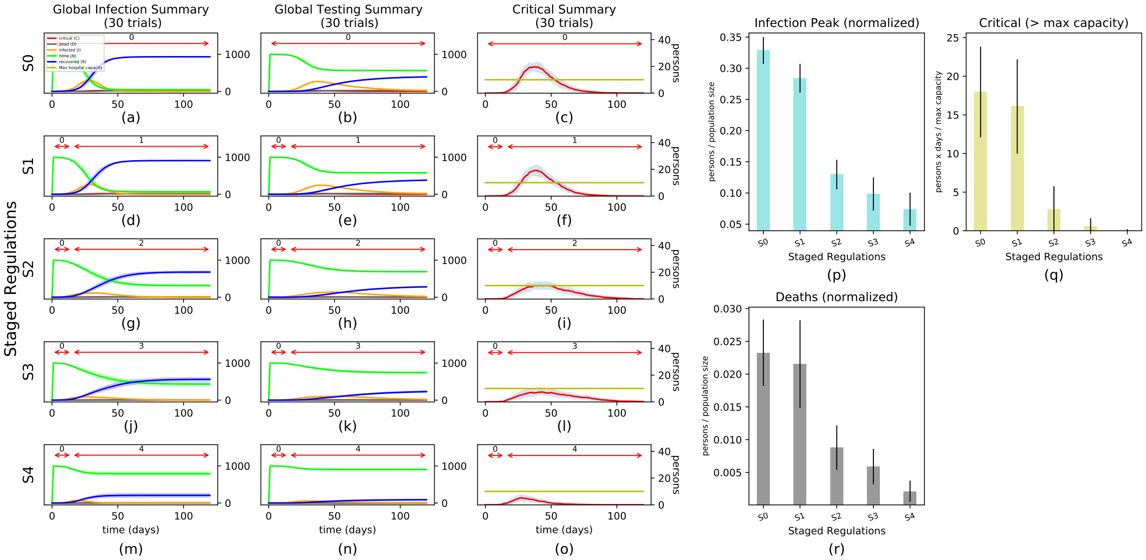


Figure 6: Simulator dynamics at different regulation stages. The plots are generated based on 30 different randomly seeded runs of the simulator. Mean is shown by a solid line and variance either by a shaded region or an error line. In the left set of graphs, the red line at the top indicates what regulation stage is in effect on any given day.

the actor see only the observed testing summary. Training is episodic with each episode lasting 120 simulated days. At the end of each episode, the environment is reset to an initial state. Refer to Table 4 for learning parameters. Once trained, only the actor is used for policy execution (Figure 4(b)), which relies only on the testing summary and current stage as inputs.

5. Experiments

The purpose of PandemicSimulator is to enable a more realistic evaluation of potential government policies for pandemic mitigation. In this section, we validate that the simulation behaves as expected under controlled conditions, illustrate some of the many analyses it facilitates, and most importantly, demonstrate that it enables optimization via RL.

Unless otherwise specified, we consider a community size of 1,000 people and a hospital capacity of 10 people.[5] To enable calibration with real data, we limit government actions to five regulation stages similar to those used by real-world cities[6] (see Section 3.6 for details), and assume the government does not act until at least five people are infected.

Figure 5 shows plots of a single simulation run with no government regulations (Stage 0). Figure 5(a) shows the number of people in each infection category per day. Without government intervention, all individuals get infected, with the infection peaking around the 25th day. Figure 5(b) shows the metrics observed by the government through the lens of testing and hospitalizations. This plot illustrates how the government sees information that is both an underestimate of the penetration and delayed in time from the true state. Finally, Figure 5(c) shows that the number of people in critical condition goes well above the maximum hospital capacity (denoted with a yellow line) resulting in many people being more likely to die. The goal of a good reopening policy is to keep the red curve below the yellow line, while keeping as many businesses open as possible.

Figure 6 shows plots of our infection metrics averaged over 30 randomly seeded runs. Each row in Figures 6(a-o) shows the results of executing a different (constant) regulation stage (after a short initial S0 phase), where S4 is the most restrictive and S0 is no restrictions. As expected, Figures 6(p-r) show that the infection peaks, critical cases and number of deaths are all lower for more restrictive stages. One way of explaining the effects of these regulations is that the government restrictions alter the connectivity of the contact graph. For example, in the experiments above, under stage 4 restrictions there are many more connected components in the resulting contact graph than in any of the other 4 cases. See Section 5.1.1 for details of this analysis.

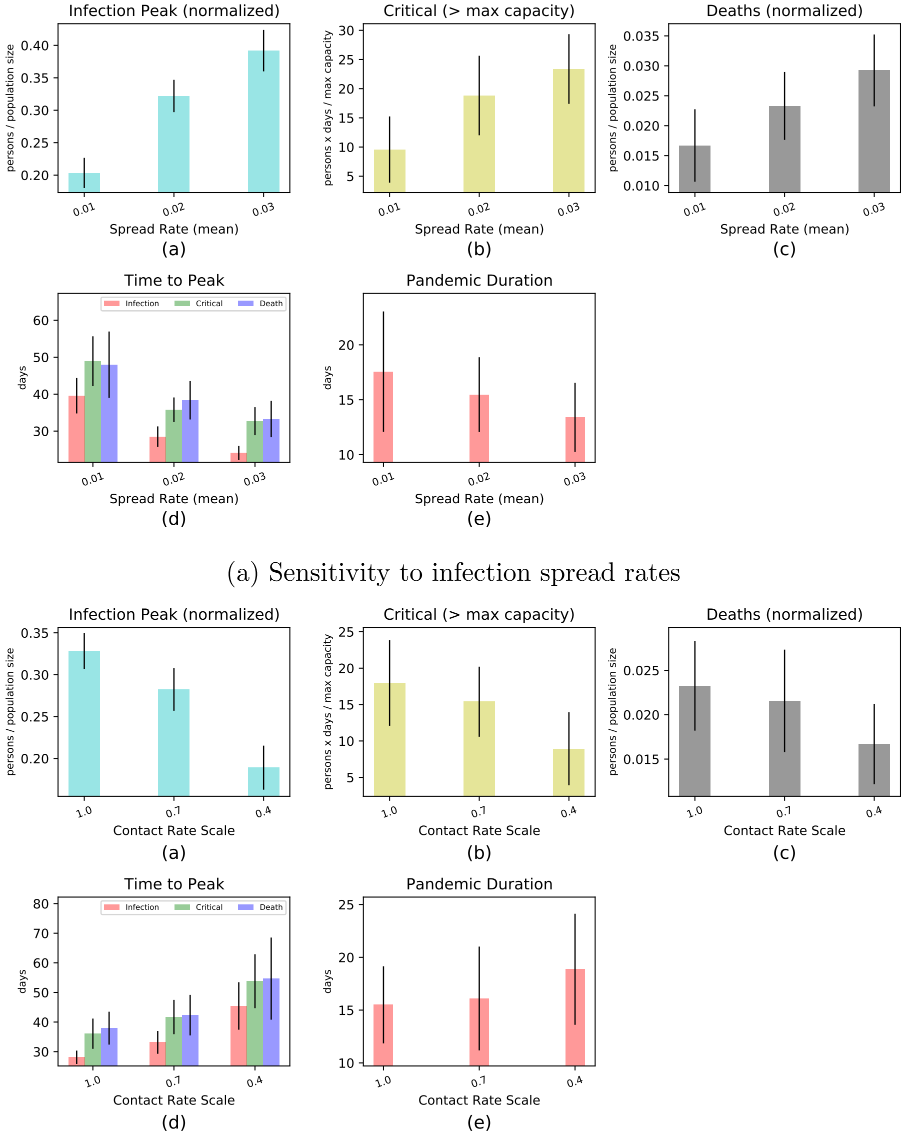
Higher stage restrictions, however, have increased socio-economic costs computed using the second objective in Eq. 3). Our RL experiments illustrate how these competing objectives can be balanced.

A key benefit of PandemicSimulator's agent-based approach is that it enables us to evaluate more dynamic policies[7] than those described above. In the remainder of this section we analyze the model's sensitivity to its parameters, we compare a set of hand constructed policies, examine (approximations) of two real country's policies, and study the impact of contact tracing. Finally, we demonstrate the application of RL to construct dynamic polices

[5] PandemicSimulator can easily handle larger experiments at the cost of greater time and computation. Informal experiments showed that results from a population of 1k are generally consistent with results from a larger population when all other settings are the same (or proportional). Refer to Table 5 for simulation times for 1k and 10k population environments. For realistic larger populations however, we would need to also incorporate venue-specific topological connectivity between locations that is not handled in this version.

