




Predicting the diversity of early epidemic spread on networks

Andrea J. Allen ^{1,*} Mariah C. Boudreau,^{1,2} Nicholas J. Roberts,¹ Antoine Allard ^{3,4,1} and Laurent Hébert-Dufresne ^{1,2,3,5}

¹Vermont Complex Systems Center, University of Vermont, Burlington, Vermont 05405, USA

²Department of Mathematics & Statistics, University of Vermont, Burlington, Vermont 05405, USA

³Département de physique, de génie physique et d'optique, Université Laval, Québec, Québec, Canada G1V 0A6

⁴Centre interdisciplinaire en modélisation mathématique, Université Laval, Québec, Québec, Canada G1V 0A6

⁵Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA



(Received 16 November 2021; accepted 26 January 2022; published 16 February 2022)

The interplay of biological, social, structural, and random factors makes disease forecasting extraordinarily complex. The course of an epidemic exhibits average growth dynamics determined by features of the pathogen and the population, yet also features significant variability reflecting the stochastic nature of disease spread. In this paper, we reframe a stochastic branching process analysis in terms of probability generating functions and compare it to continuous time epidemic simulations on networks. In doing so, we predict the diversity of emerging epidemic courses on both homogeneous and heterogeneous networks. We show how the challenge of inferring the early course of an epidemic falls on the randomness of disease spread more so than on the heterogeneity of contact patterns. We provide an analysis, which helps quantify, in real time, the probability that an epidemic goes supercritical or conversely, dies stochastically. These probabilities are often assumed to be one and zero, respectively, if the basic reproduction number or R_0 is greater than 1, ignoring the heterogeneity and randomness inherent to disease spread. This framework can give more insight into early epidemic spread by weighting standard deterministic models with likelihood to inform pandemic preparedness with probabilistic forecasts.

DOI: [10.1103/PhysRevResearch.4.013123](https://doi.org/10.1103/PhysRevResearch.4.013123)

I. INTRODUCTION

By the time of this writing, the COVID-19 pandemic had reached every corner of the world. Public health efforts are now focused on identifying new clusters of outbreaks and their risk of causing new epidemic waves, much like they did at the beginning of the pandemic. As large outbreaks soared early on in a handful of countries, sporadic clusters of confirmed cases dotted regions in the United States. Data surrounding new clusters or waves tend to consist of low numbers of cases highly sensitive to noise, sparking concern and uncertainty at the expected progression of the epidemic.

The first confirmed case of COVID-19 in the US was reported on January 21st, 2020 in the state of Washington [1]. Three subsequent cases were later identified in Washington; two hospitalizations on February 19th [2], and two deaths on February 26th, one week later [3]. Then, on February 28th, a high school closed immediately after one of its students tested positive for a strain that had been associated with the January 21st case [4]. With limited knowledge of active cases, it was

nearly impossible to predict the current and future severity of the outbreak.

One critical question in Washington after over a month with only a handful of detected cases, was whether this chain of events suggested a single tree of very few local transmissions, or multiple distinct introduction events from abroad. Despite decades of disease modeling, the community was ill-equipped to answer this question. The problem is challenging in part because of inadequate testing at the time, and also because well-established disease models often operate on deterministic mechanisms designed to describe the average behavior of large epidemics and not the random, discrete nature of small transmission chains. The looming question of whether a local COVID-19 outbreak would die off by itself or become a disaster, can only be modeled using tools capturing the stochasticity, or randomness, of person-to-person contact. To accurately model the potential outcomes of an epidemic based on limited case data, tools that capture the random nature of disease spread along with the structure of the population are required.

In this paper, we analyze the diversity of early epidemic courses. In doing so, we also hope to provide analytical tools to inform disease forecasts by accounting for the heterogeneity and stochastic nature of disease transmission.

Since the introduction of mean-field epidemic models, deterministic models of disease spread have continued to evolve in complexity and detail. Kermack and McKendrick's early work [5–7] gave rise to compartmental models, in which the population under study is divided into two or more

*Andrea.Allen@uvm.edu

states. Perhaps the most widely known of these models is the susceptible-infectious-recovered (SIR) model, where the population is divided into susceptible, infectious, and recovered states (or compartments) and the trajectory of the sizes of each compartment can be tracked analytically over time [8,9]. The standard compartmental model assumes homogeneous mixing of the population and is deterministic, meaning that a given set of initial conditions and disease transmission rates always leads to the same expected outcome. A common extension to compartmental models is to relax the assumption of homogeneous mixing. One method for doing so is to derive mean-field equations for an epidemic process over contact networks, thereby introducing heterogeneous structure into the population [10]. Similarly, it is possible to partition the population based on traits such as age, risk behaviors, or location and define how these partitions mix [11–14]. While these approaches introduce more realistic contact behavior into a model, they fail to account for the inherently stochastic nature of disease spread; something of particular importance early in an outbreak.

Models based on stochastic processes address the shortcoming of deterministic outcomes in the standard mean field compartment models. A commonly used approach is that of branching processes. Bienayme-Galton-Watson processes are one widely used example, as they provide a good approximation of more general stochastic epidemic models [15]. Beyond Bienayme-Galton-Watson processes, there exist a number of extensions such as including population structure, multiple types of hosts/pathogens, and considering time to be continuous rather than discrete [16,17]. In these branching process models the basic reproduction number R_0 , the probability of an outbreak, and the final proportion of population infected (in a “supercritical” model) are typically tractable to compute. While these are all important, a shortcoming of most branching models is the difficulty of tracking the trajectory of outbreaks through time and knowing whether it matches the continuous time dynamics of real epidemics. Stochastic differential equations are an alternative modeling approach that allow one to track outbreak trajectories, as well as often finding threshold conditions for the occurrence of an outbreak or the existence of an endemic equilibrium [18–20]. Like all models, stochastic differential equations have drawbacks; the most relevant is standard formulations do not allow for stochastic extinction if $R_0 > 1$.

Another common approach in disease modeling is times series analysis, more statistical in nature than mechanistic models. This theory can be applied to assist in estimating the parameters of compartmental models or to combine ensembles of compartmental models to increase prediction accuracy [21,22]. Independently of compartmental models, time series analysis can be used to study covariates of disease occurrence (e.g., weather), estimate the future variability in observed cases, or to make epidemic forecasts [23–25]. A necessary requirement for the effective use of many time series methods however is data. When facing sparse incidence numbers, and in the absence of historical data, the methods become problematic and thus are not suitable for emerging diseases.

Agent-based models are another family of models used for tracking epidemic progression, in which agents, or individuals in the population, are tracked throughout the course of the

epidemic. Agents are parameterized with individual attributes, capturing the heterogeneity of the population and aspects from compartmental models are used to categorize the state of each agent [26,27]. While there is great power in adjusting various attributes for different epidemic conditions and environmental factors, most of these models are computationally expensive and need a copious amount of information to generate the entire collection of agents [26–31], making them ill-suited for modeling early epidemic spread with a handful of cumulative case counts and sparsely available data.

Early in an outbreak, we often face the unique challenge of modeling disease spread while taking into account the heterogeneity of the population and the stochastic nature of disease spread, including stochastic extinction, without substantial amounts of data. The heterogeneous contact structure found in populations is accounted for by network models, and a first approximation for a relevant contact structures in a novel outbreak can be taken from past outbreaks of similar diseases. Including a sufficient number of possible states will typically account for heterogeneity in host and pathogen type. The randomness of transmission is modeled with stochastic processes, many of which easily permit stochastic extinction.

