Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

FAST STOCHASTIC EPIDEMIC SIMULATIONS AND AN ADAPTATION OF THE NEXT GENERATION MATRIX FOR A COVID-19 EPIDEMIC MODEL OF SOCIAL DISTANCING

Isabelle Rao[1], and Stephen E. Chick[2]

[1]Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, CANADA
[2]Technology and Operations Management Area, INSEAD, Fontainebleau, FRANCE

ABSTRACT

Direct stochastic simulations of medium to large scale Markovian processes with population dynamics may have runtimes that are proportional to the population size, if they account for each state transition of each individual in the population. Several approaches to speed up such simulations have been proposed. We use a discrete-time, Euler-forward type approximation for state transition functions that simulates all transitions within a given time step in an effort to improve run times, at the expense of some (potentially correctable) bias. We illustrate this with a stylized model of COVID-19 social distancing interventions in the United Arab Emirates. We also adapt the next generation matrix method of Hill and Longini (2003) to a continuous time, discrete state model. The approach accelerates simulation run times from a linear scaling of run times in population size to a constant that depends on the number of possible state transitions.

1 INTRODUCTION

This paper addresses the improvement of simulation run times of continuous time Markov chains (CTMCs) that represent discrete counts of individuals in large populations. This paper also explores connections between the next-generation method (NGM) for modeling epidemic growth (Hill and Longini 2003), a discrete-step model that has been useful for optimizing vaccine allocation resources (generation here refers to each step in a disease transmission chain), the more commonly-used compartmental model framework that is based on ordinary differential equations (ODE) in continuous time, and to analogous CTMCs.

We do so in the context of a stylized epidemic model created to inform health leaders in a Middle Eastern country during the Spring of 2020, shortly after the start of the COVID-19 pandemic. The model was designed to (a) generalize some existing, stylized deterministic models of infectious disease transmission and control to a stochastic model to account for variability in outcomes, (b) assess the effects of some public health interventions for social distancing on resource needs for the hospital system, and (c) to provide rapid run times in a context where long run times would not support dynamic decision making.

Simulation run times for such stochastic models might not scale gracefully as a function of total population size, depending on the simulation method that is chosen to run the model. One standard method for simulating CTMCs (Asmussen and Glynn 2007) is to generate the time of the next event, and then to simulate which type of event has occurred (e.g., a transition of an individual from susceptible to infected, or from infected to hospitalized). With this method, the number of such events per simulated time unit scales with the average number of individuals in the population. This is true even if only the counts of individuals in each state need be monitored – a technique that already can improve simulation run times compared to a simulation that models each and every individual explicitly, as with a discrete event simulation that might also allow for non-Markovian dynamics (Brennan et al. 2006).

Methods have been proposed through time to address this issue (e.g, see Ganyani et al. 2021). Several involved exact or approximate probability transition functions that account for all transitions over a short time interval, for example over $[t, t+\Delta t)$, for t on a lattice of times. For example, one might compute the transition probabilities exactly with Chapman-Kolmogorov equations, but this too might be computationally

challenging. Or one might use Poisson approximations for the number of transitions of individuals across each arc, and the error in transition probabilities can be reduced by shrinking Δt, but this might result in more exits from a given node than there are individuals to exit it. Another method might be to relax the discrete space constraint and simulate the process as a Gaussian diffusion process with a state-dependent drift and covariance for transitions. This can build on research from state-dependent diffusion model simulation (e.g, see Asmussen and Glynn 2007), but may lose important information if boundaries are not accounted for properly (to avoid negative population counts), and the diffusion might not be a good representation when there are small numbers of individuals in some of the nodes/compartments.

Section 2 summarizes an application context that was faced near the start of the COVID-19 pandemic and that serves as a basis for numerical studies. Section 3 presents a stylized, deterministic model of infectious disease transmission and vaccination at the start of an epidemic. We then adapt to a deterministic ODE model for our application with social distancing through an entire outbreak, then link that to a stochastic, Markov model of infection in several subpopulations with social distancing. Section 4 summarizes several methods to simulate, exactly or approximately, such stochastic simulations. Section 5 demonstrates speed/bias tradeoffs for one of those simulation methods for estimating the peak and total number of infections and the maximum number of resources used. Section 6 summarizes our results.

Although our application is to epidemic control, our intent is to discuss general tradeoffs with approximate state transition probabilities that improve run times for CTMC simulations of large populations of individuals but that may introduce biases in estimating means. Linkages between NGM, ODE, and CTMC models will be discussed mathematically, the run times and biases will be explored numerically in this paper. Our work focuses on epidemic models (see also Diekmann and Heesterbeek 2000; Pineda-Krch 2008; Allen 2017; Ganyani et al. 2021) but is also related to the rapid simulation of queue networks (Wang et al. 2024) and of other CTMCs that count populations of individuals undergoing transitions through discrete states, including birth (or exogenous arrival) and death (or departure) processes.

2 APPLICATION CONTEXT: SOCIAL DISTANCING FOR COVID-19

We applied our method to model COVID-19 dynamics in the United Arab Emirates (UAE), considering the diverse subpopulations and their adherence to social distancing measures. Specifically, we categorized the UAE population into four main subpopulations: Emirati, Professional Expats, Blue Collar, and Laborers, each with varying capacities to observe social distancing. This characterization resulted in eight dimensions of the state space, reflecting the combination of subpopulation and adherence status. In our model, we accounted for various disease states, including susceptible, infectious, recovered and dead individuals, as well as the potential need for medical resources such as hospital beds, Intensive Care Unit (ICU) beds, ExtraCorporeal Membrane Oxygenation (ECMO), and ventilators. Our primary intervention strategy focused on social distancing. This approach allows us to capture some of the complexities of COVID-19 transmission dynamics within a large and heterogeneous population like that of the UAE.

3 EPIDEMIC MODEL OF PUBLIC HEALTH INTERVENTIONS IN SUBPOPULATIONS

We adopt a previous study's model for vaccine allocation (Hill and Longini 2003), adapt it to the context of COVID-19 and social distancing in the UAE, convert it to an analogous continuous time differential equation model, then convert it to a continuous-time discrete state Markov model. Instead of vaccination, we will consider social distancing of susceptible and infected individuals. We discuss some options to simulate these models in Section 4.