[6] Such as those at `https://tinyurl.com/y3pjthyz`

[7] In this paper, we use the word "policy" to mean a function from state of the world to the regulatory action taken. It represents both the government's policy for combating the pandemic (even if heuristic) and the output of an RL optimization.

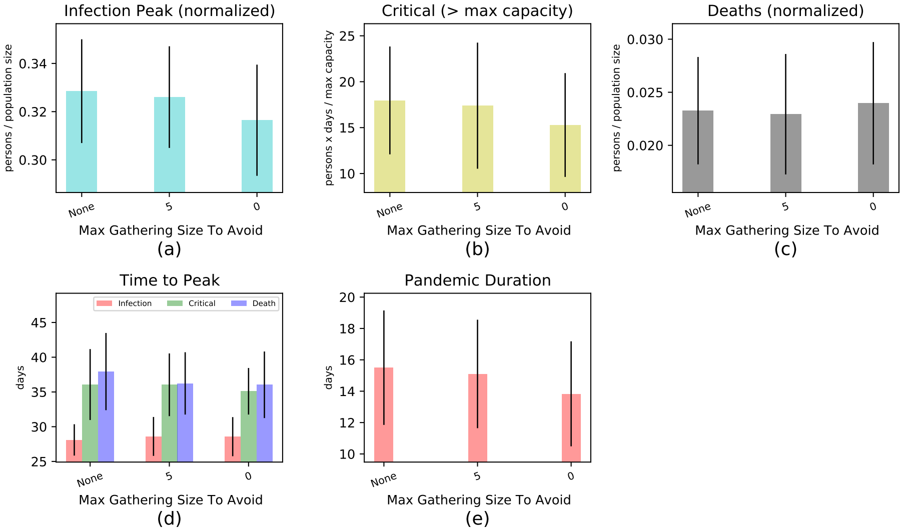


(a) Sensitivity to infection spread rates

(b) Sensitivity to location's contact rates

Figure 7: Simulator sensitivity to infection spread and contact rates. The plots are generated based on 30 different randomly seeded runs of the simulator. Mean is shown by a solid line and variance either by a shaded region or an error line.

that achieve the goal of avoiding exceeding hospital capacity while minimizing economic costs. As in Figure 6, throughout this section we report our results using plots that are generated by executing 30 simulator runs with fixed seeds. All our experiments were run on a single core, using an Intel i7-7700K CPU @ 4.2GHz with 32GB of RAM.



(a) (b) (c) (d) (e)

Figure 8: Simulator sensitivity to social gathering sizes through social events. The plots are generated based on 30 different randomly seeded runs of the simulator. Mean is shown by a solid line and variance either by a shaded region or an error line.

5.1 Sensitivity Analysis

Here we discuss a preliminary analysis of the sensitivity of PandemicSimulator to a few of its important parameters, namely: (1) spread rates for each person, (2) contact rates for each location, and (3) size of social gatherings. Specifically, we observe how uniformly scaling down or up the default values of these parameters affects the spread of the pandemic.

Figure 7a, for example, shows that increasing individual spread rates results in increased and earlier pandemic peaks. Earlier peaks also induce higher numbers of simultaneous critical cases, that easily pass the hospital capacity threshold. Consequently, we observe from our plots that spread rates are directly proportional to number of deaths in PandemicSimulator. Note that this parameter allows us to model super-spreaders on an individual basis. Thus, increasing spread rates practically means increasing the number of super-spreaders and easily infecting most of the simulator's population.

On the other hand, uniformly decreasing contact rates for all locations through a multiplier has the natural effect of reducing the spread of the pandemic, as well as the number of deaths (see Figure 7b). This effect is due to the reduction of contacts among people in the same location, which is also one the implications of social distancing. From our plots, it is possible to observe how the relation between contact rates and number of deaths / critical cases above maximum hospital capacity is non-linear. This non-linearity arises from the limited degrees of separation of our population, which is modeled as a small community where it is easy to be a 2$^{\text{nd}}$-degree contact of an infected person when contact rates are not decreased too much.

We further analyze the dynamics of PandemicSimulator as a function of the maximum allowed social gathering size. Specifically, we compare our metrics with maximum gathering size set to: none (no limit), 5, and 0 (no gatherings). As shown in Figure 8, also in this case the relation between the maximum allowed gathering size and deaths / critical cases is non-linear, for the same reasons expressed above. When compared to contact rate results, the effect of reducing gathering size is even more limited, due to the contact rates (everywhere) remaining unchanged. This observation suggests that it might not be effective to forbid social gatherings without also reducing contact rates in general and without closing other types of locations.

5.1.1 Graph Connectivity

One of the ways in which the effects of governmental restrictions can be understood is that they influence the connectivity of the contact graph among members of the population. For example, if the graph is fully connected, the average degree of separation between individuals may be increased by increasing restrictions, which presumably would then slow the spread of the disease.

To investigate the extent to which the restrictions we model influence the graph connectivity, we collected interaction data from runs corresponding to those in Figure 6. From these, we generated a graph of all interactions between the people in the simulator on each day of the simulation.

Our findings indicate that, in fact, the graph is often not connected. We therefore analyze the number of connected components in the interaction graphs during 5 separate runs, each at a different stage. The results are plotted in Figure 12.

As is apparent in the graph, on weekends (the periodic peaks in the graph), the interaction graph has many many separate components, indicating a greater degree of separation between people. Similarly during Stage 4 restrictions, the number of connected components is quite high, suggesting that lockdowns can be very effective in slowing the spread of the pandemic. Somewhat surprisingly, on weekdays at other stages, the number of connected components is relatively low, with relatively small differences between the stages.

An interesting direction for future work is to do a more in-depth graph connectivity analysis, including for interaction graphs that span multiple days.

5.1.2 Simulation Time

Although we conduct most experiments in this article on populations of 1,000 (1k) people, PandemicSimulator can easily handle larger experiments at the cost of greater time and computation. In Table 5, we report simulation times for 1k, 2k, 4k and 10k population environments. For 1k, we also report the training time it takes for our reinforcement learning algorithm to converge. While our experiments show that small population sizes are effective for learning purposes (see Section 5.4), larger simulations can be useful to analyze more complex – and possibly realistic – community interactions and connectivity graphs.

5.2 Benchmark Policies

To serve as benchmarks, we defined three heuristic policies and two policies inspired by real governments' approaches to managing the pandemic.

Table 5: Simulation time for different population sizes. The simulator was run on a single core Intel i7-7700K CPU @ 4.2GHz with 32GB of RAM.

Population size	**Simulation Time**
1k	25.4 msecs/sim-step (our RL training took about 10 hours to converge)
2k	57.9 msecs/sim-step
4k	138.5 msecs/sim-step
10k	500 msecs/sim-step

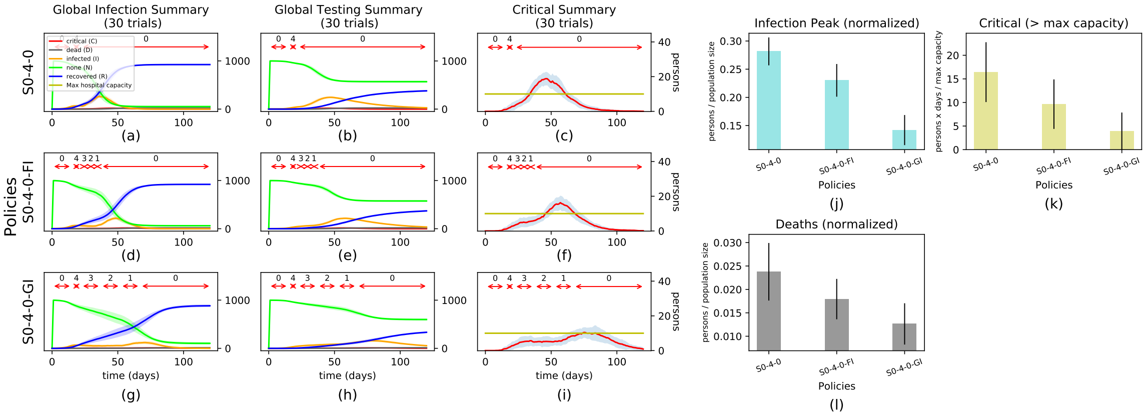


Figure 9: Simulator dynamics under different hand constructed policies.