The above considerations naturally lead to percolation theory, which can be used to analyze stochastic compartmental disease models on networks. Percolation models unite contact heterogeneity and stochasticity under a single modeling framework [32]. An underlying contact network acts as the substrate for disease to propagate through, resulting in a directed network of transmission [33–35]. The resulting epidemic percolation networks can be analyzed using branching process theory [36,37], which model stochastic transmission between individuals using an underlying offspring distribution. Branching processes are especially useful for early epidemic modeling, as they allow for stochastic behavior of spread as well as stochastic extinction [38]. Specifically, the method of probability generating functions (PGFs) can be used to analyze branching processes on percolation networks [37–39]. Consequently, there have been many recent applications of this framework designed specifically for COVID-19 [40–45].

The PGF formalism is traditionally used for estimating quantities that pertain to the predicted end of an epidemic—such as the probability of infecting a macroscopic fraction of the population and distribution of final outbreak sizes—but not how risk and outbreak sizes change dynamically over time. Kenah and Robins show how modified percolation models (epidemic percolation networks) have a final state isomorphic to a network-based SIR models [33]. Most bond percolation frameworks differ from SIR dynamics as SIR transmission events are correlated through the distribution of the infectious period of each infected individual whereas percolation models assume independent contacts and transmission events. More importantly, percolation models integrate over time to map transmission dynamics (which occur in continuous time) to discrete bond percolation (which occur in discrete time with a fixed probability of transmission).

In 2009, Noël *et al.* [46] offered a novel method for tracking the stochasticity of outbreak sizes by epidemic generations, allowing us to incorporate discrete time into the percolation-framework model. In this paper, we show how

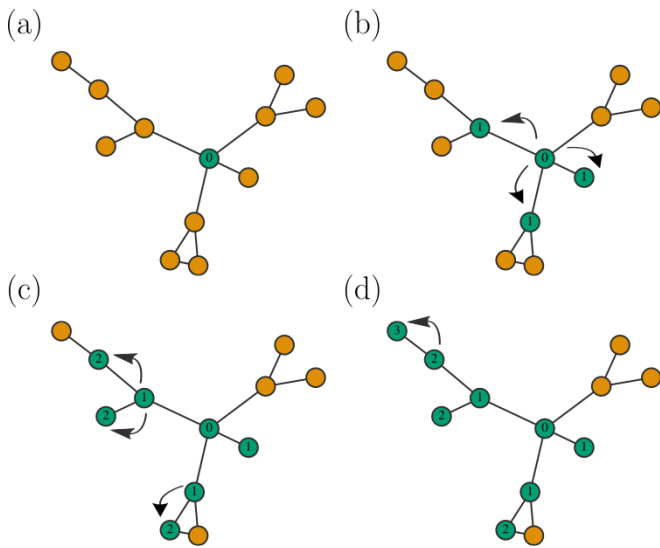


FIG. 1. Schematic of generations of infection through a network. Each node’s label corresponds to the epidemic generation in which it was infected. The initial infected node is in generation 0, any nodes they infect constitute generation 1, and so on.

the generation-based PGF formalism also succeeds in tracking emerging epidemic size in *continuous* time, by validating the PGF approach with event-driven simulations on networks. This result allows us to use PGFs and early disease data to quantify epidemic risk and survival probability.

II. THEORETICAL ANALYSIS AND SIMULATIONS

A. Probability generating functions

PGFs succinctly encode a probability distribution in a power series representation so that the methods of power series analysis can be applied [47]. PGF theory naturally extends to disease modeling, where the distribution under study encapsulates a disease transmission network, framed as a bond percolation problem where the bond occupation probability T is the probability of an infected individual infecting one of their contacts over the course of the entire epidemic [37,39]. Typically, this approach is used to solve for the average behavior of the system; we can solve for quantities such as the critical transmissibility at which the entire connected population will become infected, or the distribution of outbreak sizes. However, an increasing necessity of disease modeling is to model early epidemic spread, analyzing early cases to predict whether an outbreak will become large before it actually happens. In 2009, Noël *et al.* [46] developed the epidemic PGF modeling theory further to model the sizes of progressive epidemic *generations*, demonstrated in Fig. 1.

The foundations for both aforementioned generating function methodologies are the same, beginning with the underlying contact network. In a contact network, we represent a collection of individuals as *nodes* and their contacts between each other with *edges*. We say that two nodes are *neighbors* if they are in contact, i.e., connected by an edge. A node’s *degree* is how many neighbors it has. The *degree distribution* of a network is the probability distribution for the number of neighbors of one node. Under an SIR disease modeling

framework, nodes begin as *susceptible*, and become *infectious* if it is infected by one of its neighbors, which occurs with probability T .

The framework introduced by Noël *et al.* uses PGFs to describe generations of infection as a piece-wise generating function, which can then be studied using branching process techniques. First we introduce what an epidemic *generation* is. We say a node belongs to generation g if it became infected via a neighbor belonging to generation $g - 1$. Assuming an infinite-size random network drawn from a specific degree distribution (a process known as the configuration model [48]), each chain of infections stemming from an initial infected case, *patient zero*, can be considered uncorrelated. This uncorrelated assumption follows from configuration models having locally treelike structure, thus every subsequent case to be treated as a node that was reached by following a random edge. In this way, each node in each generation can be treated as independent from all other nodes in its generation. Thus, for each node in generation g , the PGF describing the distribution of cases that node will cause over the course of the epidemic is given by

$$G_g(x; T) = \begin{cases} G_0(x; T) & (g = 0) \\ G_1(x; T) & (g > 0) \end{cases} \quad (1)$$

where $G_g(x; T)$ is the distribution, in PGF notation, of the *secondary* cases caused by a single node in generation g . Now, we will provide the derivations used to obtain this framework using the underlying network, generating functions, and branching process theory.

Using PGF notation, we will refer to the original underlying network degree distribution as $G_0(x)$, which we write as

$$G_0(x) = \sum_{k=0}^{\infty} p_k x^k. \quad (2)$$

The k th coefficient of Eq. (2) p_k is the probability of randomly choosing a node with degree k from the network. The average degree of the network is denoted as $\langle k \rangle$, derived by the first derivative of the generating function as

$$G'_0(1) = \langle k \rangle = \sum_{k=0}^{\infty} k p_k. \quad (3)$$

To study the progression of an epidemic, we are interested in the distribution of infections from each subsequently infected node. Before introducing transmission probability, we work first with the aforementioned degree distribution to understand how many infections each node could cause through each generation. Assuming an initial infectious node, *patient zero*, we know $G_0(x)$ is the distribution of contacts for them, but that distribution is different for anyone *patient zero* infects. This phenomenon is known as the *friendship paradox*; the degree of a node chosen by following a random edge is on average, larger than the degree of the node selected at random whose edge we followed. In this context, *patient zero* has a degree distribution of $G_0(x)$, but the node who *patient zero* first infects has a degree distribution known as the *excess degree distribution*, denoted as $G_1(x)$ in PGF notation. To obtain $G_1(x)$, we are interested in the degree of nodes provided that

we arrive there by following the edge from one of its neighbors. So, this means the resulting distribution will exclude that neighbor, reducing every node's degree by 1, and multiplied by the number of ways they could have been reached, which is the original degree. This algorithm surmounts to taking the derivative of $G_0(x)$, so that we have the excess degree distribution

$$G_1(x) = \frac{\sum_k (k+1)p_{k+1}x^k}{\sum_k (k+1)p_{k+1}} = \sum_{k=0}^{\infty} q_k x^k \quad (4)$$

and where the derivative is divided by the average degree of the network $\langle k \rangle$ in order to normalize the distribution tuned to the original node. The coefficients q_k represent the probability of reaching a node with degree k from a randomly chosen edge.

