Large deviations of a susceptible-infected-recovered model around the epidemic threshold

Yannick Feld [*] and Alexander K. Hartmann

Institut für Physik, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany

We numerically study the dynamics of the SIR disease model on small-world networks by using a large-deviation approach. This allows us to obtain the probability density function of the total fraction of infected nodes and of the maximum fraction of simultaneously infected nodes down to very small probability densities like 10^{-2500}. We analyze the structure of the disease dynamics and observed three regimes in all probability density functions, which correspond to quick mild, quick extremely severe, and sustained severe dynamical evolutions, respectively. Furthermore, the mathematical rate functions of the densities are investigated. The results indicate that the so-called large-deviation property holds for the SIR model. Finally, we measured correlations with other quantities like the duration of an outbreak or the peak position of the fraction of infections, also in the rare regions which are not accessible by standard simulation techniques.

I. INTRODUCTION

The modeling of the spread of epidemic diseases has always been a central aspect in statistics, applied mathematics, and statistical mechanics [1–5]. Due to the present outbreak of the SARS-CoV-2 pandemic, interest in this field has risen even more [6–12]. Disease spreading can be modeled in many different ways, e.g., with ordinary differential equations like in the mean-field version of the *susceptible-infected-recovered* (SIR) model [13] or with agent-based approaches [14,15]. Also other fields are involved, e.g., Bayesian analysis [16] to estimate model parameters [6,9] or machine-learning approaches [17] to predict the future development of an outbreak [10,18–20]. We refer to recent review articles for a good overview of the topic [21–23].

Given the large population of humans and animals on our planet and the high number of active and potentially threatening viruses or bacteria, the actual number of pandemic diseases is surprisingly small. Thus, the outbreak of a *specific* pandemic is actually a *rare event*, i.e., occurs, looking at each single type of disease, with a very small probability. For example, the disease might be very active in one population of, e.g., bats, but much rarer contacts or rare mutations are needed to allow the transfer to another population, like humans, maybe even requiring unknown intermediate animals. Hence, it is natural to consider the application of large-deviation approaches to study disease dynamics. So far this was done only a few times, e.g., the large-deviation principle was investigated analytically [24] by generalizing an approach of Ref. [25] for simple mean-field epidemic models.

A more realistic modeling of epidemic dynamics beyond mean-field level is generally obtained by studying the dynamics on networks [3]. These networks represent the contacts between the individuals or groups of individuals. This, of course, can become arbitrarily complex, e.g., by combining several network layers, which can then represent different environments of contact [26]. Depending on the structure of a network and on the epidemic parameters, like transmission probability and recovery probability, the infection of a single node might stay contained or might lead to a pandemic outbreak. The critical value of, e.g., the transmission probability, beyond which an pandemic outbreak occurs, i.e., a percolation of the infected nodes, is called the *epidemic threshold*. For not too complex models, the epidemic threshold of disease models can be analyzed by using a variety of analytical methods, e.g., the mean-field method, its quenched version, or dynamic message passing approaches. [27–33]. Naturally, for more complex models it is even harder to obtain analytical results, thus computer simulations [34] are applied instead.

To our knowledge, for the study of disease spreading on networks with respect to large-deviations and rare events no results are available, let it be analytical or numerical. Thus, to start to establish such approaches in the field of disease dynamics, here the simple case of the SIR model on standard networks drawn from a small-world [35] ensemble is considered. This is motivated by the fact that physical contact networks between humans resemble small-world-like networks [14]. However, the methods applied here can be used for all types of networks. We apply large-deviation techniques [36–38] that are based on the Markov-chain approaches Wang Landau [39] and entropic sampling [40]. In this way we are able to explore the probability density function (pdf) of the fraction of infected nodes down to values as small as 10^{-115}. For the pdf of the maximum of the fraction of simultaneously infected nodes, we reach values as small as 10^{-2500}. For both quantities, we look at the respective mathematical *rate functions*, to verify whether the *large-deviation principle* holds [41–44]. This gives a complete description of these stochastic quantities, over the full range of the support of the distributions. First, this is desirable from a fundamental research point of view. Second, we are able to investigate cor-

---
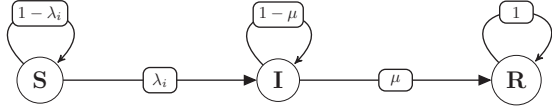
*yannick.feld@uol.de; https://www.yfeld.de.