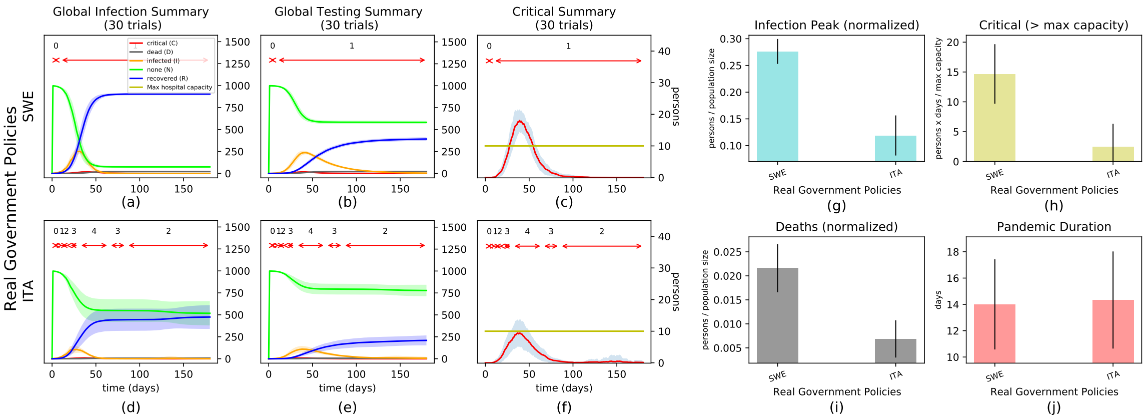


Figure 10: Simulator dynamics under Swedish and Italian government policies.

- **S0-4-0**: Using this policy, the government switches from stage 0 to 4 after reaching a threshold of 10 infected people. After 30 days, it switches directly back to stage 0;
- **S0-4-0-FI**: The government starts like S0-4-0, but after 30 days it executes a fast, incremental (FI) return to stage 0, with intermediate stages lasting 5 days;
- **S0-4-0-GI**: This policy implements a more gradual incremental (GI) return to stage 0, with each intermediate stage lasting 10 days;

- **SWE**: This policy represents the one adopted by the Swedish government, which recommended, but did not require remote work, and was generally unrestrictive.[8] Table 6 shows how we mapped this policy into a 2-stage action space.
- **ITA**: This policy represents the one adopted by the Italian government, which was generally much more restrictive.[9] Table 7 shows our mapping of this policy to a 5-stage action space.

Table 6: Swedish Covid regulations. Note that, while the Swedish government recommended, but did not require, remote work and had different recommendations for different ages of school children, we mapped the overall policy to be roughly stage 1 reported here.

Stages	**Stay home if sick, Practice good hygiene**	**Wear facial coverings**	**Social distancing**	**Avoid gathering size (Risk: number)**	**Locked locations**
Stage 0	False	False	None	None	None
Stage 1	True	False	0.00198	Low: 50, High: 50	None

Table 7: Italian Covid regulations

Stages	**Stay home if sick, Practice good hygiene**	**Wear facial coverings**	**Social distancing**	**Avoid gathering size (Risk : number)**	**Locked locations**
Stage 0	False	False	None	None	None
Stage 1	True	False	0.1	None	None
Stage 2	True	False	0.2	None	School
Stage 3	True	True	0.5	Low: 0, High: 0	School, Hair Salon, Retail Store, Bar and Restaurant
Stage 4	True	True	0.7	Low: 0, High: 0	Office, School, Hair Salon, Retail Store, Bar and Restaurant

Figure 9 compares the heuristic policies. From the point of view of minimizing overall mortality, S0-4-0-GI performed best. In particular, slower re-openings ensure longer but smaller peaks. While this approach leads to a second wave right after stage 0 is reached, the gradual policy prevents hospital capacity from being exceeded.

Figure 10 also contrasts the approximations of the policies employed by Sweden and Italy in the early stages of the pandemic (through February 2020). The ITA policy leads to fewer deaths and only a marginally longer duration. However, this simple comparison does not account for the economic cost of policies, an important factor that is considered by decision-makers.

[8] https://tinyurl.com/y57yq2x7; https://tinyurl.com/y34egdeg

[9] https://tinyurl.com/y3cepy3m

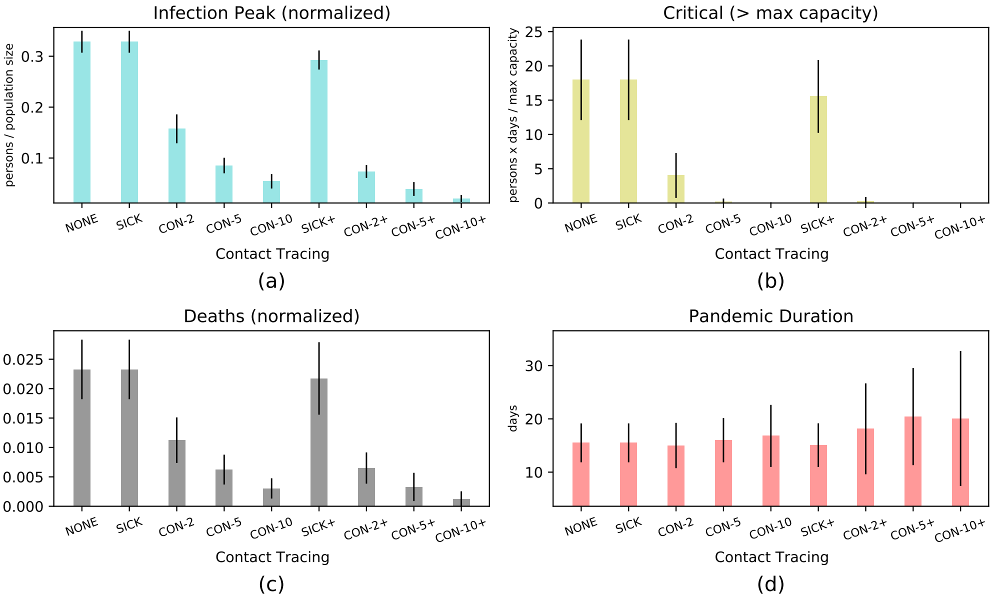


Figure 11: Comparison of various combinations of testing and contact tracing.

5.3 Testing and Contact Tracing

To validate PandemicSimulator's ability to model testing and contact tracing we compare several strategies with different testing rates and contact horizons. Specifically, we consider daily testing rates of {0.02 and 0.3 (denoted with a + symbol in our plots)} and contact tracing histories of {0, 2, 5, or 10} days. Note that, in our plots, both NONE and SICK use 0 contact tracing history (the second with self-isolation at symptom onset), while CON-N uses an N length history. The full specification of each parameter combination as well as the mapping to the label in Figure 11 is shown in Table 8. For each condition, we ran the experiments with the same 30 random seeds.

Not surprisingly, contact tracing is most beneficial with higher testing rates and longer contact histories because more testing finds more infected people and the contact tracing is able to encourage more of that person's contacts to stay home. In fact, from our experiments we observe that contact tracing, even with a small random testing rate, becomes very effective when the contact history is high (CON-5, CON-10). In this case, in fact, results in terms of death and critical cases are comparable to CON-2+ and CON-5+, where the testing rate is significantly increased.

5.4 Optimizing Reopening using RL

A major design goal of PandemicSimulator is to support optimization of re-opening policies using RL. In this section, we test our hypothesis that a learned policy can outperform the benchmark policies. Specifically, RL optimizes a policy that (a) is adaptive to the changing infection state, (b) keeps the number of critical patients below the hospital threshold, and (c) minimizes the economic cost.

Table 8: Testing and contact tracing policies.

Name	Contact tracing history (days)	Random testing rate (daily)	Stay home if sick	Stay home if positive contact
NONE	0	0.02	NO	NO
SICK	0	0.02	YES	NO
CON-2	2	0.02	YES	YES
CON-5	5	0.02	YES	YES
CON-10	10	0.02	YES	YES
SICK+	0	0.3	YES	NO
CON-2+	2	0.3	YES	YES
CON-5+	5	0.3	YES	YES
CON-10+	10	0.3	YES	YES

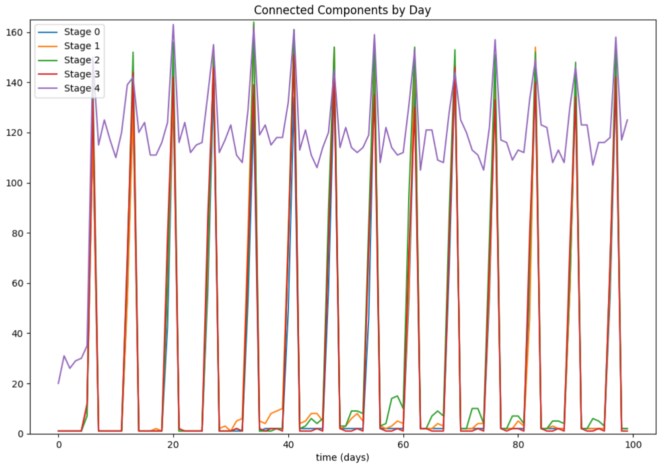


Figure 12: Graph connectivity over 5 different runs, each at a different stage.