3.1 Next Generation Matrix Model

A previous study (Hill and Longini 2003) developed a model for determining minimal vaccine allocations within a population of m heterogeneous subgroups to prevent an epidemic by reducing the reproduction number to 1. Let R_{ij} be the expected number of secondary infections in unvaccinated individuals in

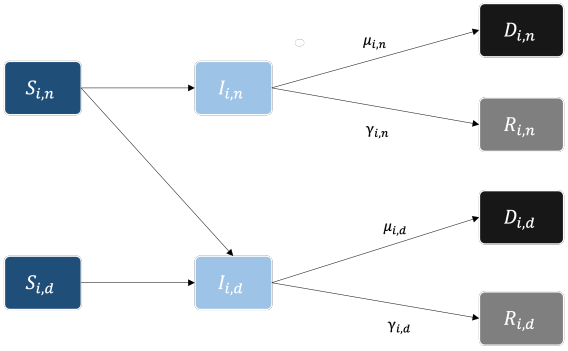


Figure 1: SIR compartmental model for one population subgroup (i).

population subgroup i from a single unvaccinated infected individual in population subgroup j in a fully susceptible population, where $i, j \in \{1, \ldots, m\}$, f_i be the proportion of individuals vaccinated in subgroup i, $1-\theta$ be the vaccine efficacy for susceptibility (the uninfected vaccinated are less likely to get infected following exposure), and $1-\phi$ be the vaccine efficacy for infectiousness (the vaccinated are less likely to transmit to others if they become infected). They model the *beginning* of the epidemic by counting the number of infected after each generation, a generation being defined by the number of transmissions since the index case.

Namely, let $y_{vi}(g)$ be the expected number of secondary infections in population i at generation g, considering individuals as unvaccinated when $v = 0$ and vaccinated when $v = 1$. We have

$$\begin{aligned} y_{0i}(g+1) &= \sum_{j=1}^{m} R_{ij}(1-f_j)y_{0j}(g) + R_{ij}\phi f_j y_{1j}(g) \\ y_{1i}(g+1) &= \sum_{j=1}^{m} R_{ij}\theta(1-f_j)y_{0j}(g) + R_{ij}\theta\phi f_j y_{1j}(g) \end{aligned} \tag{1}$$

We define the vector $y(g) = [y_{01}(g), y_{11}(g), \ldots, y_{0m}(g), y_{1m}(g)]^T$, which allows the system to be a linear recursion that approximates the epidemic size at the start of an outbreak.

We adapt the previous study's model for vaccine allocation (Hill and Longini 2003) to the context of COVID-19 and social distancing (rather than vaccination) in the UAE, with a modification that some who become infected may also choose to social distance, in addition to those that social distance at time 0. We model the spread of a disease with an SIR model with m interacting subpopulations. Figure 1 illustrates the assumed natural history of infection for one of those subpopulations. We include compartments for susceptible individuals (S), infected individuals (I), recovered individuals who are immune to the disease (R), and individuals who have died from the disease (D). We separate individuals based on whether they are social distancing. The subscript n denotes individuals who are not social distancing, and d individuals who are social distancing. These subgroups correspond to unvaccinated and vaccinated, respectively, in the model of Hill and Longini (2003). Similar to the impact of vaccination, we assume that individuals who are social distancing are less likely to become infected or to infect others. We denote by $1-\theta$ and $1-\phi$ the reduction in susceptibility and infectivity for individuals who are social distancing.

Let R_{ij} be the expected number of secondary infections of individuals in subgroup i from an infected individual in subgroup j. We further denote by α_i the duration of infection for individuals in subgroup i for $i \in \{1, \cdots m\}$ and $\mu_{i,x}$ the proportion of individuals who die after infection in subgroup (i,x) for $i \in \{1, \cdots m\}$ and $x \in \{n, d\}$. Here, n and d map to 0 and 1 in (1).

We will assume that at time 0, a fraction f_i of subgroup i observes social distancing, so that $f_i = S_{i,d}(0)/(S_{i,n}(0) + S_{i,d}(0))$ for $i \in \{1, \cdots m\}$. We assume that no infected individuals initially are social distancing, such that $I_{i,d}(0) = 0$. In contrast with the model of Hill and Longini (2003), we assume that some susceptible individuals will start to social distance once they become infected. We denote by ξ_i the

proportion of newly infected individuals in $S_{i,n}$ who start social distancing and move to compartment $I_{i,d}$. With this adaptation, we modify (1) for our application to:

$$\begin{aligned} y_{0i}(g+1) &= \sum_{j=1}^{m} \Big(R_{ij} y_{0j}(g) + R_{ij}\phi y_{1j}(g) \Big)(1-\xi_i) \\ y_{1i}(g+1) &= \sum_{j=1}^{m} \Big(R_{ij}\theta y_{0j}(g) + R_{ij}\theta\phi y_{1j}(g) \Big) + \sum_{j=1}^{m} \Big(R_{ij} y_{0j}(g) + R_{ij}\phi y_{1j}(g) \Big)\xi_i \end{aligned} \tag{2}$$

where y_{0i} corresponds to $I_{i,n}$ and y_{1i} corresponds to $I_{i,d}$ in Figure 1.

3.2 Conversion to an Ordinary Differential Equation Model

We now convert the next generation matrix model of the start of an outbreak for our application in (2) to a continuous time compartmental model, an ordinary differential equation (ODE) model, of the entire outbreak. This will allow us to model infection transmission and control dynamics through time.