Returning to the percolation problem, we incorporate disease transmissibility T to transform the excess degree distribution into a *secondary case distribution*. The probability that a single infectious node infects l neighbors given it has degree k , or k neighbors, is given by

$$p_{l|k} = \binom{k}{l} T^l (1-T)^{k-l}. \quad (5)$$

From this we can derive the PGF for the number of infections caused by patient zero, which we denote $G_0(x; T)$ for short, given by

$$\begin{aligned} G_0(x; T) &= \sum_{l=0}^{\infty} \sum_{k=l}^{\infty} p_k p_{l|k} x^l \\ &= \sum_{k=0}^{\infty} \sum_{l=0}^k p_k \binom{k}{l} T^l (1-T)^{k-l} x^l \\ &= G_0(1 + (x-1)T). \end{aligned} \quad (6)$$

From $G_0(x; T)$, $G_1(x; T)$ can be calculated in a parallel fashion as $G_1(x)$ is from $G_0(x)$. The PGF $G_1(x; T)$ is now the probability distribution of the number of infections caused by a single node, i.e., the secondary case distribution.

We now present how to study the evolution of the distribution of cumulative cases for the percolation model following Noël *et al.* Let s be the number of cumulative cases at generation g and let m be the number of infectious nodes strictly belonging to generation g . (Note that in this way, s is the sum of all m values from generation 0 up to and including generation g .) We let the probability of having s total infections by the end of the g th generation with m becoming infected (and thus being infectious) during that generation be denoted as ψ_{sm}^g [46]. This has an associated probability generating function, given by

$$\Psi_0^g(x, y) = \sum_{s,m} \psi_{sm}^g x^s y^m \quad (7)$$

over all s, m .

We know the distribution of infections following from a single infectious node in generation $g-1$ is generated by $G_{g-1}(1 + (x-1)T)$ [from Eq. (6)]. The PGF of a finite sum of independent processes is the product of their PGFs, and as discussed above, each node in generation $g-1$ can be treated independently. Thus, if we assume the state in

generation $g-1$ is given by the pair (s', m') , then the probability of spawning m new infectious nodes in generation g is generated by

$$\sum_m P(m|s', m') x^m = [G_{g-1}(x; T)]^{m'} \quad (8)$$

where the equality occurs as a result of the right side describing the probability of m infectious nodes in generation g assuming m' such nodes at $g-1$ from branching process theory.

For a given state (s', m') in generation $g-1$, m new infections will result in $s' + m$ cumulative infections in generation g . So, having m new infections occurs with probability $\psi_{s'm'}^{g-1} P(m|s', m')$, where the $\psi_{s'm'}^{g-1}$ term is the probability of being in the state (s', m') at generation $g-1$. Now, we can rewrite the entire PGF for the state space of (s, m) at generation g as

$$\begin{aligned} \Psi_0^g(x, y) &= \sum_{s,m} \psi_{sm}^g x^s y^m = \sum_{s',m} \psi_{s'm}^g x^{s'} (xy)^m \quad (9) \\ &= \sum_{s',m'} x^{s'} \sum_m \psi_{s'm'}^{g-1} P(m|s', m') (xy)^m \\ &= \sum_{s',m'} \psi_{s'm'}^{g-1} x^{s'} \sum_m P(m|s', m') (xy)^m \\ &= \sum_{s',m'} \psi_{s'm'}^{g-1} x^{s'} [G_{g-1}(xy; T)]^{m'} \\ &= \Psi_0^{g-1}(x, G_{g-1}(xy; T)). \end{aligned} \quad (10)$$

This defines a recurrence relation when $g \geq 1$. Taking $\Psi_0^0 = xy$ as the assumption that there is only one initial infectious individual, then $\psi_{sm}^0 = \delta_{s1} \delta_{m1}$.

Our primary focus in this paper will be on the distribution of cumulative infections s in each generation g . We derive a generating function for this quantity by taking the marginal distribution over y of Eq. (9). We let the coefficient p_s^g be defined as the probability of having s cumulative cases at generation g . To derive p_s^g , we wish to take the sum over all values of m for which the state s, m holds at generation g . To do so, we set the counting variable y of new cases simply equal to 1. As such, the coefficients p_s^g are generated by

$$\Psi_0^g(x, 1) = \sum_{s,m} \psi_{sm}^g x^s = \sum_s \sum_m \psi_{sm}^g x^s = \sum_s p_s^g x^s. \quad (11)$$

Now the generating function in Eq. (11) defines a probability distribution over s for each generation g , and is our main quantity under study. The analytical distributions are illustrated in Fig. 2 along with event-driven simulations to validate the theory.

B. Simulations of continuous SIR dynamics

For a realistic model of the spread of disease in a population, we simulate a stochastic disease process of an SIR epidemic on synthetic contact networks in continuous time [49]. We use an event-driven framework, which is advantageous for epidemic modeling, because it is much faster compared to a brute-force time-step simulation due to its

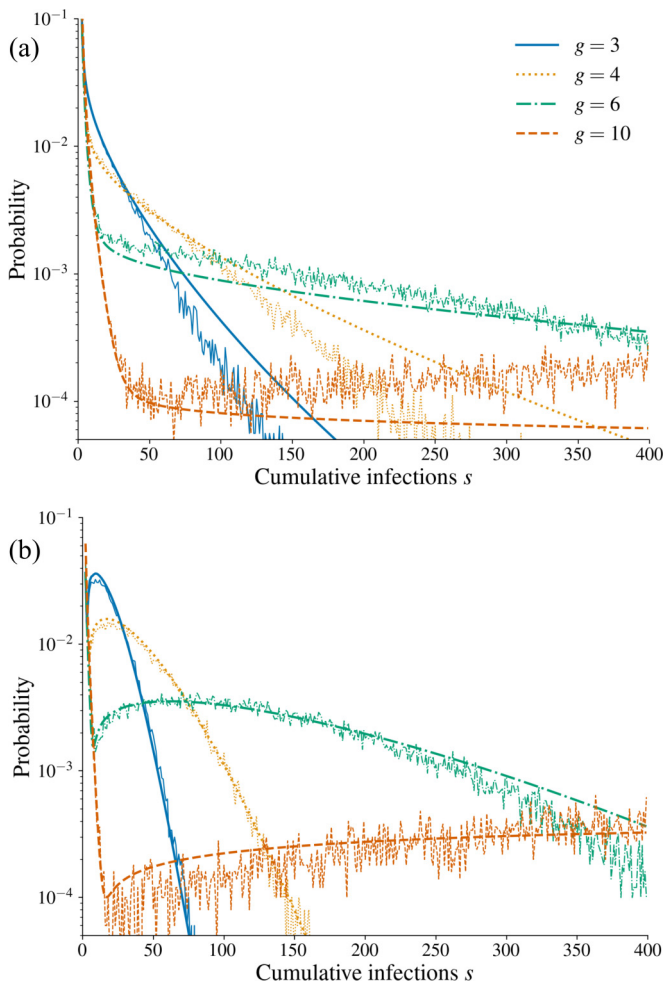


FIG. 2. Time evolution of epidemics on homogeneous and heterogeneous networks. We show the probability of having s cumulative cases by and including generation g for select generations [Eq. (11)]. Panel (a) shows the results on a modified power-law random networks with degree distribution given by $p_k = k^{-2}e^{-k/10}$ with average degree $\langle k \rangle = 1.79$, average excess degree $\langle q \rangle = 3.04$, $\beta = 0.004$, and $\gamma = 0.001$ such that $R_0 = T \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = \frac{\beta}{\beta + \gamma} \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = 2.44$. The smooth lines show the theoretical prediction for the probability distribution of cumulative infections. The distributions are validated by 75 000 simulations performed on 150 random network realizations with 10 000 nodes, following the process outlined in Sec. II B. Panel (b) shows the results of equivalent analysis and simulations on Erdős-Rényi random networks with $\langle k \rangle = 2.5$, $\beta = 0.004$, and $\gamma = 0.001$ such that $R_0 = 2.0$.

leveraging of the Markovian dynamics of infectious and recovery periods of individuals [50–52]. Recall in the SIR model that nodes inhabit the susceptible, infectious, and recovered states as the disease progresses, where nodes become infected if one of their infectious neighbors transmits to them. The standard SIR model is governed by two rate parameters; β , the rate per unit time of an infectious node transmitting to other nodes, and γ , the rate per unit time of an infected node recovering. In a continuous time event-driven simulation, infection, and recovery are Poisson processes occurring at rates β and γ respectively, and relate back to the percolation framework by defining transmissibility $T = \beta/(\beta + \gamma)$.