We ran experiments using the 5-stage regulations defined in Table 3; trained the policy by running RL optimization for roughly 2 million training steps; and evaluated the learned policies across 30 randomly seeded initial conditions. Figures 13(a-f) show results comparing our best heuristic policy (S0-4-0-GI) to the learned policy. The RL optimized policy (Opt_daily) is better across all metrics as shown in Figures 13(m-p). Further, we can see how the learned policy reacts to the state of the pandemic; Figure 13(f) shows different traces through the regulation space for 3 of the trials. The learned policy briefly oscillates between Stages 2 and 3 around day 40. To minimize such oscillations, we evaluated the policy at an action frequency of one action every 3 days (bi-weekly; labeled as Opt_biweekly) and every 7 days (weekly; labeled as Opt_weekly). Figure 13(p) shows that both variants perform equally well. To test robustness to scaling, we also evaluated the learned policy (with daily actions) in a town with a population of 10,000 (Opt_10x) and found that the results transfer well. This success hints at the possibility of learning policies quickly even when intending to transfer them to large cities.

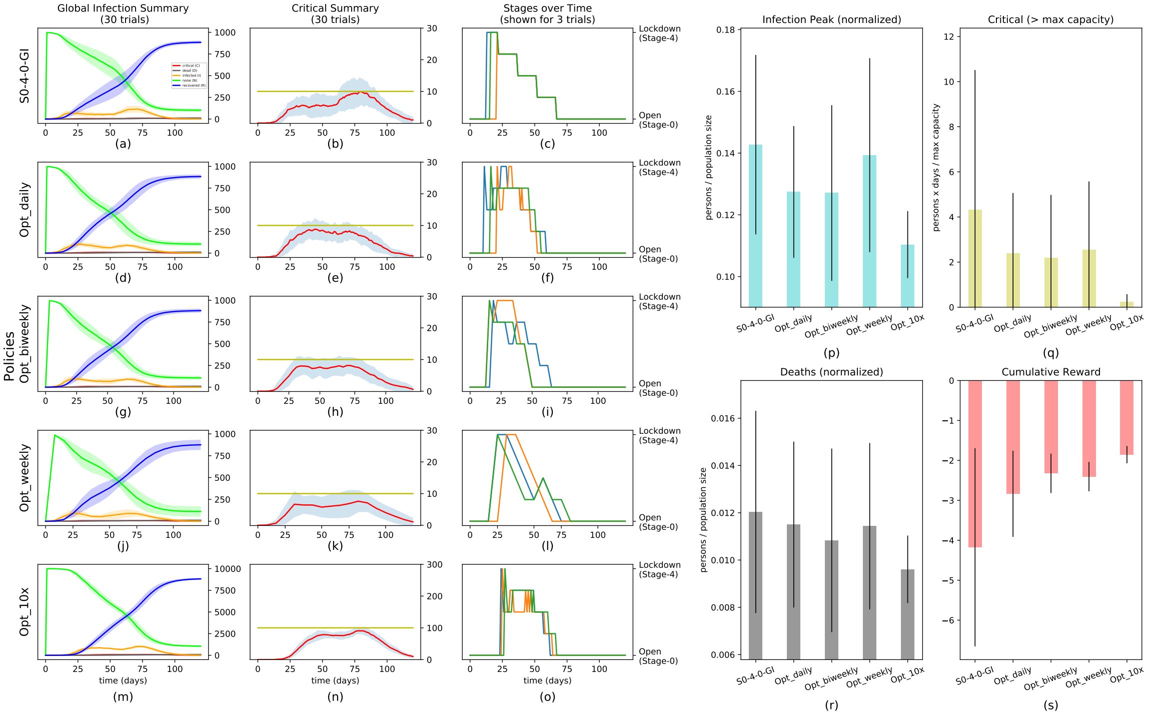


Figure 13: Simulator runs comparing the S0-4-0-GI heuristic policy with a learned policy evaluated at different action frequencies (daily, biweekly and weekly) and in a larger population (10x) environment.

This section presented results of applying RL to optimize reopening policies. An interesting next step would be to study and explain the learned policies as simpler rule based strategies to make it easier for policy makers to implement. For example, in Figure 13(f), we see that the RL policy waits at stage 2 or 3 before reopening schools to keep the second wave of infections under control. Whether the wait is specific to school reopening or is a result of a combined effect of multiple triggers, is one of many interesting questions that this type of simulator allows us to investigate.

6. Individual Infection Likelihood

Probabilistic knowledge regarding who is currently infected can be useful for applying targeted policies. For instance, isolating a subset of the population that is affiliated with higher infection probabilities can mitigate the epidemic progression similarly to a complete lockdown while imposing a smaller economic burden. In this section, we introduce a novel method, based on Hidden Markov Models, for identifying the individuals in the population with the highest likelihood of being infected.

6.1 Problem Definition: Individual Infection Likelihood

We consider a scenario where a population, S, is exposed to an infectious disease. At every time step, t, a subset of the population, $I_t \subset S$, is infected. The median individual infection period (in days) is known, and denoted d. We define an infection event for individual s_i at time t by $s_i \in I_{t+1}$ and $s_i \notin I_t$. Once an individual is infected, they will remain so (and infectious to others) until they recover. Each individual can be tested at any time step to determine if they are infected. The test false positive and false negative rates are known and denoted F^+_{test} and F^-_{test} respectively. Every individual can also be showing symptoms (or not) at every time step. The probability of showing symptoms and not being infected is known and denoted F^+_{symp}. Similarly, the probability of showing no symptoms and being infected is known and denoted F^-_{symp}. We define the *Infection Probability Inference* problem as follows.

Goal: compute an infection belief state over the entire population. That is, compute a vector of probabilities, $B = \mathbb{R}^{|S|}$, where $B[i]$ represents the probability that individual i is currently infected (and infectious).

Input: The following observations are provided at every time step, t:

- Test results for a subset of the population.
- The existence of symptoms for each individual.
- Contact graph as a symmetric matrix, $C_t = \mathbb{R}^{|S|\times|S|}$. $C_t[i, j]$, represents the transmission probability between individuals i and j during time step t.

In a real-world scenario some of these inputs might be unknown. For instance, some individuals might refuse to report existing symptoms or who they were in contact with. We consider such partial observability in the empirical study.

Desiderata: The proposed solution should **(a)** maximize prediction accuracy regarding infected individuals, and **(b)** avoid storing the full contact, testing, and symptoms history due to computational and memory limitations as well as privacy concerns.

6.2 Individual Infection Likelihood Inference

We address the infection likelihood inference problem as a hidden Markov model (Eddy, 2004) where the infection probabilities define the belief space. At each time step, the infection probabilities are updated according to test results, symptoms, and contact observations. At each state transition the infection probabilities are updated according to a recovery probability.

Applying the Markov property to the affiliated belief space seems counter-intuitive since updating the infection probabilities for the current state impacts the infection probabilities in previous time steps. For example, if s_i tested positive today, the infection probability for its previous physical contacts should be increased. In order to maintain the Markov property, we include a compressed representation of the contact history in each state. Such a compressed representation also complies with the desired feasibility and privacy restrictions that prohibit explicitly storing the contact history. Algorithm 1 details our proposed solution.

Algorithm 1: Infection Likelihood Inference

```
   Input: daily contacts, test results, and symptoms report
   Result: Daily individual infection probabilities, $B$
1  Initialization:
2      Init belief state as a vector of probabilities: $B = \mathbb{R}^{|S|}$ where
       $\forall s_i \in S$ , $B[i] = |I|/|S|$;
3      Set the decay rate based on the infection median length ($d$): $\gamma = \sqrt[d]{0.5}$;
4      Init double decay contact history as a symmetric matrix of probabilities:
       $C_{\gamma^2} = \mathbb{R}^{|S|\times|S|}$;
5      Init triple decay contact history as a symmetric matrix of probabilities:
       $C_{\gamma^3} = \mathbb{R}^{|S|\times|S|}$;
6  foreach step t, (day) do
7      Decay the infection probabilities: $B = \gamma B$;
8      Update contact matrices based on today's reported contacts ($C_t$):
       $C_{\gamma^2} = \gamma^2 C_{\gamma^2} + C_t$ and $C_{\gamma^3} = \gamma^3 C_{\gamma^3} + C_t$;
9      foreach individual, $s_i \in S$ do
10         If tested positive, then: $B[i] = 1 - (1 - B[i])\, F^{+}_{test}$ ;
11         Else if tested negative, then: $B[i] = B[i] F^{-}_{test}$;
12         If showing symptoms, then: $B[i] = 1 - (1 - B[i])\, F^{+}_{symp}$;
13         Else (no symptoms), then: $B[i] = B[i] F^{-}_{symp}$ ;
14     end
15     foreach individual, $s_i \in S$ do
16         $B_{next}[i] = 1 - \prod_j \left(1 - B[j]\left(C_{\gamma^2}[i,j] - B[i] C_{\gamma^3}[i,j]\right)\right)$;
17     end
18     Normalize $B_{next}$ ;
19     $B = B_{next}$;
20 end
```