We visualize the structure of the epidemic model using a network of nodes. The nodes consist of different combinations of epidemiological states and population subgroups ($S_{i,n}, I_{i,n}, R_{i,n}, D_{i,n}, S_{i,d}, I_{i,d}, R_{i,d}, D_{i,d}$). The nodes are connected by arcs, and we define the flow rates for each directed arc in Table 1. There are 7 arcs connecting the nodes, each representing a type of state transition: i) $S_{i,n} \to I_{i,n}$, ii) $S_{i,n} \to I_{i,d}$, iii) $S_{i,d} \to I_{i,d}$, iv) $I_{i,n} \to R_{i,n}$, v) $I_{i,n} \to D_{i,n}$, vi) $I_{i,d} \to R_{i,d}$ and vii) $I_{i,d} \to D_{i,d}$.

The number of individuals at each node, which is discrete-valued in practice, is approximated by a real-valued quantity in such compartmental models. In our example, the instantaneous flow rates for the associated ODE along each directed arc in Figure 1 are given in Table 1. The ODE that determines the dynamics of the outbreak are defined by adding the flow rates into a given node, and subtracting the sum of flow rates out of a given node. For example, the node $I_{i,n}$ that represents the number of infected in subpopulation i that are not socially distanced, has the following dynamic

$$\begin{aligned} \frac{dI_{i,n}}{dt} &= \underbrace{(1-\xi_i)\frac{S_{i,n}(t)}{N_i(t)} \sum_{j=1}^{m} \Big(\frac{R_{ij}}{\alpha_j} I_{j,n}(t) + \frac{R_{ij}}{\alpha_j}\phi I_{j,d}(t) \Big)}_{\text{flow in}} - \underbrace{\Big(\frac{1-\mu_{i,n}}{\alpha_i} I_{i,n}(t) + \frac{\mu_{i,n}}{\alpha_i} I_{i,n}(t) \Big)}_{\text{flow out}} \\ &= (1-\xi_i)\frac{S_{i,n}(t)}{N_i(t)} \sum_{j=1}^{m} \Big(\frac{R_{ij}}{\alpha_j} I_{j,n}(t) + \frac{R_{ij}}{\alpha_j}\phi I_{j,d}(t) \Big) - \frac{1}{\alpha_i} I_{i,n}(t) \end{aligned} \tag{3}$$

This model spreads out the infections by an individual from subgroup j to individuals in subgroup i by spreading them out over the duration of infectivity (the division by α_j in the 'flow in'), noting that a fraction $S_{i,n}(t)/N_i(t)$ of contacts are with susceptible individuals. Simlarly, the 'flow out' accounts for the average duration of infectivity α_j. The equations for the other nodes with outbound arcs follow similarly.

For nodes with no outbound arc, only inbound terms are included, e.g., $dD_{i,n}/dt = (\mu_{i,n}/\alpha_i) I_{i,n}(t)$. Such nodes are terminal (or sinks) for flows of individuals.

By combining such equations for all nodes, we have a system of (nonlinear) differential equations that describe the dynamics for the state vector $\mathbf{X}_t = (S_{i,n}, I_{i,n}, R_{i,n}, D_{i,n}, S_{i,d}, I_{i,d}, R_{i,d}, D_{i,d})_{i=1}^{m}$, which monitors the number if individuals in each compartment/node (one node for each subpopulation, social-distancing status combination). We write the overall dynamics compactly as:

$$\frac{d\mathbf{X}_t, n}{dt} = \mathscr{G}(\mathbf{X}_t, t) \tag{4}$$

where in this case the rate function vector $\mathscr{G}(\mathbf{X}_t, t)$ is quadratic in $\mathbf{X}_t$ due to the nature of terms like those in (3). That function is not dependent on t if there is no time-dependent intervention done.

3.3 Conversion to Continuous Time Discrete State Markov Model

It is known that when ODE epidemic models may be misleading for quantifying average infection levels or the nature of the best decisions for epidemic control when the large-population limit behind the ODE is not valid (Koopman et al. 2002). This can happen, for example, when the number of individuals in some compartments of the model is small. (Colloquially, small might be understood relative to whether a continuous-time, continuous state diffusion approximation to a CTMC epidemic model 'hits the boundary often' during the simulateions of interest.) Moreover, ODE models do not model stochastic variability explicitly, and variability may be important for resource allocation decisions regarding care processes.

We therefore convert this model to a continuous time discrete state Markov model (CTMC). The model will account for each individual state transition of any individual from one node to another – that is, each recover and infection and death event will be modeled. This is stochastic, and discrete vector-valued (unlike the deterministic, continuous space ODE model). For our specific application, we start with the same initial state as the ODE model. For the dynamics, the overall flow rates from the ODE model in Table 1 become instantaneous transition probabilities for the associated CTMC. For example, the instantaneous transition probability for a given state $\mathbf{X}_t = (S_{i,n}, I_{i,n}, R_{i,n}, D_{i,n}, S_{i,d}, I_{i,d}, R_{i,d}, D_{i,d})_{i=1}^m$ to a state $\mathbf{X}_t = (S_{i,n}-1, I_{i,n}+1, R_{i,n}, D_{i,n}, S_{i,d}, I_{i,d}, R_{i,d}, D_{i,d})_{i=1}^m$, meaning an infection of a non-social distanced individual who remains non social distanced, is the rate in the first row of Table 1. Such a transition is associated with the directed arc $(S_{i,n}, I_{i,n})$, and the probability that this happens in time interval $[t, t+\Delta t)$ is

$$\text{Prob}(S_{i,n} \to I_{i,n} \mid \mathbf{X}_t) = \Delta t\, r_{S_{i,n},I_{i,n}} + o(\Delta t) = \Delta t (1-f_i) \frac{S_{i,n}(t)}{N_i(t)} \sum_{j=1}^{m} \left(\frac{R_{ij}}{\alpha_j} I_{j,n}(t) + \frac{R_{ij}}{\alpha_j} \phi I_{j,d}(t) \right) + o(\Delta t), \quad (5)$$

where $r_{j,\ell}$ is the infinitesimal transmission probability / flow rate on the directed arc from node j to node ℓ, and $o(\Delta t)$ is a function such that $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$.

More generally, let $\mathscr{N}$ be the set of nodes for our CTMC representing population flows (compartments in an epidemic model, or nodes in a queue network). We let $\mathscr{N}$ also contain a special node n_b which can generate new arrivals (births) to the system, and/or a special node n_d to represent system departures.