We draw a random network from a given degree distribution, and begin the simulation algorithm by assuming a random initial infectious node, patient zero, with degree k_0 . Patient zero could either recover before transmitting to any of its neighbors, or infect one or more of its neighbor nodes. The stochastic process governing the behavior of a single infected node is the superposition of $\hat{k} + 1$ Poisson processes, where \hat{k} is the number of susceptible neighbors, and with one extra process governing the time until recovery. Say patient zero infects Neighbor 1, who has k_1 neighbors. Then with two infectious nodes, the stochastic process encompassing all possible events is a Poisson process with rate $(\hat{k}_0 - 1)\beta + \hat{k}_1\beta + 2\gamma$, and so on as more nodes become infected.

Each possible event given by the subprocesses is the first to occur with probability $i/(\hat{k}\beta + \gamma)$ where $i \in \{\beta, \gamma\}$, with the Poisson process rate term from \hat{k} reducing if an infection event occurs, and stopping entirely if the contagious node recovers. The disease process for the whole population is a natural extension of that described above, with each node assumed identical apart from degree. The evolution of the unmitigated disease process from here is intuitive, either eventually all the infectious nodes recover or the whole connected population becomes infected.

Computationally, the above process is simulated by generating a random network from a given degree distribution using a large enough number of nodes N , such that average degree $k \ll N$. As we cannot simulate numerically on an infinite network, the best choice for N is the largest value the numeric simulation can support. A node is randomly selected to be patient zero, and the disease spread proceeds via stochastic event-driven simulation, often known as the Gillespie algorithm [53]. Continuous time is tracked using a random variable τ , known as the waiting time, which is exponentially distributed with parameter the sum of the rates of all the potential infection and recovery events. Each competing process is the first to occur with probability of its own rate divided by the sum of all rates of that process type, as described by the Poisson process above. The simulation is advanced via this algorithm until either there are no more infectious nodes or until there are no more susceptible nodes, and allows for obtaining the resulting evolution of the disease spread in terms of both generations of infection and continuous time.

III. RESULTS

We employ the generational size distribution theory to explore the evolution of epidemic size on a variety of network structures, and compare the generating function theory against continuous-time simulations. We use the event-driven simulation framework so that we can track the progression of the epidemic in both continuous time as well as the generation sizes corresponding with the branching process, which allows us to validate the theoretical distributions, as well as introduce a preliminary prediction for the expected continuous time emergence of successive generations. Then, we use the PGF framework to measure the probability of an epidemic surviving, or continuing on, past an arbitrary generation, depending on the characteristics of the network and disease.

A. Time evolution on homogeneous and heterogeneous networks

In Fig. 2 we show the probability distributions of cumulative infections by the specific generation for two network models. It is noteworthy that this modeling method holds for configuration model networks with varying types of degree distributions. Here, we show the results on a modified power law network and an Erdős-Rényi (ER) network both used in Ref. [46]. The ER network has mean degree and excess degree $\langle k \rangle = \langle q \rangle = 2.5$, while the modified power law has mean degree $\langle k \rangle = 1.79$ and average excess degree $\langle q \rangle = 3.04$, a more heterogeneous distribution. We demonstrate that the distributions of outbreak size appear to be more a result of the stochastic nature of the disease spread, rather than the structure of the network, though the structure does play a role in the shape of the distribution.

Our results convey that there is not one clear trajectory of a typical large outbreak, in contrast to traditional results with deterministic modeling. Instead, the stochastic nature of epidemic size is captured by a long tail in the distribution of cumulative cases over each epidemic generation. One unique aspect of this paper is that we validate this result using continuous-time simulations showing the same shape and long tail in outbreak size distributions as our analytical results. We do anticipate the simulated distributions and analytical distributions to vary from each other due to a few factors including the finite-size effects of simulated networks, and the fact that we compare a discrete analysis with a continuous-time process, but the general behavior appears consistent throughout the different generations.

We also find that on both the heterogeneous network and the homogeneous network, there is a high probability of an outbreak going extinct before growing large, however, if it does take off, the distribution levels off over the space of epidemic size. That is to say, if indeed an epidemic takes off and has arrived at generation six, via a transmission chain of length six, there is an almost equal probability of having anywhere from 50 to 500 cumulative cases by the time generation six is reached. We emphasize that these results display the unpredictability in early stages of epidemics, even ignoring the difficulty of estimating model parameters, it is near impossible to infer with much confidence how many infections there may actually be in the population.

B. Generations of infection in continuous time

While the behavior of the epidemic in our formalism is described by generations of infection, most applications of disease models desire descriptions of the dynamics in continuous time. We find early agreement from our model of generational infections with a distribution in continuous time, described in terms of the expected time of emergence of an arbitrary generation g . The agreement is surprising since one might not expect a consistent relationship between a generation number and the expected time of its emergence given the observed heterogeneity of early spread in Fig 2. Yet, by defining the *emergence* of generation g as the time its first member is infected, we find a simple linear relationship that allows us to map the PGF framework to continuous time.

We can show that the expected time of emergence of an arbitrary generation g is given by

$$\mathbb{E}[t(g)] = \frac{g}{\langle q \rangle \beta}$$

where $\langle q \rangle = G'_1(1)$ is the average excess degree of the network. We arrive at this expression for $\mathbb{E}[t(g)]$ via a simple argument over the Poisson process governing how nodes in generation $g - 1$ can lead to the first cases of generation g . Each node of generation $g - 1$ can recover at rate γ but also has on average $\langle q \rangle$ neighbors they can infect at rate β . Therefore, the first event around them will occur at a combined rate $\alpha = \langle q \rangle \beta + \gamma$ and will lead to a case in generation g with probability $T_q = \langle q \rangle \beta / (\langle q \rangle \beta + \gamma)$. The first infectious node in generation $g - 1$ can therefore lead to the emergence of generation g after $1/\alpha$ with probability T_q ; if not, or the second node in generation $g - 1$ could lead to the emergence of generation g with probability $T_q(1 - T_q)$ after $2/\alpha$ (approximate delay between the first and second node of generation $g - 1$ plus the expected time to generation g); and so on for the third node and beyond. This sequence of possibilities can be summarized by an arithmetico-geometric sum,

$$\begin{aligned} \mathbb{E}[t(g) - t(g-1)] &= \frac{T_q}{\alpha} \sum_{k=1}^{\infty} (1 - T_q)^{k-1} k \\ &= \frac{T_q}{\alpha} \frac{1}{T_q^2} = \frac{1}{\langle q \rangle \beta}. \end{aligned} \quad (12)$$

In Fig. 3, we demonstrate in practice how the expected time of emergence of consecutive generations falls in line with the predicted time measure. To show intuitively why we see this phenomenon, we show the time evolution of the active epidemic generations. We track time in two ways; in continuous time following the event-driven process discussed in Sec. II B, and also in terms of the expected time of emergence of each generation g , in the form $t = g/\langle q \rangle \beta$. We define a generation to be *active* if it contains one or more nodes who are not recovered and have susceptible neighbors at time t in the simulation. We illustrate the number of total and active generations over time, as well as the number of active *nodes* belonging to each generation, which helps clarify the roles each generation plays in causing the next wave of infection over a given interval in continuous time.