At every time step, three data-fields are updated and stored per individual, s_i. These are, infection probability ($B[i]$), double decayed contact history ($C_{\gamma^2}[i]$), and triple decayed contact history ($C_{\gamma^3}[i]$). Each of the decayed contact histories is stored as a symmetric matrix with an entry per (unordered) pair of individuals in the community. We denote these matrices as double and triple decayed contact histories since they are decayed by factors γ^2 and γ^3 respectively. The intuition behind the need for these decay factors is not straightforward. This need is derived from the mathematical representation of the problem under a set of assumptions that are discussed in the "Theoretical Analysis" section. Both contact matrices are initialized as a zero matrix in Lines 4 and 5. The diagonals of both contact matrices are set to a constant at $C_{\gamma^2}[i,i] = 1$ and $C_{\gamma^3}[i,i] = 0$. The specified data-fields define the state space in the affiliated HMM representation. A decay rate, γ, is set such that an initial probability of 1 would decay to 0.5 after d days (Line 3). This decay rate represents a recovery probability of 0.5 by the median infection period (d). Such a decay rate assumes a constant per day probability of recovering (see Assumption 1 in Section 6.3) Line 16 updates both (double and triple) decayed contact history to include the contacts reported for the current time step (C_t). Recall that the entries of matrices C_{γ^2}, C_{γ^3}, and C_t represent probabilities. As a result, every entry is capped at 1.

The individual infection probabilities are updated based on reported test results and symptoms observations. For instance, if individual s_i tested positive at the current time step, we update its infection probability to equal the complementary event for both not-being infected **and** falsely testing positive (Line 10). Note that this algorithm makes a simplifying assumption that test observation and symptoms observation are conditionally independent given infection. If this assumption is violated, as suggested for COVID-19 (Grushka-Cohen et al., 2020), then lines 10-13 should specify unique and mutually excluding cases per outcome combination. For example, **if** s_i tested positive **and** is showing symptoms, then $B[i] \leftarrow 1-(1-B[i])F^{++}_{test\&symp}$ where $F^{++}_{test\&symp}$ is the probability of not being infected when both test results and symptoms indicate infection (false positive-positive rate). Similarly, $F^{+-}_{test\&symp}$, $F^{-+}_{test\&symp}$, and $F^{--}_{test\&symp}$ will need to be defined.

Next, infection probabilities are updated according to both decayed contact matrices (Line 16). Specifically, Line 16 computes the probability that s_i was infected by some individual s_j over the past days and did not recover since. The derivation of this update formula is provided later in the "Theoretical Analysis" section. Setting the diagonals of the two contact matrices $C_{\gamma^2}, C_{\gamma^3}$ as, ones and zeros respectively results in self infection probability of 1 from the previous day. That is, unless recovered, an infected individual will remain infected. The reader can verify that for these values, $B[i](C_{\gamma^2}[i,i] - B[i]C_{\gamma^3}[i,i])$ results in $B[i]$ (Line 16). Note that, Lines 15-16 can, and should, be computed more efficiently using matrix operations. The iterative form is provided for ease of presentation.

Finally, the set of probabilities is normalized to fit the estimated disease spread in the community (Line 18). Specifically, B is scaled such that $\sum B = |\hat{I}|$ where $|\hat{I}|$ is the estimated number of infected individuals. We assume that I can be evaluated using positivity test rates and random serological/PCR tests.

6.3 Theoretical Analysis

The following simplifying **assumptions** are used for justifying the update rule in Line 16 of Algorithm 1.

1. A constant per day recovery probability $(1-\gamma)$ for infected individuals.
2. If s_i was infected on some day then it cannot get infected on subsequent days (events of infection are mutually exclusive over days).
3. The probability that any individual was previously infected and recovered is practically zero.
4. For any individual, s_i, the daily a priori infection probability is equal between past days.

The reader should note that, in many real-world scenarios, these assumptions are not guaranteed to hold. For example, evidence regarding the COVID-19 pandemic (Lauer et al., 2020) do not support Assumption 1. Furthermore, Assumption 3 is mainly relevant during the initial stages of the outbreak. Nonetheless, the reader should keep in mind that the early stages of the outbreaks are exactly those where inference is most important as it enables reducing the maximal number of concurrent active cases (the epidemic spread peak) by applying better confinement strategies.

Let p_i^t be the probability that individual s_i is infected on day t. Similarly, p_i^{t-k} is the probability the individual s_i was infected k days before t. Following Assumption 1, we get:

Proposition 1. *The probability that s_j infected s_i on day $t-k$ and s_i did not recover since is:*

$$p_j^{t-k}C_{t-k}[i,j]\gamma^k \tag{4}$$

When considering Assumption 2, Proposition 1 must be updated to include the requirement that s_i was not already infected at day $t-k$.

Proposition 2. *The probability that s_j infected s_i on day $t-k$ **and** s_i was not already infected **and** s_i did not recover since is:*

$$p_j^{t-k}C_{t-k}[i,j](1-p_i^{t-k})\gamma^k \tag{5}$$

Note that Proposition 2 does not take into account a case where s_i was infected and fully recovered before day $t-k$. Recall that such scenarios have a probability of 0 according to Assumption 3. Also note that Proposition 2 defines infection events that are mutually exclusive over days. It is well known that the probability that no mutually exclusive event happens is one minus the sum of the events probabilities. Following Proposition 2, we can write the probability that individual i is currently infected as one minus the probability that no unrecovered infection occurred between s_i and any other individual, s_j, at any past day, k. And so:

Proposition 3. *The probability that individual i is currently (time step t) infected is:*

$$1-\prod_j\left(1-\sum_k p_j^{t-k}\gamma^k C_{t-k}[i,j](1-p_i^{t-k})\right) \tag{6}$$

Lemma 1. *Assumptions 1-4 imply:*

$$p_i^{t-k}=p_i^t\gamma^k \tag{7}$$

Proof. Let p_i^t+ represent the event where s_i is infected at time step t. Similarly, p_i^t- represent the event where s_i is **not** infected at time step t. Let $P(A|B)$ be the conditional probability of event A given event B.

$$p_i^{t-k}=^{(1)}p_i^t\cdot P(p_i^{t-k}+|p_i^t+)+(1-p_i^t)(p_i^{t-k}+|p_i^t-)$$

$$=^{(2)} p_i^t \cdot P(p_i^{t-k}+|p_i^t+) =^{(3)} p_i^t \cdot P(p_i^t+|p_i^{t-k}+) =^{(4)} p_i^t \gamma^k$$

$=^{(1)}$ by definition.
$=^{(2)}$ follows from Assumption 3.
$=^{(3)}$ follows from Bayes Theorem and Assumption 4.
$=^{(4)}$ follows from Assumption 1. □

Combining Proposition 3 with Lemma 1, we get:

Proposition 4.

$$\begin{aligned} p_i^{t+1} &= 1 - \prod_j \left(1 - \sum_k p_j^t \gamma^k \gamma^k C_{t-k}[i,j](1 - p_i^t \gamma^k)\right) \\ &= 1 - \prod_j \left(1 - p_j^t \left(\sum_k \left(\gamma^{2k} C_{t-k}[i,j]\right) - p_i^t \sum_k \left(\gamma^{3k} C_{t-k}[i,j]\right)\right)\right) \end{aligned} \tag{8}$$

$\sum_k \gamma^{2k} C_{t-k}[i,j]$ is an exponentially decayed moving average that is stored as C_{γ^2} in Algorithm 1. That is, there is no need for explicitly storing the full contact history. The same goes for $\sum_k \gamma^{3k} C_{t-k}[i,j]$ that is stored as C_{γ^3}.

The reader can verify that Line 16 from Algorithm 1 follows from Equation 8 in Proposition 4.

6.4 Empirical Study

To complement our theoretical analysis, we evaluate the effectiveness of the proposed approach via experiments in our custom-built agent-based pandemic simulator. Note that in the simulator, the simplifying assumptions upon which the theoretical analysis relied do not hold. Namely, the recovery probability is not constant (in contrast to Assumption 1), but rather a function of the infection duration and individual attributes; the probability that any individual was previously infected and recovered grows as time progresses (in contrast to Assumption 3); and for any individual, the daily a priori infection probability changes as a function of contact with infected individuals (in contrast to Assumption 4).

Our empirical study addresses the following questions.