We let $\mathscr{A}$ denote the set of directed arcs with possible flows. One way to model births to a given node $j \in \mathscr{N}$ is to create a flow from node n_b that does not depend on the capacity of n_b at a given rate $r_{n_b,j}$. The value of $r_{n_b,j}$ might be constant if there is a constant stream of arrivals, might be proportional to the sum of some of the dimensions of $\mathbf{X}_t$ if this is a birth process, or might (implicitly) depend on t if there is a time-varying arrival process. Departures might be explicitly modeled by absorbing node(s), as do the $2m$ death compartments $D_{i,\cdot}$ in our model, or by allowing n_d to serve as the model's sole absorbing node.

We let $\mathscr{G} = (\mathscr{N}, \mathscr{A})$ denote the digraph of such a model, and let $\mathbf{r} = (r_a)_{a \in \mathscr{A}}$ be the vector of its transition probabilities (or flow rates). Similarly, we let $\mathbf{r}^{n',\cdot}$ be the flow rates associated with set of directed arcs that start from node n', and we let $\mathbf{r}^{\cdot,n}$ be the flow rates associated with set of directed arcs that terminate at node n. It will be useful to refer to the sum of elements of a vector, and we use the 1-norm to do so, e.g., $\| \mathbf{r}^{\cdot,n} \|_1$ is the sum of flow rates into node n. These flow rates may be time dependent and state dependent (a function of the state vector $\mathbf{X} = (X_n)_{n \in \mathscr{N}}$).

We note that such a $\mathscr{G}, \mathbf{r}$ combination can also determine a deterministic ODE model, as in Section 3.2, in addition to a CTMC, as here in Section 3.3. Such ODE models can be large-population limits, in some sense, of the CTMC model, given technical conditions (Kurtz 1971; Diekmann and Heesterbeek 2000). We will compare simulation algorithms and results for these models in the coming sections.

In what follows, it will be useful to refer to the specific dimension of the state vector for node n by number. Therefore, we will also write $\mathbf{X}_t = (X_1, X_2, \ldots, X_{|\mathscr{N}|})$ and refer to the count of individuals in node n by X_n, rather than referring to the names of the nodes in 1. The t is implicit in the notation.

Table 1: Transition rates for arcs in the simplified COVID-19 model of Figure 1.

Directed Arc	Flow (or Transition) Rate	Meaning
$S_{i,n} \to I_{i,n}$	$(1-\xi_i)\frac{S_{i,n}(t)}{N_i(t)}\sum_{j=1}^{m}\left(\frac{R_{ij}}{\alpha_j}I_{j,n}(t)+\frac{R_{ij}}{\alpha_j}\phi I_{j,d}(t)\right)$	Infection of non social distanced individual who remains non social distanced
$S_{i,n} \to I_{i,d}$	$\xi_i\frac{S_{i,n}(t)}{N_i(t)}\sum_{j=1}^{m}\left(\frac{R_{ij}}{\alpha_j}I_{j,n}(t)+\frac{R_{ij}}{\alpha_j}\phi I_{j,d}(t)\right)$	Infection of non social distanced individual who starts social distancing
$S_{i,d} \to I_{i,d}$	$\theta\frac{S_{i,d}(t)}{N_i(t)}\sum_{j=1}^{m}\left(\frac{R_{ij}}{\alpha_j}I_{j,n}(t)+\frac{R_{ij}}{\alpha_j}\phi I_{j,d}(t)\right)$	Infection of social distanced individual who keeps social distancing
$I_{i,n} \to R_{i,n}$	$\frac{1-\mu_{i,n}}{\alpha_i}I_{i,n}(t)$	Recovery of non social distanced individual
$I_{i,n} \to D_{i,n}$	$\frac{\mu_{i,n}}{\alpha_i}I_{i,n}(t)$	Death of non social distanced individual
$I_{i,d} \to R_{i,d}$	$\frac{1-\mu_{i,d}}{\alpha_i}I_{i,d}(t)$	Recovery of social distanced individual
$I_{i,d} \to D_{i,d}$	$\frac{\mu_{i,d}}{\alpha_i}I_{i,d}(t)$	Death of social distanced individual

4 SOME OPTIONS FOR SIMULATING THE EPIDEMIC MODEL

We recall some algorithms to simulate a general epidemic model, be it the CTMC version, $\mathscr{G}, \mathbf{r}$, in Section 3.3 or its ODE analog in Section 3.2. Some but not all of these models (the Euler-multinomial method below), are discussed in an instructive tutorial by Allen (2017). That work also discusses other models and an illustration with malaria, rather than our use of a running example of COVID-19.

Example 1 In our COVID-19 social distancing application, $\mathscr{N}$ has 32 nodes, which represent the 8 nodes in Figure 1 per subpopulation $i = 1,2,3,4$, $\mathscr{A}$ has 28 directed arcs, which represent the 7 directed arcs in Figure 1 for each $i = 1,2,3,4$, and r_{ℓ_1,ℓ_2} is specified in Table 1 for each $(\ell_1,\ell_2) \in \mathscr{A}$. An enhanced version of the model would have additional nodes to account for infected individual's transitions into states requiring added resources (hospital beds, ICU care, ventilators, or ECMO machines).

4.1 Deterministic ODE Model Simulation

The ODE model of Section 3.2 and specified in (4) is often simulated with Euler-forward or higher-order approximations that can give even better numerical accuracy (Press et al. 1992). We will use an Euler-forward approach, which essentially computes the state vector $\mathbf{X}_t$ for a lattice of times $t = i\Delta t$, where $\Delta t > 0$ is a small time step, T is a time horizon of interest, and $i = 0,1,2,\ldots,\lfloor T/\Delta t \rfloor$, with

$$\mathbf{X}_{(i+1)\Delta t} = \mathbf{X}_{i\Delta t} + \mathscr{G}(\mathbf{X}_{i\Delta t}, i\Delta t)\Delta t. \tag{6}$$

Naive application of this Euler-forward algorithm may allow for the contents of a compartment to become negative. Therefore, we have adapted this approach so that the time step Δt is dynamically shrunk on a given iteration of the algorithm so as to prevent the contents of any compartment from becoming negative.