Having an understanding of the time at which a generation will emerge acts as a complement to the probabilities of extinction and cumulative cases discussed in Secs. III A and III C. Equipped with the distributions describing the stochasticity of outbreaks, the expected time mapping can be a tool for analysis of the dynamics of the worst-case scenarios when an outbreak does occur.

C. Probability of pandemics or stochastic extinction

The PGF generational theory can also be used to measure the probability that an emerging epidemic has a chance of dying off on its own, or “surviving”. Deterministic models always predict that an epidemic will occur if $R_0 > 1$, that is, if the average number of secondary infections caused by an infectious individual is more than one. In reality, there is a nonzero chance the outbreak will die off by chance, shown in

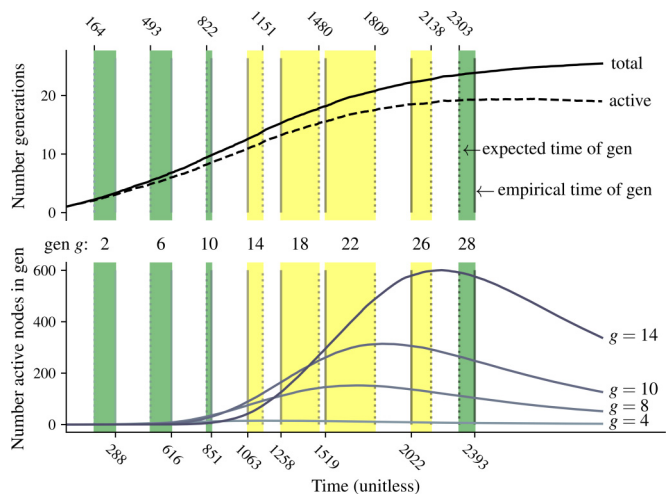


FIG. 3. Time evolution of the active epidemic generations and emergence times. Top panel (curves): Average number of total and active generations at time t for the modified power-law network with degree distribution $p_k = k^{-2}e^{-k/10}$. Bottom panel (curves): Average number of active nodes belonging to each generation shown over time to accompany the top panel. The tick marks in the top panel (and dotted-vertical lines) correspond to increments of $g/(q)\beta$, the predicted generational emergence times, and the bottom tick marks (and solid-vertical lines) correspond to the average empirical time at which that generation g emerged, for an example network. If the average time of emergence was greater than its respective $g/(q)\beta$ value, that is, after the predicted time, the difference is highlighted in green. If the average empirical time was less than that predicted, the difference is highlighted in yellow.

Fig. 4. Branching process models have been used in theoretical epidemiology for estimating such probabilities [56–58]. However, simple branching process models are Markovian in the number of active infections m . This is problematic in an applied setting as cumulative cases s are often the available data. Moreover, we show that conditioned on reaching generation g , the probability of the outbreak going extinct after generation g rather than becoming an epidemic is path dependent in the sense that the value of s at g changes the extinction probability, shown in Fig. 5.

To utilize the extinction probabilities, we want to look specifically at the variable ρ_s^g , the probability that given s cumulative cases at generation g the epidemic will go extinct, or die off, sometime afterwards. Given that the evolution of m occurs as a branching process with the offspring PGF given by Eq. (6), one can easily compute the probability of extinction of a single infection chain p_e as the solution of $p_e = G_1(p_e; T)$ using branching process theory [38]. The distribution of probabilities of reaching (s, m) in the state space for each g is given by ψ_{sm}^g , as discussed in Sec. II A. We define a new distribution, that of the probability of the outbreak still being in existence in generation g , by

$$\tilde{\psi}_{sm}^g = \begin{cases} \frac{\psi_{sm}^g}{\sum_{s', m' > 0} \psi_{s'm'}^g} & m > 0 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

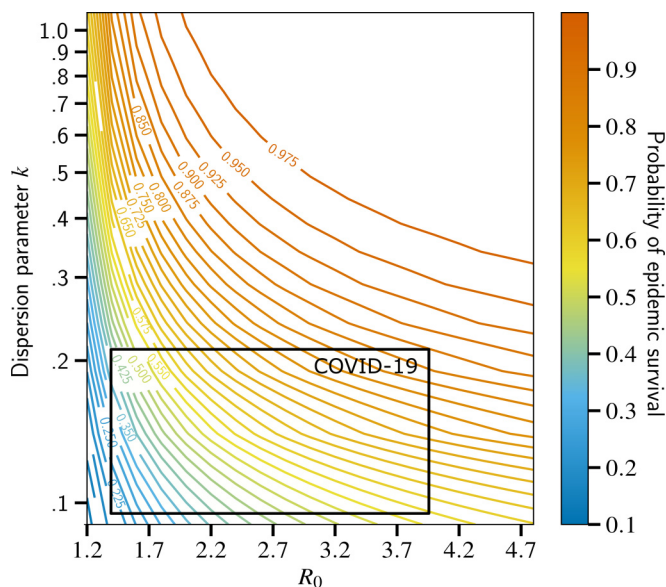


FIG. 4. Probability of epidemic survival as a function of contact structure. The contour plot shows the initial probability of epidemic survival for negative binomial distributions of infections over a range of possible R_0 values (average transmissions per case) and dispersion parameter k (inverse of heterogeneity). The box highlights estimates for COVID-19 based on data from Wuhan, China [54]. We assume an epidemic generation of $g = 4$ and $s = 16$ cases, which corresponds to the epidemic growing from 1 case to 16 over 4 generations. Using a serial interval of 4 days, the average of the estimated range for COVID-19 [55], this tracks to roughly over two weeks of spread. Similarly, in the state of Washington, the first recorded case of COVID-19 occurred on January 21st, 2020 but following cases were only identified on February 19th and increased to 18 by March 2nd. This figure illustrates how these cumulative case data could have been used in real time with our theoretical tools to estimate epidemic risk.

Thus, ρ_s^g , the probability of the epidemic going extinct given it has arrived at s cases by generation g is given by

$$\rho_s^g = \sum_m \frac{\tilde{\psi}_{sm}^g}{\sum_{m'} \tilde{\psi}_{sm'}^g} p_e^m \quad (14)$$

The probability of epidemic survival for an epidemic being active in generation g with s cumulative infections is then given by $1 - \rho_s^g$. We illustrate an example of how the survival probabilities change depending on the underlying network and disease parameters in Fig. 4.