1. Can the proposed inference approach proactively identify infected individuals better than existing approaches?

2. When comparing to existing approaches, can the proposed inference approach reduce epidemic progression when combined with a simple testing and quarantine policy?

3. How do different levels of observability regarding contact tracing affect the efficiency of the proposed approach?

The reported results support the following answers: yes, yes, and better contact tracing leads to better inference.

6.4.1 Experimental Settings

For the following set of experiments, three levels of contact tracing are considered.

1. **Passive tracing** - home and work/school addresses are known.

2. ***x*% tracing** - $x\%$ of the population are actively traced (e.g., by a relevant cellphone app). When two such individuals occupy the same building, a contact event is registered along with the contact duration.

3. **Active tracing** - denotes 100% tracing.

For passive tracing, the daily reported contacts $C_t[i, j]$ was set to equal 0 and then +0.5 if s_i, s_j shared the same house and +0.1 for sharing the same school/work place. For active tracing, $C_t[i, j]$ was set proportional to the contact length between s_i and s_j in hours divided by 24, i.e., the fraction of time steps that they were in the same location. For $x\%$ tracing, $C_t[i, j]$ was set according to the active tracing rule if both s_i and s_j are actively traced; otherwise it was set according to the passive tracing rule. Not that tracing is separate from testing. In active tracing, 100% tracing only provides full information on contacts without testing the course of the disease is unknown.

For these experiments we followed the 4,000 person town parameters, which directly scales the 1k population parameters in Table 1 by a factor of 4. Members of the population move between their homes, school, work, and their leisure activities as usual. Individuals designated for quarantine temporarily cease contact with others. Decay rate is set based on a median infection length of 5 days. The same false positive and false negative rates listed for testing parameters in Table 1 were used. Further, it was assumed that not all of those whom are symptomatic would report their status. In these experiments 30% of the population dutifully reported their symptomatic status.

The false positive symptomatic rate (showing symptoms yet are not infected), F^{+}_{symp}, was set to 0.0655 following the average workdays loss ratio (pre COVID-19) due to sickness in Japan (Chimed-Ochir et al., 2019). The false negative symptomatic rate (infected but asymptomatic), F^{-}_{symp}, was set to 0.6 following the "Current Best Estimate" (September-8, 2020) of the US Centers for Disease Control and Prevention (CDC, 2020).

6.4.2 Baseline

Our baseline for comparison follows the risk score method presented by Grushka et al. (Grushka-Cohen et al., 2020) for ranking COVID-19 positive testing probabilities. According to the reported correlations, the following ranking is inferred (lower rank number implies higher infection probability).

1. Individuals who tested positive.

2. Individuals who were in contact with a confirmed case during the last 7 days (exposed) **and** are showing symptoms.

3. Exposed individuals.

4. Individuals showing symptoms.

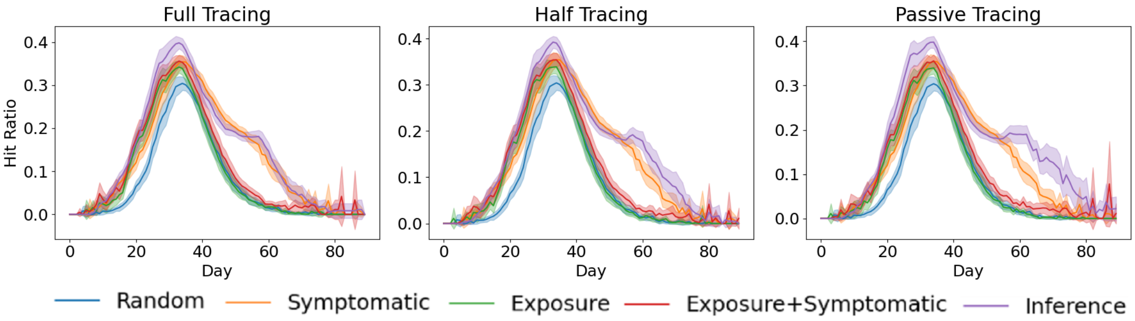


Figure 14: Hit-ratio as a function of time. Shaded areas represent 95% confidence interval over 30 trials.

5. All others.

Each individual in the community is assigned to the highest rank (where 1 is the highest and 5 the lowest) that fits its status. That is, if an individual is both showing symptoms and was exposed, it is assigned to rank 2. We consider four unique baselines that are derived from the above categories.

- **Exposure+symptoms**, infection probability ordering follows the above ranking as suggested by Grushka et al. i.e., $1 \prec 2 \prec 3 \prec 4 \prec 5$.
- **Exposure**, infection probability ordering follows the ranking $1 \prec (2 \equiv 3) \prec 4 \prec 5$.
- **Symptomatic**, infection probability ordering follows the ranking $1 \prec (2 \equiv 4) \prec 3 \prec 5$.
- **Random,** infection probability ordering follows no ranking (excluding those tested positive), i.e., $1 \prec (2 \equiv 3 \equiv 4 \equiv 5)$.

When querying for the n most probable infected individuals, each baseline method returns the n highest ranked individuals while breaking ties randomly.

6.4.3 Inference Accuracy

The first set of experiments aims to address research question #1: Can the proposed inference approach proactively identify infected individuals better than the baseline methods?

In order to allow fair comparison between the baselines and the inference approach, no quarantine operations were used and the testing policy was purely random (sampling 1% of the population each day). The baselines and the inference approach were provided the exact same information (test results and contact tracing) within the exact same run. Doing so allowed us to compare how accurately each approach managed to guess the subset of infected individuals. It is important to note that the compared prediction approaches did not influence the simulation progression in any way (they simply observed and reported predictions).

Let I_t be the set of actively infected individuals at day t. Let $S^n(B_t)$ be the set of n individuals with the highest infection probability according to belief state B_t. Define hit-ratio for day t as $\frac{|I_t \cap S^n(B_t)|}{|I_t|}$ with $n = |I_t|$.

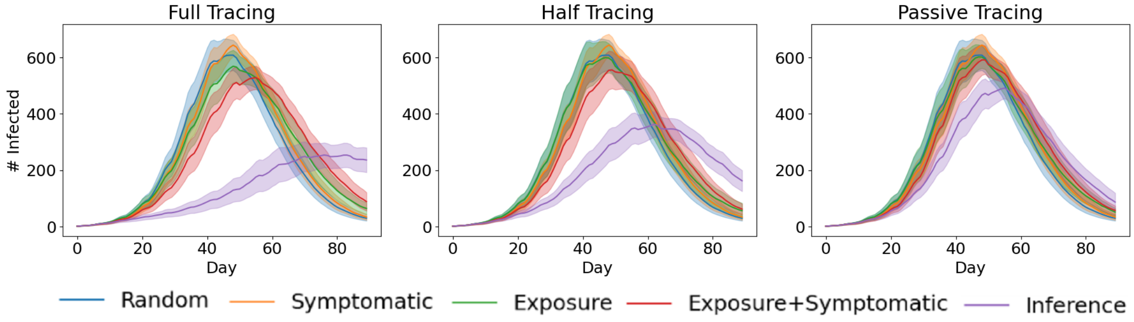


Figure 15: Number of actively infected individuals as a function of time when quarantined following various policies. Shaded areas represent 95% confidence interval over 30 trials.

Figure 14 presents the hit-ratio over time for our inference approach and the four baseline approaches. Three scenarios are considered with regards to contact tracing, namely, active, 50%, and passive. Note that over different contact tracing scenarios (between the subfigures) the 'Random' curve is showing the exact same trend and this is also true for the 'Symptomatic' curve. Neither the 'Random' or 'Symptomatic' approaches consider contact tracing, so all the scenarios are the same from their perspective. However, several trends regarding our proposed inference approach can be observed.

First, in all scenarios and all time steps the proposed inference approach performs at least as well as all the baseline approaches (equal or higher hit-ratios). Nonetheless, even small advantage in prediction accuracy can result in significant advantage once combined with a suitable testing and quarantining policy as shown in the next section.

Another observed trend is the significant advantage for the 'Symptomatic' and 'Inference' curves after the infection count peak (around day 37). Blindly prioritizing exposed individuals ('Exposure', 'Exposure+Symptomatic') does not perform well in such cases due to "herd immunity", i.e., most contacts are with recovered individuals who do not get infected. However, the reader should note that post-peak infection prediction has little to no impact when aiming to "flatten the curve," i.e., reduce the peak's magnitude with respect to number of infected individuals.

6.4.4 Impact on Test and Quarantine Policies

The second set of experiments aims to address research question #2: Can the proposed inference approach reduce epidemic progression when combined with an appropriate testing and quarantine policy?