4.2 Stochastic CTMC Model Simulation

4.2.1 Exact Simulation Methods

Two methods for simulating CTMC epidemic models formulated as in Section 3.3 are to simulate every state transition, or to use an Euler-forward type approach with Chapman-Kolmogorov forward equations to determine the exact transition functions over a given time of duration Δt.

Simulating every state transition can be done with standard CTMC algorithms (Asmussen and Glynn 2007). In epidemiology, this is also known as the Gillespie algorithm (Ganyani et al. 2021). See Algorithm 1. It proceeds by computing the time to the next event, which is exponentially distributed with rate equal

the sum of flow rates of all arcs in the graph. It then determines which arc is activated, by sampling from a multinomial distribution with one trial (one transition) over all arcs, with probabilities proportional to the flow rates of each arc. The resulting outcome, called arcflow, has all zeros except for one 1 that corresponds to the directed arc that is activated. We let arcnum(arcflow) identify that directed arc. The transition associated with that arc is processed, the clock is updated, and the process repeats.

Chapman-Kolmogorov (CK) equations can be used to compute the probability transmission matrix which contains transition probabilities from all states $\mathbf{X}_t$ at time t to all states $\mathbf{X}_{t+\Delta t}$ at time $t+\Delta t$ (Ross 1983). Although this gives exact transmission probabilities over intervals $[t, t+\Delta t]$, and can account for multiple state transitions of individuals over that time interval, we choose not to simulate with this method – it would involve solving a high-dimensional matrix ODE for the transition probabilities at each time step.

4.2.2 Approximate Simulation Methods

The Poisson Method combines the time step of the Euler forward method, and the characterization of the Poisson distribution as a certain limit of Bernoulli random variables. Colloquially, suppose that the flow rate $r_{n,n'}$ does not vary much over the time interval $[t, t+\Delta t)$. From (5) and the characterization of the Poisson distribution as a certain limit of Bernoulli random variables, one expects that the number of transitions from node n to node n' on arc (n,n') during time interval $[t, t+\Delta t)$ would have (approximately) a Poisson distribution with mean $r_{n,n'}\Delta t$.

An algorithm to implement this approximation could iterate at each time t on the lattice $0, \Delta t, 2\Delta t, \ldots$ and sample independent Poisson random variables for each arc, with mean according to Δt and the flow rate, then each flow would be executed, with state updates made for the next time step.

The Gaussian Diffusion Method is similar to the Poisson method, except that the dimensions of the state vector are no longer required to be integers. They can be continuous-valued. Rather than using Poisson increments, each arc can have a normally distributed flow with the same mean and standard deviation as the Poisson increment (both are $r_{n,n'}\Delta t$, here). The resulting continuous-time continuous state (CTCS) process can be simulated on a lattice of times, as with the Euler forward for ODEs or the Poisson method, but the state is continuous, as with the ODE. This model does account for variability, which is a benefit beyond the ODE, and this can be useful when ther are multiple subpopulations with very different rates for each subpopulations (such as the counts of viruses in an environment on the one hand, and numbers of infected individuals on the other hand, e.g., see Chick et al. 2004). These types of simulations are regularly used in financial modeling, among other areas, and there are tools to control the error as a function of the time step (e.g., see Whitt 2002; Asmussen and Glynn 2007). The resulting CTCS process can be represented by a stochastic differential equation (SDE), and the diffusion model is a discretization of the continuous-time SDE to facilitate numerical simulations.

One important problem with these Poisson and Gaussian diffusion approaches is that the computed outflow from a given node at a given time step may exceed the contents of the node. This results in

Algorithm 1: Gillespie method for exactly simulating a CTMC epidemic model.

Input: $\mathscr{N}, \mathscr{A}, \mathbf{r}, T$
Set time $t = 0$
while $t < T$ **do**
  Compute duration of time to next event, $\tau \leftarrow -\ln(U)/\sum_{a\in\mathscr{A}} r_a$.
  Determine which arc is active in that event: arcflow $\sim$ Multinomial$(1, \mathbf{r}_t/\sum_{a'\in\mathscr{A}} r_{a'})$
  Find nodes of that arc: $(\ell_1, \ell_2) \leftarrow$ arcnum(arcflow)
  Action the transition for that arc: decrement ℓ_1 by 1 and increment ℓ_2 by 1
  $t \leftarrow t + \tau$
end
Report simulation results

Algorithm 2: Euler-multinomial method for approximately simulating an epidemic model with nodes $\mathscr{N}$ and transition arcs $\mathscr{A}$, and transition rates $\mathbf{r}$, with time step $\Delta t > 0$.

Input: $\mathscr{N}, \mathscr{A}, \mathbf{r}, \Delta t, T$
Set time $t = 0$
while $t < T$ **do**
  Let $h \leftarrow \min(\Delta t, 1/(\max_n \| \mathbf{r}^{n,\cdot} \|_1))$ (shrink time step if needed)
  for $n \in \mathscr{N}$ **do**
    Let X_n be the number of individuals at node n at time t
    Compute flows on arcs leaving n: $\text{outflow}_n \leftarrow \text{Multinomial}(X_n, [h\mathbf{r}^{n,\cdot}, 1 - h\| \mathbf{r}^{n,\cdot} \|_1])$
  end
  Update the simulated time, $t \leftarrow t + h$
  for *each arc* $(n, n') \in \mathscr{A}$ **do**
    Decrement X_n by $\text{outflow}_n(n')$, and increment $X_{n'}$ by $\text{outflow}_n(n')$
  end
end

nonfeasible flows on some sample paths. The probability of such nonfeasible flows can be made smaller by choosing $\Delta t > 0$ smaller, but the probability is not zero. Moreover, a smaller Δt implies a longer run time (more iterations until the time horizon is reached).