D. Epidemic probability and COVID-19 data

We now apply the epidemic survival probability theory to early incidence of COVID-19 cases in the US. This allows us to look at the evolution over time of public health risk, while taking into account the stochastic elements of the early spread. We assume a distribution of secondary infections parameterized as a negative binomial with R_0 , the basic reproductive number, and k , the dispersion parameter of the contact network [54]. Together, these parameters determine the average behavior of disease spread where k is responsible for the variation in secondary cases, in turn affecting the likelihood

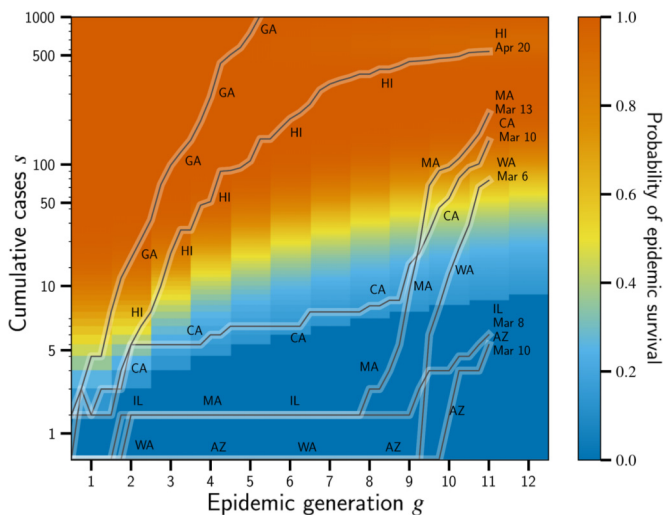


FIG. 5. Probability of epidemic continuing on as a function of case counts and time. As a simple comparison, we use early data from the COVID-19 pandemic and show a selection of U.S. states following unique timelines from the first recorded case onward. This simple visualization is not meant as a validation but only to explore how quickly our predictions for the probability of the epidemic not dying off changes as an epidemic grows. To calculate these probabilities, we use a negative binomial distribution of secondary infections with $k = 10^{-1}$, $R_0 = 2.5$ along with data from the COVID-19 Repository by the CSSE at Johns Hopkins [64,65]. The first data point for each state shown correspond to the first date on which 1 or more cases were recorded. Raw data of cumulative case counts are used, and plotted on the same range of epidemic generations for purposes of comparison, despite an evident variability in the duration of generation length. Using a serial interval of 4 days, progressive generations are shown along the horizontal axis (generation two corresponds to 8 days, for example). On the vertical axis, the cumulative case counts for each state are plotted. We see how a state’s proclivity to the epidemic taking off changes over the course of successive generations. Several states such as California, Massachusetts, and Washington had a lower probability of epidemic survival early on, then crossed the band into a higher likelihood over a short time span. Although the data used in this figure does not take into account factors such as missing count data, it serves as a visualization of how sharply the interplay of generation of epidemic and cumulative cases demarcate the probability of the epidemic continuing.

of superspreading events [59–61]. A low dispersion parameter k (high heterogeneity) means that a select few cases may cause the majority of secondary infections [62], which in our framework here might correspond to a single case leading to an extreme increase in cases in the next generation. For that reason, it is often assumed that the early spread of an epidemic is highly sensitive to superspreading events [63]. Yet, as shown in Fig. 2, heterogeneity in contact structure actually has less of an impact on the distribution of outcomes than the inherent stochasticity of transmission.

In Fig. 4 we show the probability of epidemic survival (that is, the probability of an epidemic continuing to grow) with a fixed generation $g = 4$ and fixed cumulative cases $s = 16$ over a range of R_0 and k values, highlighting parameter estimates for COVID-19 [62]. Despite the relatively low number of cases after several generations, clearly affected by the lack

of testing resources at the time, the chances of the epidemic stochastically dying out were already close to a simple coin flip. In Fig. 5, we show the inverse problems: fixing disease parameters and varying temporal variables. We set $R_0 = 2.5$ and $k = 0.1$, falling within the range of values for COVID-19, and track seven US states over time to observe where their disease progression state falls in the probability space of epidemic survival.

Guided by the results shown in Fig. 3, we proceed knowing that our model predicts generations to emerge in linear increments of time. We use the serial interval of 4 days, taken from the window for COVID-19 [55] to correspond with successive generational emergence. We observe that several states hovered around a low probability of epidemic survival at low early cases, but very quickly crossed to a much higher bracket where natural extinction of the disease spread is virtually impossible. The states of Washington and Massachusetts each took only two generations to cross from sub to supercritical epidemic survival probability, even derived from limited data and poor testing at the time. The extraordinary leap in epidemic probability from just one generation to the next explain, in part, why it was so hard for public health systems to react and adapt to the spread of COVID-19.

IV. DISCUSSION

Temporal models of disease spread often fall in one of three categories. (i) Compartmental models that are deterministic in nature as they rely on ordinary differential equations, where uncertainty only stems from our imperfect knowledge of model parameters, rather than from the inherent randomness of disease transmission. (ii) Complicated agent-based models that lose the tractability of analytical models, which require significant amount of data to parametrize and do not produce explicit likelihood of outcomes. (iii) Time series analyses that can produce probabilistic forecasts. This last approach can produce useful predictions by ignoring transmission mechanisms or contact structure, but that perspective also precludes it from evaluating potential interventions that affect individual parameters or contact structure.

In this paper, we have shown that analysis of branching processes often used to only study the final state of epidemic models can actually combine the strengths of these different approaches by including stochasticity, contact heterogeneity and even individual characteristics [33,39,66]. The reason this framework is usually used to solely predict the probability and final size of an epidemic is that the mathematical treatment involves integrating over contacts and therefore time [54]. However, we provided a first demonstration that the predictions made over generations by the branching process are actually very close approximation of continuous time epidemic dynamics on equivalent contact networks. This result alone justifies a large body of work and creates a foundation for analytical, probabilistic, epidemic forecasts based on PGFs.

Our probabilistic and temporal forecasts allowed us to uncover the diversity of epidemic courses, in the form of an unusually broad distribution of potential transmission trees over time. We have also shown that these flat distributions

emerge on both homogeneous (e.g., Erdős-Rényi graphs) and heterogeneous (e.g., scale-free) contact networks. This phenomenon is therefore driven by the stochasticity of disease transmission rather than by the complexity of the contact structure. This broad likelihood of early disease incidence justifies our use of a stochastic branching process, whereas deterministic models would typically track only the average or expected number of cases, which is a poor description of flat distributions.

Our framework currently rests on a few assumptions. By building our framework on a configuration model, we ignore potentially important structural correlations. The PGF framework itself can be extended, data permitting, to include such correlations like degree-degree assortativity [67,68], clustering [69,70], and more general structures [71,72]. All of these generalizations of the PGF framework still rely, at some level, on a treelike approximation, but this approach has been shown to capture most important network features [73].

We also assume that there are a finite number of active generations at any given time and that the distribution of contacts and transmission probability do not change over time. This first assumption was tested in Fig. 3 where we show that a simple network-based serial interval provides a reasonable approximation for time of emergence of epidemic generations in the continuous dynamics, illustrating both why and how we can align the generation-based branching process with the underlying temporal dynamics.

Our assumptions on the constant contact patterns and transmissibility provide a great road map for future work. In Eq. (1), we formulate our PGFs on a per generation basis, which would allow us to change these patterns over time to model adaptive behavior or top-down interventions (e.g., lockdowns limiting contacts or masks reducing transmissibility). Certain network interventions have been shown to alter the dynamics of epidemic outcomes in interesting ways, such as contact tracing [45,74] or vaccination roll-outs [75,76]. Specifically, when interventions are targeted around key individuals (e.g., hubs [77]) or affect different subset of the population differently [66], one can see the emergence of smeared transitions when epidemics mostly spread in specific subgraphs with subcritical spillover in other populations [78]. Modeling interventions under a generational PGF framework would provide probabilistic forecasts not only of disease

dynamics but also of the impact and timing of particular interventions.

Importantly, our results on the diversity of epidemic courses highlight how little information can actually be gathered from early incidence data. In Fig. 2, we see that the same disease in the same population can be roughly as likely to produce 40 or 400 cases after 10 epidemic generations.

Finally, our results on epidemic survival show how quickly a situation can move from an uncertain outbreak to supercritical exponential growth. Due to both the randomness of disease spread and the imperfect COVID-19 testing protocols from early 2020, most states in the US moved from below 20% survival probability of the epidemic to above 80% in about two epidemic generations (2 weeks or less).

Altogether, our results stress the danger of justifying a lack of intervention with slow trends in early disease spread data. Little can be learned about transmission mechanisms and dynamics from the first few epidemic generations. The distribution of epidemic courses is mostly driven by the inherent randomness of transmission, and the window in which the dynamics settle into their subcritical or supercritical behavior tends to be unfortunately narrow, which leaves little room for fast adaptive responses.