A simple testing policy was implemented where tests are assigned to the most probable infected individuals using either our inference approach or the baseline approaches. The number of tests per day was set to 1% of the population or 40 in total. A complimentary quarantine policy was implemented where the most probable infected individuals were isolated and had no active contacts in successive days. For the *Random* baseline, those tested positive were isolated. Isolation lasts 14 days after which normal behavior is resumed. Isolated individuals are not considered for being tested. In order to for allow a fair comparison, the number of individuals that are sent to be isolated per day is capped at 2%

Table 9: Maximal number of infected individuals in a single day for different caps of tests and quarantine orders per day. Note that quarantine orders are in effect for 14 days so 1% orders per day can accumulate to 14% of the population being quarantined simultaneously. Asterisk in front of a value denotes a 95% statistically significant difference over 30 trials.

	Active tracing					**Passive tracing**				
Test cap (%)	0.5	1	2	4	10	0.5	1	2	4	10
	cap 1% quarantine/day									
Baseline	896	869	736	757	498	913	912	847	844	680
Improvement ratio	*1.04	1.09	1.12	*1.29	*1.82	*1.01	1.05	*1.10	*1.08	*1.14
	cap 2% quarantine/day									
Baseline	597	529	387	314	78	615	591	539	484	267
Improvement ratio	*1.58	*2.08	*3.83	*15.7	*4.59	*1.26	*1.20	*1.35	*1.44	*2.78
	cap 4% quarantine/day									
Baseline	16	16	13	14	9	18	17	15	17	10
Improvement ratio	*1.23	1.6	1.08	1.27	0.90	1.06	*1.06	*1.07	1.30	1.11

of the simulated population (80 individuals). Note that more than 80 individuals can be isolated simultaneously if they were initially sent to isolation on different days.

Figure 15 presents the number of infected individuals over time for our inference approach and the four baseline approaches. As in Figure 14, three scenarios are considered and presented regarding the contact tracing, namely, active, 50%, and passive. The symptomatic baseline achieves a higher hit rate than random in Figure 14, but lower effect than random in Figure 15. This is due to in part infected individuals being capable of spreading the disease prior to becoming symptomatic. Symptomatic testing also has a reduced effect because it can not remove people likely to become sick. Random testing has a non-zero chance to cut contact for future cases, improving its efficacy. When full contact tracing is considered, our inference approach shows a significant advantage over the baseline approaches (the error intervals are not overlapping). The advantage is apparent when aiming to “flatten the curve”, that is, when seeking to reduce the maximal number of concurrent infected individuals. On the other hand, when passive tracing is considered, our inference approach shows little to no advantage. For half tracing, our inference approach shows a significant advantage however it is not as prominent as in the full tracing case.

6.4.5 Sensitivity Analysis

Next we examine the inference procedure’s performance sensitivity to the number of available tests and available quarantine orders per day. This set of experiments is motivated by the assumption that the government would prefer to minimize or cap the number of quarantine orders so as to keep the economy as open as possible.

Table 9 compares the best performing baseline approach (Exposure+Symptomatic) with our inference approach for different tests and quarantine caps. The table entries report the maximal number of concurrent infected individuals over the entire simulation run for the baseline as well as the ratio of improvement over the baseline for our inference approach (“improvement ratio”). Results are averaged over 30 runs with similar random seeds for both the baseline and our inference. Results that are 95% statistically significant, using a paired t-test (similar random seeds were used over the compared approaches), are denoted by an asterisk.

We observe that, in general, more testing yields a greater advantage to our inference approach. This trend, however, is not apparent when the baseline method can halt the disease progression (values ≤ 20). In such cases the baseline is sufficient to stop the disease progression, meaning that it does not spread over the entire community. As a result, the advantage from our (better) inference method is limited. This phenomenon is more apparent in the 4% cap quarantine/day. Such an aggressive quarantining policy results in slightly more than 50% of the population concurrently isolated (as opposed to 14% for a 1% cap). Consequently, the epidemic dies out in most runs. Nonetheless, our inference approach can still provide significant advantage when paired with active tracing by stopping the epidemic progression earlier.

6.4.6 Conclusions from Experimental Results

Several general conclusions can be drawn from our empirical study.

- The proposed inference approach can better predict the infected set of individuals prior to the infection peak when compared to the baseline approaches.
- When paired with a simple testing and quarantine policy, the proposed inference approach can significantly reduce the number of concurrent infected cases (flatten the curve). This advantage can reach a factor of $314/20 = 15.7$ (for testing and quarantine caps of 4% and 2% respectively).
- In all of our experiments, other than the aggressive 4% quarantine/day policy, incorporating the inference with active contact tracing resulted in statistically significant improvement over half tracing, which significantly improved on passive tracing.
- When applying an aggressive quarantine policy (4% quarantine/day), identifying infected individuals has little to no advantage as most of the population ends up isolated and the epidemic dies out.

It is important to note that these conclusions are relevant to PandemicSimulator. Discrepancies between the simulated model and the real-world might influence these general conclusions. An important direction for future work is to examine the extent to which the above conclusions hold in other simulation models that allow contact tracing. Ultimately, the reported trends ought to be examined in a real-world scenario.

7. Conclusion

Epidemiological models aim at providing predictions regarding the effects of various possible intervention policies that are typically manually selected. In this article, we instead introduce a Markov modeling methodology for optimizing adaptive mitigation policies aimed at minimizing the number of concurrent infected individuals, or at least keeping it below the hospital capacity, while also minimizing restrictions on personal freedom, for instance by avoiding personal and business lockdown orders. To this end, we implement an open-source agent-based simulator, where pandemics can be generated as the result of the contacts and interactions between individual agents in a community. We analyze the sensitivity of the simulator to some of its main parameters and illustrate its main features, while also showing that adaptive policies optimized via RL achieve better performance when compared to heuristic policies and policies representative of those used in the real world. Moreover, we demonstrate how probabilistic inference can be applied for proactively identifying infected individuals. Such inference, when paired with straightforward testing and quarantine policies achieve significant reductions in the maximal number of concurrent infections.

While our work opens up the possibility to use machine learning and probabilistic inference to explore fine-grained policies in this context, PandemicSimulator still has limitations and could be expanded and improved in several directions. One important direction for future work is to perform a more complete calibration of our simulator against real-world data, while analyzing and visualizing model uncertainty. Moreover, in order to experiment with realistic larger populations, it would be useful to also incorporate venue-specific topological connectivity between locations. Finally, one could also implement and analyze additional testing and contact tracing strategies to contain the spread of pandemics, along with vaccination strategies. In particular, while PandemicSimulator currently assumes app-based contact tracing, it is also important to consider imperfections in this process due to people forgetting contacts or not cooperating.

Ethics Statement

This paper is intended as a proof of concept that Reinforcement Learning algorithms have the potential to optimize government policies in the real world. We acknowledge that the question of what policies to enact is a highly polarizing issue with many political and socio-economic implications. As described in detail in the paper, the simulator introduced here has many free parameters that can dramatically affect its behavior. While we have made an effort to calibrate it to some real-world data, this effort was mainly for the purpose of showing that the simulator *can* be calibrated. If it is to be used to inform any real world policy decisions, it will be essential for these parameters to be calibrated to match historical data in the community in question, in conjunction with local experts. Similarly, the available government actions would need to be set according to the options available to local policy-makers. Even so, it would be important to recognize that the simulator encodes several assumptions and is inherently approximate in its projections. Policymakers must be fully informed of these assumptions and limitations before they draw any conclusions or take any actions based on our experiments or any future experiments in PandemicSimulator.

These cautionary considerations notwithstanding, we consider the contributions of this paper to be an important first step towards the prospect of optimizing pandemic response policies via RL. We would like nothing more than for this work to be continued (by us or by others) to the point where it can be used to good effect for the purpose of saving lives and/or improving the economic health in real world communities.

References

Aleta, A., Martin-Corral, D., y Piontti, A. P., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N. E., Halloran, M. E., Longini Jr, I. M., Merler, S., et al. (2020). Modelling the impact of testing, contact tracing and household quarantine on second waves of covid-19. *Nature Human Behaviour*, *4*(9), 964–971.

Almaraz, E., & Gómez-Corral, A. (2018). On sir-models with markov-modulated events: Length of an outbreak, total size of the epidemic and number of secondary infections. *Discrete & Continuous Dynamical Systems-B*, *23*(6), 2153–2176.

Bansal, S., Grenfell, B. T., & Meyers, L. A. (2007). When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of the Royal Society Interface*, *4*(16), 879–891.

Britton, T., & Pardoux, E. (2019). *Chapter 2 Inference for Markov Chain Epidemic Models*, pp. 343–362. Springer International Publishing, Cham.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., & Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

CDC (2020). COVID-19 pandemic planning scenarios. `https://www.cdc.gov/coronavirus/2019-ncov/hcp/planning-scenarios.html`. Accessed: 2020-09-8.