FIG. 1. Showing transition probabilities for node i at a given time step.

relations between different quantities, e.g., how the fraction of infected nodes corresponds to how quickly the disease dies out. By using a large-deviation technique, we are able to study these correlations much beyond the typical behavior. Thus, we can also analyze extremely severe as well as extremely mild disease progression, and try to identify their possible causes through looking at their correlations. The much broader understanding gained in this way could be one piece to help to better prevent pandemic outbreaks in the future, in particular, if it is applied not to the general model but for a specific case tailored to the epidemic under scrutiny, respectively.

The paper is organized as follows: First, the SIR model is introduced and its dynamics and the main measurable quantities are defined. Next, we define the ensemble of networks we use. In the main methodological section, we present the algorithms used for sampling the rare events and how we have to set up the simulation of the SIR dynamics to embed it into the large-deviation scheme. Our results come in three parts. We begin by investigating the ensemble with standard techniques to identify interesting points in parameter space. For these points, large-deviation simulations are performed to obtain the distributions of the total fraction C of infected individuals and of the maximum fraction M of simultaneously infected individuals, respectively. We finish with a summary and an outlook.

II. SIR MODEL

Let there be a given *connected* network with N nodes, where the nodes represent individuals and the edges contacts. The term *connected* here means that there is only one connected component, i.e., all nodes can be reached from all other nodes through paths along edges. Each node is in one of the three states *susceptible* (S), *infected* (I), or *recovered* (R).

For any given configuration of states in the network, at each time step a node can change its state as follows: The probability of an infected node infecting a specific susceptible neighbor is given by the *transmission probability* $\lambda = $ const. The probability of an infected node recovering in a given time step is given by $\mu = $ const. A node in the recovered state remains recovered forevermore.

We consider a node i in S state, which has A_i adjacent infected nodes. Since each infected neighbor has a probability of λ to infect node i, the probability for node i to become infected in a time step is

$$\lambda_i = 1 - (1 - \lambda)^{A_i}. \tag{1}$$

All possible transitions between the states of a node are shown in Fig. 1.

For all disease dynamics we consider, as initial state at discrete time $\tau = 0$, one particular node (in the following called node 0) set to the infected state, while all other nodes

are susceptible. Our simulations [34] are performed at discrete times $\tau \rightarrow \tau + 1$ by applying the above mentioned rules in a parallel fashion to all nodes. This is repeated until the disease dies out, i.e., no infected nodes remain, or if a maximum chosen time is reached. Such a development we call the *time evolution of an outbreak* from here on out. Note that the outbreak might be very small, with just node 0 being initially infected and recovering after some time before any other node is infected. Clearly, unless $\lambda \ll \mu$, this will not occur too often.

Due to the probabilistic nature of the problem, multiple outbreak simulations will generally lead to different results.

To describe the time evolution of an outbreak, let us introduce a few quantities: Let $s(\tau)$, $i(\tau)$, and $r(\tau)$ be the fractions of susceptible, infected, and recovered nodes at time step τ, respectively. Let $c(\tau) = i(\tau) + r(\tau)$ be the fraction of the total, i.e., cumulative infections, which have occurred up to time step τ. These quantities depend on the time step. To describe the global characteristics of an outbreak, the following two quantities are introduced:

$$C = \max_\tau [c(\tau)] \equiv c(\infty) \tag{2}$$

describes the fraction of the network that caught the disease during the outbreak and is therefore a measure for its severity. This is the standard quantity to distinguish between a local outbreak and a pandemic.

$$M = \max_\tau [i(\tau)] \tag{3}$$

denotes the peak fraction of nodes that happen to simultaneously be in the infected state during an outbreak and is, therefore, a relevant quantity for the health care system.