Faced now with emergence of variants of COVID-19 around the world, the current situation is reminiscent of the scenario in the state of Washington during January of 2020—sporadic clusters of cases with an unclear growth trajectory. We see from the data in Washington, as well as many other states and countries, how quickly cases explode and what that means for the likelihood of controlling the epidemic without external intervention efforts. Slow initial disease growth does not preclude a rapid increase shortly thereafter.

ACKNOWLEDGMENTS

A.J.A. and L.H.-D. acknowledge financial support from the National Institutes of Health 1P20 GM125498-01 Centers of Biomedical Research Excellence Award. M.C.B. is supported as a Fellow of the National Science Foundation under NRT Award No. DGE-1735316 and N.J.R. is supported by the University of Vermont. A.A. acknowledges financial support from the Sentinelle Nord initiative of the Canada First Research Excellence Fund and from the Natural Sciences and Engineering Research Council of Canada (Project No. 2019-05183).

-
- [1] First Travel-related Case of 2019 Novel Coronavirus Detected in United States, <https://www.cdc.gov/media/releases/2020/p0121-novel-coronavirus-travel-case.html>, 2020.
- [2] D. Oxley and J. Ryan, “Volatile and unpredictable”: Life Care Center speaks publicly for the first time since COVID-19 outbreak, *KUOW News and Inf.*, 2020, March 7, <https://www.kuow.org/stories/volatile-and-unpredictable-life-care-speaks-publicly-for-the-first-time-since-covid-19-outbreak>.
- [3] O. Sullivan, Coronavirus death toll rises to nine in Washington, *Kirkland Reporter*, 2020, March 3, <https://www.kirklandreporter.com/news/coronavirus-death-toll-rises-to-eight-in-washington/>.

- [4] A. Sundell, New coronavirus cases confirmed in Snohomish, King counties, *KING-TV*, 2020, February, <https://www.king5.com/article/news/health/coronavirus/washington-coronavirus-update/281-e73682dc-dad7-4b6e-b0ec-d2234ff9e2e0>.
- [5] W. O. Kermack and A. G. McKendrick, A contribution to the mathematical theory of epidemics. I, *Proc. R. Soc. London A* **115**, 700 (1927).
- [6] W. O. Kermack and A. G. McKendrick, A contribution to the mathematical theory of epidemics. II.—The problem of endemicity, *Proc. R. Soc. London A* **138**, 55 (1932).
- [7] W. O. Kermack and A. G. McKendrick, A contribution to the mathematical theory of epidemics. III.—Further studies of

- the problem of endemicity, *Proc. R. Soc. London A* **141**, 94 (1933).
- [8] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, Oxford, 1991).
- [9] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, Princeton, 2007).
- [10] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Epidemic processes in complex networks, *Rev. Mod. Phys.* **87**, 925 (2015).
- [11] H. Iwana, Threshold and stability results for an age-structured epidemic model, *J. Math. Biol.* **28**, 411 (1990).
- [12] W. Huang, K. L. Cooke, and C. Castillo-Chavez, Stability and bifurcation for a multi-group model for the dynamics of HIV/AIDS transmission, *SIAM J. Appl. Math.* **52**, 835 (1992).
- [13] B. Bolker and B. Grenfell, Space, persistence and dynamics of measles epidemics, *Phil. Trans. R. Soc. London B* **348**, 309 (2015).
- [14] A. L. Lloyd and V. A. A. Jansen, Spatiotemporal dynamics of epidemics: Synchrony in metapopulation models, *Math. Biosci.* **188**, 1 (2004).
- [15] F. Ball and P. Donnelly, Strong approximations for epidemic models, *Stoch. Process. Their Appl.* **55**, 1 (1995).
- [16] F. Ball, D. Mollison, and G. Scalia-Tomba, Epidemics with two levels of mixing, *Ann. Appl. Probab.* **7**, 46 (1997).
- [17] L. S. J. Allen, *Stochastic Population and Epidemic Models* (Springer, New York, 2015).
- [18] L. S. J. Allen, A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis, *Infect. Dis. Model.* **2**, 128 (2017).
- [19] W. Wang, Y. Cai, Z. Ding, and Z. Gui, A stochastic differential equation SIS epidemic model incorporating Ornstein–Uhlenbeck process, *Physica A* **509**, 921 (2018).
- [20] D. A. Gray, L. Greenhalgh, L. Hu, X. Mao, and J. Pan, A stochastic differential equation SIS epidemic model, *SIAM J. Appl. Math.* **71**, 876 (2011).
- [21] B. F. Finkenstädt and B. Grenfell, Time series modelling of childhood diseases: A dynamical systems approach, *Appl. Stat.* **49**, 187 (2000).
- [22] Z. Zhan, W. Dong, Y. Lu, P. Yang, Q. Wang, and P. Jia, Real-time forecasting of hand-foot-and-mouth disease outbreaks using the integrating compartment model and assimilation filtering, *Sci. Rep.* **9**, 2661 (2019).
- [23] R. Allard, Use of time-series analysis in infectious disease surveillance, *Bull. World Health Organ.* **76**, 327 (1998).
- [24] B. Lopman, B. Armstrong, C. Atchinson, and J. J. Gray, Host, weather and virological factors drive norovirus epidemiology: Time-series analysis of laboratory surveillance data in England and Wales, *PLoS One* **4**, e6671 (2009).
- [25] W. Hu, S. Tong, K. Mengersen, and B. Oldenburg, Rainfall, mosquito density and the transmission of Ross River virus: A time-series forecasting model, *Ecol. Model.* **196**, 505 (2006).
- [26] P. C. L. Silva, P. V. C. Batista, H. S. Lima, M. A. Alves, F. G. Guimarães, and R. C. P. Silva, COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions, *Chaos Solitons Fractals* **139**, 110088 (2020).
- [27] N. M. Gharakhanlou and N. Hooshangi, Spatio-temporal simulation of the novel coronavirus (COVID-19) outbreak using the agent-based modeling approach (case study: Urmia, Iran), *Inform. Med. Unlocked* **20**, 100403 (2020).
- [28] E. Cuevas, An agent-based model to evaluate the COVID-19 transmission risks in facilities, *Comput. Biol. Med.* **121**, 103827 (2020).
- [29] N. Hoertel, M. Blachier, C. Blanco, M. Olfson, M. Massetti, F. Limosin, and H. Leleu, Facing the COVID-19 epidemic in NYC: A stochastic agent-based model of various intervention strategies, *medRxiv* (2020), doi: 10.1101/2020.04.23.20076885.
- [30] A. Staffini, A. K. Svensson, U.-I. Chung, and T. Svensson, An agent-based model of the local spread of SARS-CoV-2: Modeling study, *JMIR Med. Inform.* **9**, e24192 (2021).
- [31] V. Srikrishnan and K. Keller, Small increases in agent-based model complexity can result in large increases in required calibration data, *Environ. Model. Softw.* **138**, 104978 (2021).
- [32] L. Meyers, Contact network epidemiology: Bond percolation applied to infectious disease prediction and control, *Bull New Ser. Am. Math. Soc.* **44**, 63 (2007).
- [33] E. Kenah and J. M. Robins, Second look at the spread of epidemics on networks, *Phys. Rev. E* **76**, 036113 (2007).
- [34] J. C. Miller, Epidemic size and probability in populations with heterogeneous infectivity and susceptibility, *Phys. Rev. E* **76**, 010101(R) (2007).
- [35] E. Kenah and J. C. Miller, Epidemic percolation networks, epidemic outcomes, and interventions, *Interdiscip. Perspect. Infect. Dis.* **2011**, 543520 (2011).
- [36] K. B. Athreya and P. E. Ney, *Branching Processes* (Springer-Verlag, Berlin, 1972).
- [37] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distributions and their applications, *Phys. Rev. E* **64**, 026118 (2001).
- [38] J. C. Miller, A primer on the use of probability generating functions in infectious disease modeling, *Infect. Dis. Model.* **3**, 192 (2018).
- [39] M. E. J. Newman, Spread of epidemic disease on networks, *Phys. Rev. E* **66**, 016128 (2002).
- [40] J. Levesque, D. W. Maybury, and R. H. A. D. Shaw, A model of COVID-19 propagation based on a gamma subordinated negative binomial branching process, *J. Theor. Biol.* **512**, 110536 (2021).
- [41] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge, The challenges of modeling and forecasting the spread of COVID-19, *Proc. Natl. Acad. Sci. USA* **117**, 16732 (2020).
- [42] I. A. Mitrofanis and V. P. Koutras, A branching process model for the novel coronavirus (Covid-19) spread in Greece, *Int. J. Mod. Opt.* **11**, 63 (2021).
- [43] L. Zhang, H. Wang, Z. Liu, X. F. Liu, X. Feng, and Y. Wu, A heterogeneous branching process with immigration modeling for COVID-19 spreading in local communities in China, *Complexity* **2021**, 6686547 (2021).
- [44] M. Akian, L. Ganassali, S. Gaubert, and L. Massoulié, Probabilistic and mean-field model of COVID-19 epidemics with user mobility and contact tracing, *arXiv:2009.05304*.
- [45] S. Kojaku, L. Hébert-Dufresne, E. Mones, S. Lehmann, and Y.-Y. Ahn, The effectiveness of backward contact tracing in networks, *Nat. Phys.* **17**, 652 (2021).
- [46] P.-A. Noël, B. Davoudi, R. C. Brunham, L. J. Dubé, and B. Pourbohloul, Time evolution of epidemic disease on finite and infinite networks, *Phys. Rev. E* **79**, 026101 (2009).