The Euler-multinomial method operates on a time grid, like the Euler forward method, but allows for stochastic flows of discrete counts of individuals, like the Poisson method. Unlike the Poisson method, which may allow computed flows to exceed the amount that is able to flow from a node, this method uses a multinomial distribution to ensure that computed flows do not violate mass conservation constraints. See also related binomial model proposals (Pineda-Krch 2008; Allen 2017; Ganyani et al. 2021).

It does this, for each node n with contents X_n at time t, by computing a multinomial random variable, with parameter X_n for the number of items to categorize into bins, one bin for each arc leaving n and one bin to allow contents of the node to remain in node n for the next $\Delta t > 0$ units of time. The probabilities for the flow on each arc are proportional to the flow rates of each arc leaving n and are also proportional to Δt. The probability of remaining in node n is 1 less the sum of those outflow probabilities. This gives the first for loop in Algorithm 2. Then, simulated time is incremented by Δt, so that outflows that are computed at time t result in state changes Δt time units later. Those state changes are computed, arc by arc, in the second for loop of Algorithm 2.

For Algorithm 2 to be feasible, $\Delta t > 0$ must be chosen sufficiently small so that all calls to the Multinomial distribution have feasible probabilities: namely, the probability of flowing out of arc n at each time t for each possible state must be less than one, $\Delta t \| \mathbf{r}^{n,\cdot} \|_1 \leq 1$. For applications where the maximum (of the time dependent) flow rates are computable for every possible state, Δt can be chosen to be 1 over the sum of the maximum of all such rates for all nodes at each time. In practice, however, we will allow Δt to be somewhat bigger than that, and we may dynamically shrink the time step if needed, with $h \leftarrow 1/(\max_n \| \mathbf{r}^{n,\cdot} \|_1)$ taking the role of Δt.

4.3 Issues with Approximate Simulation Methods

For each of the simulation methods in this section, except for the exact/Gillespie method, the choice of the time step parameter Δt plays an important role. This is true for both the deterministic and stochastic simulations. A smaller time step $\Delta t > 0$ means that more time steps are required to simulate the process to a time horizon T. This causes a longer run time. A smaller time step $\Delta t > 0$ may also result in a more accurate approximation to the underlying process. The choice of Δt, therefore, has implications for the tradeoff between computational speed and biases in estimated quantities of interest.

Table 2: Values for model parameters for a UAE-like population near start of the COVID-19 pandemic.

Parameter	*Description*	*Emirati*	*White Collar*	*Blue Collar*	*Laborers*
$S_n(0)$	Starting susceptible population (non social distancing)	493800	483924	1037071	4956288
$S_d(0)$	Starting susceptible population (social distancing)	494500	484610	692300	1240206
$I_n(0)$	Starting infected population (non social distancing)	700	686	1379	4536
$I_d(0)$	Starting infected population (social distancing)	0	0	0	0
$R_n(0)$	Starting recovered population (non social distancing)	0	0	0	0
$R_d(0)$	Starting recovered population (social distancing)	0	0	0	0
$D_n(0)$	Starting dead population (non social distancing)	0	0	0	0
$D_d(0)$	Starting dead population (social distancing)	0	0	0	0
f	Fraction of susceptibles who initially social distance	0.5	0.5	0.4	0.2
ξ	Fraction of newly infected that switch to social distancing	0.1	0.1	0.1	0.1
α	Average duration of infection (days)	14	14	14	14
μ	Fraction of infected who die	0.01	0.01	0.01	0.01
h	Fraction of infected that require hospitalization	0.47	0.2	0.16	0.06
d_h	Average number of hospitalization days	12.7	12.5	12.	12.8
i	Fraction of infected that require ICU	0.13	0.08	0.06	0.02
d_i	Average number of ICU days	12.8	12.7	12.5	19.4
v	Fraction of infected in ICU that require ventilators	0.45	0.45	0.40	0.42
d_v	Average number of days on ventilators	16	10	14	23
e	Fraction of infected in ICU that require ECMO machines	0.03	0.05	0.04	0.1
d_e	Average number of days on ECMO machines	15	15	12	38

5 ILLUSTRATIVE NUMERICAL EXPERIMENTS

In Section 5.1 we describe data for the UAE social distancing application introduced in Section 2. We assess speed-bias tradeoffs for simulated estimates as a function of the simulation time step Δt and population size in Section 5.2. Section 5.3 presents histograms for resource requirements during simulated outbreaks.

5.1 Data for the Model

We extended the model depicted in Figure 1 to account for four interacting subpopulations. The sub population cohort data was obtained from UAE census data from 2018. We used initial transmission data from March 2020 to estimate the R_{ij} values for each sub population in (1).

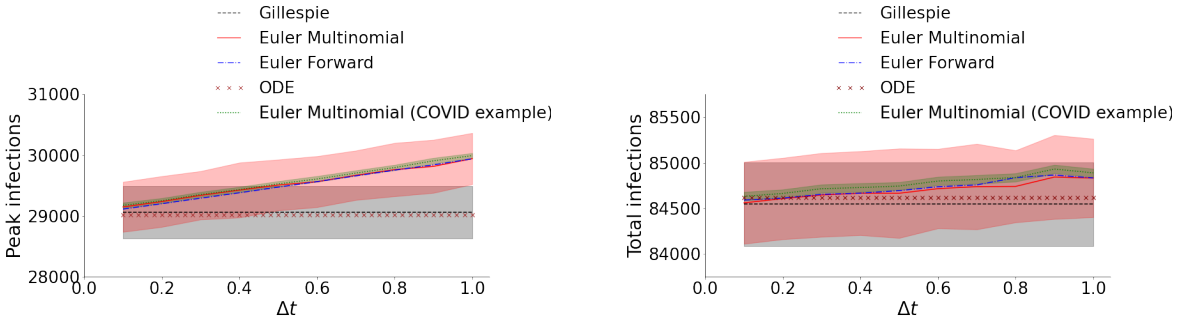
$$R = \begin{bmatrix} 1.70 & 0.07 & 0.05 & 0.01 \\ 0.20 & 1.91 & 0.05 & 0.10 \\ 0.20 & 0.07 & 2.53 & 0.10 \\ 0.02 & 0.07 & 0.05 & 5.00 \end{bmatrix}.$$

In addition, that model was extended to account for the chance that infected individuals might need hospitalization, or ICU care, or other resources (ventilators, ECMO machines) in the ICU. The capacity of these resources was initially ignored – one hope for the model was to learn the evolution of the outbreak, assuming sufficient capacity. Parameters related to resources were also determined based on hospitalized COVID-19 patients from March 2020, data available at the time, or expert guidance.