III. ENSEMBLE

In our work, we investigate a small-world network ensemble [35,45–47], because contact networks between individuals are highly connected small-world-like networks [14]. We use the same implementation as we used in Ref. [48].

The network is constructed as follows: Let there be N nodes $i = 0, \ldots, N - 1$. First, the nodes are arranged in a ring structure, meaning every node is connected to its next and second next neighbor by the edges $\{i, i+1\}$ and $\{i, i+2\}$ (nodes N and $N + 1$ are identified with index 0 and 1, respectively).

To gain small-world characteristics, next some of the edges created in the first step are made *long-range*, i.e., each edge will be rewired with probability p. To rewire an edge $\{i, j\}$, with $j = i + 1$ or $j = i + 2$, a random node $j' \neq i$ is drawn and the edge is changed to $\{i, j'\}$. Throughout this paper, $p = 0.1$ is used.

As mentioned, we only consider *connected* networks, i.e., networks where there exists a path of edges between any two nodes. We use depth first search to verify, whether any created network is connected or not. If the network is not connected, then the whole network is discarded and the construction process repeated, until a connected network is generated.

IV. ALGORITHMS

The straightforward way to perform the outbreak simulations outlined in Sec. II, often called *simple sampling*, is to

start with the initial state and, while performing the iterations, draw the necessary random numbers independently "on demand." This will generate typical outbreaks, i.e., when performing K independent runs, one can efficiently sample events which occur with probabilities not smaller than $O(1/K)$.

We are interested in the large-deviation properties of the outbreaks, i.e., want to access events that occur with much smaller probabilities. To achieve this, we have to control the dynamics of the outbreaks. This works by biasing them in a suitable way within a Markov chain Monte Carlo (MCMC) simulation [38], as explained in Sec. IV B. But to use the outbreak simulation as the basic element within an MCMC simulation, we have to make it accessible for control, as explained in the next section.

All large-deviation simulations are for a fixed given network. As stated below, we average only over few networks, or, for a large number N of nodes, only one given network is considered due to assumed self-averaging. Thus, we are *not* interested in rare properties induced by rare network structures. This is justified, because the contact network of a population of individuals is usually given. Thus, what we are interested in are typical and rare dynamical processes taking place on typical networks.

A. Outbreak simulation

To analyze this model with large-deviation methods down to very small probabilities, we need a way to manipulate the randomness of the spread of disease in a controlled fashion. This is done by manipulating the random numbers utilized within the simulation.

An easy way to achieve this, is to draw the random numbers beforehand, store them in one or several vectors [38,49] and pick numbers from the vector whenever needed. That means, an educated guess is required about how many time steps $t_{\max}$ are needed for the simulation to make the vectors large enough. Clearly, the choice of $t_{\max}$ will depend on the values of λ and μ and will be determined below.

Now, the random numbers to be drawn beforehand are contained in two arrays $\xi_\lambda[l]$ and $\xi_\mu[l]$ with $l = 0, 1, \ldots t_{\max}N$. The entries shall be drawn uniformly between 0 and 1 each. The MCMC approach will manipulate these two vectors to control the outbreak simulation. The basic assumption used in the MCMC approach is that the state of a system, here the entire evolution of an outbreak, changes only slightly, if the random numbers are changed only slightly. For this purpose each random number will be assigned a specific purpose or use. This implies that any random number can occasionally be ignored.

Now the use of the random numbers in one outbreak simulation is detailed. Let $\tau \geqslant 0$ be the current time step. To calculate the states of the nodes for the next iteration, we first iterate over all susceptible nodes i, that have at least one infected neighbor. The probability for i to be infected is λ_i as shown in Eq. (1). To decide whether the node should be flagged for becoming infected at time $\tau + 1$ the random number $\xi_\lambda[\tau N + i]$ is used, i.e., it will be flagged to become infected if $\xi_\lambda[\tau N + i] < \lambda_i$. Of course this means that for all nodes i which have no infected neighbors, the corresponding entries of ξ_λ are ignored.

Next, we iterate over all infected nodes i. We use the random number stored at $\xi_\mu[\tau N + i]$, to flag the state of node i to be recovered in the next time step, which occurs with probability μ.