Chimed-Ochir, O., Nagata, T., Nagata, M., Kajiki, S., Mori, K., & Fujino, Y. (2019). Potential work time lost due to sickness absence and presence among japanese workers. *Journal of occupational and environmental medicine*, *61*(8), 682–688.

Claeson, M., & Hanson, S. (2021). Covid-19 and the swedish enigma. *The Lancet*, *397*(10271), 259–261.

Cobey, S. (2020). Modeling infectious disease dynamics. *Science*.

Cohen, J., & Kupferschmidt, K. (2020). Countries test tactics in 'war'against covid-19..

Del Valle, S. Y., Mniszewski, S. M., & Hyman, J. M. (2013). Modeling the impact of behavior changes on the spread of pandemic influenza. In *Modeling the interplay between human behavior and the spread of infectious diseases*, pp. 59–77. Springer.

Drakopoulos, K., Ozdaglar, A., & Tsitsiklis, J. N. (2014). An efficient curing policy for epidemics on graphs. *IEEE Transactions on Network Science and Engineering*, *1*(2), 67–75.

Duque, D., Morton, D. P., Singh, B., Du, Z., Pasco, R., & Meyers, L. A. (2020). Covid-19: How to relax social distancing if you must. *medRxiv*.

Eddy, S. R. (2004). What is a hidden markov model?. *Nature biotechnology*, *22*(10), 1315–1316.

Grefenstette, J. J., Brown, S. T., Rosenfeld, R., DePasse, J., Stone, N. T., Cooley, P. C., Wheaton, W. D., Fyshe, A., Galloway, D. D., Sriram, A., et al. (2013). Fred (a framework for reconstructing epidemic dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations. *BMC public health*, *13*(1), 1–14.

Grushka-Cohen, H., Cohen, R., Shapira, B., Moran-Gilad, J., & Rokach, L. (2020). A framework for optimizing covid-19 testing policy using a multi armed bandit approach. *arXiv preprint arXiv:2007.14805*.

Gudbjartsson, D. F., Helgason, A., Jonsson, H., Magnusson, O. T., Melsted, P., Norddahl, G. L., Saemundsdottir, J., Sigurdsson, A., Sulem, P., Agustsdottir, A. B., et al. (2020a). Spread of sars-cov-2 in the icelandic population. *New England Journal of Medicine*.

Gudbjartsson, D. F., Helgason, A., Jonsson, H., Magnusson, O. T., Melsted, P., Norddahl, G. L., Saemundsdottir, J., Sigurdsson, A., Sulem, P., Agustsdottir, A. B., et al. (2020b). Spread of sars-cov-2 in the icelandic population. *New England Journal of Medicine*.

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870.

Halliday, J. E., Hampson, K., Hanley, N., Lembo, T., Sharp, J. P., Haydon, D. T., & Cleaveland, S. (2017). Driving improvements in emerging disease surveillance through locally relevant capacity strengthening. *Science*, *357*(6347), 146–148.

He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., et al. (2020). Temporal dynamics in viral shedding and transmissibility of covid-19. *Nature medicine*, *26*(5), 672–675.

Hoertel, N., Blachier, M., Blanco, C., Olfson, M., Massetti, M., Sánchez Rico, M., Limosin, F., & Leleu, H. (2020). A stochastic agent-based model of the sars-cov-2 epidemic in france. *Nature Medicine*.

Hoffmann, J., & Caramanis, C. (2018). The cost of uncertainty in curing epidemics. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, *2*(2), 1–33.

Kaplan, E. H., & Forman, H. P. (2020). Logistics of aggressive community screening for coronavirus 2019. In *JAMA Health Forum*, Vol. 1, pp. e200565–e200565. American Medical Association.

Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Hart, G., Rosenfeld, K., Selvaraj, P., Nunez, R. C., Hagedorn, B., George, L., et al. (2020). Covasim: an agent-based model of covid-19 dynamics and interventions. *medRxiv*.

Khadilkar, H., Ganu, T., & Seetharam, D. P. (2020a). Optimising lockdown policies for epidemic control using reinforcement learning. *Transactions of Indian National Academy of Engineering*.

Khadilkar, H., Ganu, T., & Seetharam, D. P. (2020b). Optimising lockdown policies for epidemic control using reinforcement learning. *arXiv preprint arXiv:2003.14093*.

Kleinman, R. A., & Merkel, C. (2020). Digital contact tracing for covid-19. *CMAJ*, *192*(24), E653–E656.

Larremore, D. B., Wilder, B., Lester, E., Shehata, S., Burke, J. M., Hay, J. A., Tambe, M., Mina, M. J., & Parker, R. (2020). Test sensitivity is secondary to frequency and turnaround time for covid-19 surveillance. *MedRxiv*.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, *172*(9), 577–582.

Lefèvre, C., & Simon, M. (2019). Sir-type epidemic models as block-structured markov processes. *Methodology and Computing in Applied Probability*, 1–21.

Li, M., Dushoff, J., & Bolker, B. M. (2018). Fitting mechanistic epidemic models to data: A comparison of simple markov chain monte carlo approaches. *Statistical Methods in Medical Research*, *27*(7), 1956–1967. PMID: 29846150.

Libin, P., Moonens, A., Verstraeten, T., Perez-Sanjines, F., Hens, N., Lemey, P., & Nowé, A. (2020). Deep reinforcement learning for large-scale epidemic control. *arXiv preprint arXiv:2003.13676*.

Liu, C. (2020). A microscopic epidemic model and pandemic prediction using multi-agent reinforcement learning. *arXiv preprint arXiv:2004.12959*.

Liu, Q.-H., Ajelli, M., Aleta, A., Merler, S., Moreno, Y., & Vespignani, A. (2018). Measurability of the epidemic reproduction number in data-driven contact networks. *Proceedings of the National Academy of Sciences*, *115*(50), 12680–12685.

Marwa, Y. M., Mwalili, S., & Mbalawata, I. S. (2018). Markov chain monte carlo analysis of cholera epidemic. *J. Math. Comput. Sci.*, *8*(5), 584–610.

Metcalf, C. J. E., & Lessler, J. (2017). Opportunities and challenges in modeling emerging infectious diseases. *Science*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Rivers, C. M., & Scarpino, S. V. (2018). Modelling the trajectory of disease outbreaks works. *Nature*.

Salathé, M., Althaus, C. L., Neher, R., Stringhini, S., Hodcroft, E., Fellay, J., Zwahlen, M., Senti, G., Battegay, M., Wilder-Smith, A., et al. (2020). Covid-19 epidemic in switzerland: on the importance of testing, contact tracing and isolation.. *Swiss medical weekly*, *150*(11-12), w20225.

Shoer, S., Karady, T., Keshet, A., Shilo, S., Rossman, H., Gavrieli, A., Meir, T., Lavon, A., Kolobkov, D., Kalka, I., et al. (2020). Who should we test for covid-19? a triage model built from national symptom surveys. *medRxiv*.

Song, S., Zong, Z., Li, Y., Liu, X., & Yu, Y. (2020). Reinforced epidemic control: Saving both lives and economy. *arXiv preprint arXiv:2008.01257*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tindale, L., Coombe, M., Stockdale, J. E., Garlock, E., Lau, W. Y. V., Saraswat, M., Lee, Y.-H. B., Zhang, L., Chen, D., Wallinga, J., et al. (2020). Transmission interval estimates suggest pre-symptomatic spread of covid-19. *MedRxiv*.

Tolles, J., & Luong, T. (2020). Modeling epidemics with compartmental models. *JAMA*.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P. G., Fu, H., et al. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*.

Walensky, R. P., & del Rio, C. (2020). From Mitigation to Containment of the COVID-19 Pandemic: Putting the SARS-CoV-2 Genie Back in the Bottle. *JAMA*, *323*(19), 1889–1890.

Willem, L., Abrams, S., Libin, P. J., Coletti, P., Kuylen, E., Petrof, O., Møgelmose, S., Wambua, J., Herzog, S. A., Faes, C., et al. (2021). The impact of contact tracing and household bubbles on deconfinement strategies for covid-19. *Nature communications*, *12*(1), 1–9.

Xiao, Y., Yang, M., Zhu, Z., Yang, H., Zhang, L., & Ghader, S. (2020). Modeling indoor-level non-pharmaceutical interventions during the covid-19 pandemic: a pedestrian dynamics-based microscopic simulation approach. *arXiv preprint arXiv:2006.10666*.

Zhang, J., Litvinova, M., Wang, W., Wang, Y., Deng, X., Chen, X., Li, M., Zheng, W., Yi, L., Chen, X., et al. (2020). Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside hubei province, china: a descriptive and modelling study. *The Lancet Infectious Diseases*.