The model was then simulated for a range of values for fraction that social distance vs non-social distance and ϕ and θ values. Table 2 summarizes the model parameter values for the numerical experiments. We show the results for $\theta = 1$, $\phi = 0.25$.

5.2 Experiments that Explore Speed-Bias Tradeoffs

To illustrate the COVID example using the Gillespie algorithm, we scale down the starting population size in Table 2 by a factor of 100, rounding to the nearest integer. We will refer to this population as the smaller



(a) Mean *peak* number of infections.

(b) Mean *total* number of infections.

Figure 2: Estimated mean peak and mean total number of infections for Euler forward ODE, stochastic Euler-multinomial and Gillespie simulations as function of time step Δt. In the legend, "COVID example" refers to the example as described in Table 2, while the others refer to the smaller population size.

population. We compare some choices of Δt, and discuss the estimated bias of the Euler-multinomial in several estimates of interesting metrics (e.g. max infected and total infected). We compare expected output for range of Δts and total population sizes, and compare with the associated limiting ODE, Euler forward ODE and CTMC Gillespie mean.

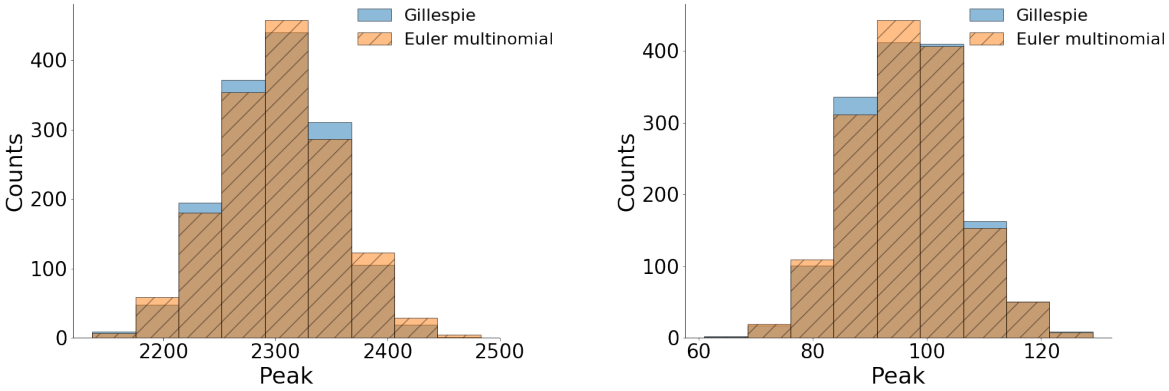
We consider two population sizes: the original COVID example population (as detailed in Table 2) and the smaller population. For Gillespie, Euler forward and the ODE, we simulate only with the smaller population. For Euler-multinomial, we simulate with both population sizes, and scale down the outcomes with the COVID example by 100. Figure 2 compares the mean peak and total number of infections for all five different scenarios. We vary Δt between 0.1 and 1 with increments of 0.1. We iterate our Euler-multinomial and the Gillespie method 250 times for each Δt. The solid lines represent the mean of the outcomes, and the shaded bands represent the 95% confidence intervals.

Bias. We find that for both the peak and total number of infections, the difference in outcomes between the Euler-multinomial (with both population sizes) and Gillespie algorithms increases roughly proportional to Δt for small Δt. The Euler-multinomial algorithm with the (larger) COVID example population follows the same trend as the Euler-multinomial algorithm with the smaller population, but the confidence interval is 10 times smaller, which might be expected with large population limits for such scaled population dynamics models. The bias associated with the time step has a greater influence in this example than the bias associated with scaling the population size by a factor of 100 in this example.

Speed and Bias. Suppose that Δt is a sufficiently small value for the time step Δt so that the time step h is typically not shrunk in the Euler-multinomial method of Algorithm 2. Further cutting Δt in half would double the time steps, and therefore double the number of iterations of the while loop in Algorithm 2. Heuristically speaking, then, for sufficiently small Δt, the bias in estimated means is cut in half when Δt but the run time is doubled (ignoring the fixed setup costs of the simulation run). We see this in Table 3, where we show iterations as run times are mostly proportional to iterations. We note that a rigorous proof of the monotonic increase of the bias in Δt is beyond the scope of this work.

Table 3: Average speed and estimated bias for several values of Δt for the Euler-multinomial algorithm in estimating the mean total number of infections.

Time step Δt	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Iterations	250000	125000	83500	62500	50000	41750	35750	31250	28000	25000
Bias	27.7	50.9	66.0	109.3	145.6	192.6	211.9	283.1	344.0	293.0



(a) Maximum number of hospitalized patients.

(b) Maximum number of ECMO machines in use.

Figure 3: Histogram of the maximum resource usage during outbreak for Euler-multinomial ($\Delta t = 0.1$). Vertical lines represent the mean values from the Gillespie, Euler forward and Euler-multinomial algorithms.

For small to medium Δt, then, one might try to estimate the Gillespie mean at time t, which may require a lot of computation for large populations, with two Euler-multinomial runs with time steps Δt and $2\Delta t$, with outputs $\hat{\mu}_{\Delta t}$ and $\hat{\mu}_{2\Delta t}$, respectively, then reduce what appears to be first order bias by estimating the Gillespie mean with an arbitrarily small time step by $\lim_{\varepsilon \to 0} \hat{\mu}_{t+\varepsilon} \approx 2\hat{\mu}_{t+\Delta t} - \hat{\mu}_{t+2\Delta t}$. This estimates the mean of the exact Gillespie algorithm, with a first-order Taylor-series type bias correction and two Euler-multinomial runs with approximate transition probabilities. This can be useful if the Euler-multinomial simulations can be run much more quickly. A more careful analysis of this heuristic correction is warranted.