Afterward all nodes that are currently flagged to become infected are set to infected.

Note that, technically, one could store all needed random numbers from ξ_λ and ξ_μ in one single array. We found this splitting more convenient, in particular because it allows to easily manipulate the arrays in different ways.

The underlying network is not changed during the simulation, it represents a typical society. Furthermore, because the actual outbreak always starts with only node 0 being infected, no other randomness is present except the one contained in the two vectors of random numbers. Thus, the dynamic evolution and any measurable quantity are deterministic functions $f(\xi_\lambda, \xi_\mu)$.

B. Large-deviation sampling

Our goal is to calculate the probability density function (pdf) $P(E)$ for a given network $\mathcal{G}$ and given values of λ and μ. Here, E stands for a measurable quantity of the spread of disease, in our case either $E \equiv C$ or $E \equiv M$. In the following, E will be referred to as *energy*.

To calculate the pdf over a large range of the support, possibly over its full support, one usually must be able to obtain it in the region of very small probability densities as well. To achieve this within numerical simulations [34], specific large-deviation algorithms [36] can be applied. Such approaches have been used to study various equilibrium and nonequilibrium problems like alignment scores of protein sequences [37,50,51], nucleation [52], properties of random networks [53–55], dynamics of the totally asymmetric exclusion process [56,57], traffic models [58], calculation of partition functions [59], dynamics of model glasses [60], dynamics of Ising ferromagnets [38,61], statistics of negative-weight percolation [62], and RNA work processes [63].

Various large-deviation algorithms exist. Here, we applied an approach based on the Wang-Landau (WL) algorithm [39]. Although the general approach is well known, we present the main steps along with the details that are necessary to reproduce our results.

The algorithm starts with a nonnormalized estimate $g(E)$ of the density of states for the energy E. In case one does not have any prior information, like here, one starts with an unbiased estimate $g(E) = 1 \, \forall E$. The algorithm will iteratively refine $g(E)$ to converge closely to the true pdf. This is achieved by creating a Markov chain in the space of all possible outbreaks for a given network and given initial state S, I or R of each node. Since, as shown in the previous section, each outbreak has a one-to-one correspondence to the two arrays ξ_λ and ξ_μ, the Markov chain is actually performed in the space of all possible assignments of random number entries from [0,1] to these two arrays. We denote by $(\xi_\lambda^n, \xi_\mu^n)$ the current configuration at Markov step n. For each of such a configuration, a full outbreak simulation is performed and the energy, i.e., the cumulative or peak fraction of infections, is read off. Thus, as mentioned, this energy is just a deterministic function of the configuration: $E_n = E(\xi_\lambda^n, \xi_\mu^n)$.

To perform the Markov chain, we use in particular the Metropolis-Hastings MCMC method [64–66]. Therefore, each step in the Markov chain consist of generating a *trial configuration* $(\tilde{\xi}_\lambda, \tilde{\xi}_\mu)$ from the current configuration, which will be accepted or rejected, as detailed more below.

First, we explain how the trial configurations are generated here. A combination of three different possible moves is used. The moves are all based on the current configuration, i.e., they start with $(\tilde{\xi}_\lambda, \tilde{\xi}_\mu) = (\xi_\lambda^n, \xi_\mu^n)$. One of the three following change operations is randomly selected:

With a high probability of 98%, we just perform *random changes* as follows: We randomly choose one of the two arrays $\tilde{\xi} \in \{\tilde{\xi}_\lambda, \tilde{\xi}_\mu\}$, draw a random index k and a random number $\chi \in [0, 1]$ uniformly and set $\xi[k] = \chi$. This is repeated B times. As a rule of thumb, B should be chosen such that about 50% of the trial configurations are accepted, which is what we aim for. The actual numbers are stated in the results section. Note that the correctness of the method does not depend on the acceptance rate, however, it does affect efficiency. It is clear that this move alone can reach all possible configurations of (ξ_λ, ξ_μ), which means that ergodicity is fulfilled. Nevertheless, for a better convergence, we include two more moves:

With a probability of 1%, we perform a *rotation*, i.e., $\tilde{\xi}_\lambda$ and $\tilde{\xi}_\mu$ are rotated by N elements to the left or to the right, with periodic boundaries. This roughly corresponds to shifting the resulting time series of the outbreak by one time step to the left or right.