- [47] H. S. Wilf, *Generating Functionology* (CRC Press, Boca Raton, FL, 2005).
- [48] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, Configuring random graph models with fixed degree sequences, *SIAM Rev.* **60**, 315 (2018).
- [49] A. Allen, “andrea-allen/epintervene v1.0.3 EpIntervene Release (v1.0.3)”, Zenodo, doi: [10.5281/zenodo.5514401](https://doi.org/10.5281/zenodo.5514401) (2021).
- [50] I. Z. Kiss, J. C. Miller, and P. L. Simon, *Mathematics of Epidemics on Networks: From Exact to Approximate Models* (Springer, New York, 2019).
- [51] J. C. Miller and T. Ting, EoN (Epidemics on Networks): A fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks, *J. Open Source Softw.* **4**, 1731 (2019).
- [52] P. Bauer, S. Engblom, and S. Widgren, Fast event-based epidemiological simulations on national scales, *Inter. J. High Perform. Comput. Appl.* **30**, 438 (2016).
- [53] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions, *J. Phys. Chem.* **81**, 2340 (1977).
- [54] L. Hébert-Dufresne, B. M. Althouse, S. V. Scarpino, and A. Allard, Beyond R0: heterogeneity in secondary infections and probabilistic epidemic forecasting, *J. R. Soc., Interface* **17**, 20200393 (2020).
- [55] Z. Du, X. Xu, Y. Wu, L. Wang, B. J. Cowling, and L. A. Meyers, Serial interval of COVID-19 among publicly reported confirmed cases, *Emerg Infect Dis.* **26**, 1341 (2020).
- [56] N. G. Becker, Estimation for discrete time branching processes with applications to epidemics, *Biometrics* **33**, 515 (1977).
- [57] N. G. Becker, On parametric estimation for mortal branching processes, *Biometrika* **61**, 393 (1974).
- [58] O. Diekmann, H. Heesterbeek, and T. Britton, *Mathematical Tools for Understanding Infectious Disease Dynamics* (Princeton University Press, Princeton, 2013).
- [59] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz, Superspreading and the effect of individual variation on disease emergence, *Nature (London)* **438**, 355 (2005).
- [60] C. L. Althaus, Ebola superspreading, *Lancet Infect. Dis.* **15**, 507 (2015).
- [61] A. J. Kucharski and C. L. Althaus, The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission, *Eurosurveillance* **20**, (2015).
- [62] J. Riou and C. L. Althaus, Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020, *Eurosurveillance* **25**, 2000058 (2020).
- [63] B. M. Althouse, E. A. Wenger, J. C. Miller, S. V. Scarpino, A. Allard, L. Hébert-Dufresne, and H. Hu, Superspreading events in the transmission dynamics of SARS-CoV-2: Opportunities for interventions and control, *PLoS Bio.* **18**, e3000897 (2020).
- [64] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, <https://github.com/CSSEGISandData/COVID-19>, 2021.
- [65] E. Dong, H. Du, and L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, *Lancet Infect. Dis.* **20**, 533 (2020).
- [66] A. Allard, B. M. Althouse, S. V. Scarpino, and L. Hébert-Dufresne, Asymmetric percolation drives a double transition in sexual contact networks, *Proc. Natl. Acad. Sci. USA* **114**, 8969 (2017).
- [67] A. Vázquez and Y. Moreno, Resilience to damage of graphs with degree correlations, *Phys. Rev. E* **67**, 015101 (2003).
- [68] L. Hébert-Dufresne, A. Allard, J.-G. Young, and L. J. Dubé, Percolation on random networks with arbitrary k-core structure, *Phys. Rev. E* **88**, 062820 (2013).
- [69] M. E. J. Newman, Random Graphs with Clustering, *Phys. Rev. Lett.* **103**, 058701 (2009).
- [70] A. Allard, L. Hébert-Dufresne, P.-A. Noël, V. Marceau, and L. J. Dubé, Bond percolation on a class of correlated and clustered random graphs, *J. Phys. A* **45**, 405005 (2012).
- [71] B. Karrer and M. E. J. Newman, Random graphs containing arbitrary distributions of subgraphs, *Phys. Rev. E* **82**, 066118 (2010).
- [72] A. Allard, L. Hébert-Dufresne, J.-G. Young, and L. J. Dubé, General and exact approach to percolation on random graphs, *Phys. Rev. E* **92**, 062807 (2015).
- [73] S. Melnik, A. Hackett, M. A. Porter, P. J. Mucha, and J. P. Gleeson, The unreasonable effectiveness of tree-based theory for networks with clustering, *Phys. Rev. E* **83**, 036112 (2011).
- [74] A. K. Rizzi, A. Faqeeh, A. Badie-Modiri, and M. Kivelä, Epidemic spreading and digital contact tracing: Effects of heterogeneous mixing and quarantine failures, [arXiv:2103.12634](https://arxiv.org/abs/2103.12634).
- [75] G. Burgio, B. Steinegger, and A. Arenas, Homophily impacts the success of vaccine roll-outs, [arXiv:2112.08240](https://arxiv.org/abs/2112.08240).
- [76] T. Hiraoka, A. K. Rizzi, M. Kivelä, and J. Saramäki, Herd immunity and epidemic size in networks with vaccination homophily, [arXiv:2112.07538](https://arxiv.org/abs/2112.07538).
- [77] R. Cohen, K. Erez, D. ben Avraham, and S. Havlin, Breakdown of the Internet under Intentional Attack, *Phys. Rev. Lett.* **86**, 3682 (2001).
- [78] L. Hébert-Dufresne and A. Allard, Smear phase transitions in percolation on real complex networks, *Phys. Rev. Research* **1**, 013009 (2019).