If Δt is too large, then the time step in Algorithm 2 will be dynamically shrunk from Δt to some smaller h, if needed, so that flows can be computed with a multinomial distribution (probabilities of outflows on each arc must not sum to more than 1). When Δt is large, simulated clock time will pass more quickly when there is no need to shrink Δt, but when the rates of outflows is high for a given state and time, the simulated clock time will automatically be advanced in increments h smaller than Δt. This would suggest that the bias would not be linear for larger Δt even if the true underlying bias only had first-order Taylor-expansion bias terms. However, it is worth nothing that Δt only starts to shrink for values higher than 6.8 days. At this threshold, the bias of the Euler-multinomial algorithm would become very high.

5.3 Experiments to Explore Resource Needs

We now explore the volatility in resource needs under the assumption that the available equipment can sufficiently accommodate the demand. We run 1500 iterations of the Euler-multinomial and Gillespie algorithm with the smaller population example. Figure 3 shows the distribution of maximum resource needs for number of hospital beds and ECMO machines. In this example, the standard deviations are approximately the square root of the estimated mean, which would be expected if data were Poisson distributed. Our analysis shows that the distributions for all four considered resources generated with the Gillespie and Euler-multinomial algorithms closely align. The percentage difference between the mean and standard deviation is less than 5%. For example, for the number of hospitalizations, we have $(\bar{X}, \hat{\sigma}) = (2302.2, 51.08)$ with Gillespie, compared to $(2302.7, 49.62)$ with Euler-multinomial. These findings highlight the usefulness of the Euler-multinomial algorithm in capturing essential characteristics of the epidemic dynamics while being computationally more efficient. This can be a guide for assessing the overage or underage costs at peak loading during the outbreak.

6 DISCUSSION AND CONCLUSIONS

Simulations of epidemic models can be accelerated by the use of approximate simulation methods. It is known that a larger Δt for either the Euler-forward method for simulated deterministic ODE improves run times but may increase error in estimating maximums on disease trajectories or cumulative number infected during an entire outbreak. We observed a similar phenomenon for the stochastic Euler-multinomial approach presented above and applied to an epidemic control model: cutting Δt in half cut the bias for such estimates in half, but cuts run time in half (heuristically speaking). If one runs with two values of Δt, one might get fast run times with an ability to reverse engineer the means predicted for very small values of Δt with modest computational cost, using first-order bias reduction techniques.

ACKNOWLEDGMENTS

We are thankful for the support of Dr. Madhu Sasidhar, formerly Chief Medical Officer at Cleveland Clinics Abu Dhabi, and of Suma Krishnaprasad, data scientist at Cleveland Clinics Abu Dhabi. The authors acknowledge the support of Dr Simba Gill and Sabi Dau to the INSEAD Healthcare Management Initiative.

REFERENCES

Allen, L. 2017. "A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis". *Infect Dis Model* 2(2):128–142.

Asmussen, S. and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. New York: Springer.

Brennan, A., S. E. Chick, and R. Davies. 2006. "A Taxonomy of Model Structures for Economic Evaluation of Health Technologies". *Health Economics* 15(12):1295–310.

Chick, S. E., S. Soorapanth, and J. S. Koopman. 2004. "Microbial Risk Assessment for Drinking Water". In *Operations Research and Health Care: Handbook of Methods and Applications*, 497–494. Kluwer Academic Publishers.

Diekmann, O. and J. Heesterbeek. 2000. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis, and Interpretation*. Chichester: Wiley.

Ganyani, T., C. Faes, and N. Hens. 2021. "Simulation and Analysis Methods for Stochastic Compartmental Epidemic Models". *Annual Review of Statistics and Its Application* 8(1):69–88.

Hill, A. N. and I. M. Longini, Jr.. 2003. "The critical vaccination fraction for heterogeneous epidemic models". *Mathematical Biosciences* 181:85–106.

Koopman, J. S., S. E. Chick, C. P. Riolo, C. P. Simon and G. Jacquez. 2002. "Stochastic Effects of Disseminating Versus Local Transmission of Infection". *Mathematical Biosciences* 180:49–71.

Kurtz, T. G. 1971. "Limit Theorems for Sequences of Jump Markov Processes Approximating Ordinary Differential Processes". *Journal of Applied Probability* 8:344–356.

Pineda-Krch, M. 2008. "GillespieSSA: Implementing the Gillespie Stochastic Simulation Algorithm in R". *Journal of Statistical Software* 25(12):1–18.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. *Numerical Recipes in C*. Second ed. Cambridge, USA: Cambridge University Press.

Ross, S. M. 1983. *Stochastic Processes*. New York: John Wiley & Sons, Inc.

Wang, T., Y. Song, and J. Hong. 2024. "Fast Approximation to Discrete-Event Simulation of Markovian Queueing Networks". In *Proc. Winter Simulation Conference*, WSC '23, 3613–3623. IEEE Press. https://doi.org/10.1109/WSC60868.2023.10407737.

Whitt, W. 2002. *Stochastic Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues*. New York: Springer.

AUTHOR BIOGRAPHIES

ISABELLE RAO is an Assistant Professor at the University of Toronto. She started this work as a post-doctoral researcher at INSEAD – Europe Campus. Her research focuses on developing novel mathematical models to inform critical decisions in public health and personalized medicine. Her email address is isabelle.rao@utoronto.ca and her website is https://www.isabellerao.com/.

STEPHEN E. CHICK is a Professor of Technology and Operations Management, and Novartis Chair of Healthcare Management, at INSEAD – Europe Campus. His research interests include stochastic simulation, Bayesian inference, sequential learning, and applications in health care innovation and public health decisions. His email address is stephen.chick@insead.edu and his website is https://www.insead.edu/faculty/stephen-e-chick.