Also with a 1% probability we perform a *swap*. Here, we draw two random indices ι and ν and swap the values $\xi_\lambda[\iota] \leftrightarrow \xi_\lambda[\nu]$ and $\xi_\mu[\iota] \leftrightarrow \xi_\mu[\nu]$. This is repeated B times to create one trial configuration.

Note that these moves do not skew the probability of the resulting random-number vectors in any direction, since all entries are always uniformly drawn from [0,1].

For the trial configuration of random numbers a complete outbreak simulation has to be performed again, resulting in the corresponding energy $\tilde{E} = E(\tilde{\xi}_\lambda, \tilde{\xi}_\mu)$.

The trial configuration will now be accepted, i.e., $(\xi_\lambda^{n+1}, \xi_\mu^{n+1}) = (\tilde{\xi}_\lambda, \tilde{\xi}_\mu)$ and therefore $E_{n+1} = \tilde{E}$, with a Metropolis-Hastings probability

$$p_{\mathrm{acc}} = \min\left[1, \frac{g(E_n)}{g(\tilde{E})}\right]. \tag{4}$$

If the trial configuration is rejected, then the current configuration is kept, i.e., $(\xi_\lambda^{n+1}, \xi_\mu^{n+1}) = (\xi_\lambda^n, \xi_\mu^n)$, and therefore $E_{n+1} = E_n$.

As usual for the WL algorithm, next the density estimate g is updated using a multiplicative factor $f > 1$, i.e., $g(E_{n+1}) \to f g(E_{n+1})$, while for all other values of E, $g(E)$ remains the same. One can start with a rather large factor like $f = e \approx 2.71$. The factor is then reduced toward 1 during the simulation.

However, the saturation of the final error becomes a problem for the original WL algorithm ([67]; see also Refs. [68–70]). The algorithm introduced by Belardinelli and Pereyra [68] is used to circumvent the problem, since it was shown [70], that error saturation does not become a problem for this alternative algorithm. The main difference between

this algorithm and the original WL is in how the factor f is updated during the simulation, for details see the citations.

Still, the WL algorithm and its variants do not fulfill detailed balance. Therefore, we perform entropic sampling [40] afterward. We start with the estimate $g(E)$ as computed by WL. Entropic sampling is very similar to WL. The same method is used to generate a Markov chain and accept the states based on the probability Eq. (4). This time, however, we do not update g, but instead maintain a histogram $H(E)$ of visited states. We always employ entropic sampling for the same number of steps, as were used for the preceding Wang-Landau runs, respectively.

To finish the entropic sampling simulation, the desired pdf can be calculated. First, a nonnormalized pdf is calculated,

$$\tilde{P}(E) = g(E)H(E), \tag{5}$$

for all bins, where $H(E) > 0$. For all other bins the pdf would be unknown. Then the pdf is normalized,

$$P(E) = \frac{\tilde{P}(E)}{\int_{\hat{E}} \tilde{P}(\hat{E})d\hat{E}}. \tag{6}$$

During the simulation using the entropic sampling, we occasionally sample, i.e., store trajectories of outbreaks, which can be analyzed later on. This will lead to a rather uniform sampling of the trajectories with respect to the measured energy, C or M.

Calculating the pdf over the whole regime at once can be rather challenging. To make it more feasible, the E range is split in multiple overlapping intervals [71,72]. For each of those intervals we performed a WL and an entropic sampling simulation. Finally the resulting pdf are merged to obtain a full pdf. This can be done, because the pdfs of the overlapping regions have to match, at least within statistical fluctuations [37,39].

For some of these overlapping intervals we had problems with the ergodicity, which can be observed if not the full interval is visited, or if the distribution from neighboring intervals does not match well. To circumvent the problem, we use a replica exchange Wang-Landau (REWL) algorithm [73–75] for the affected pdfs, which works similar to the Wang-Landau algorithm described above, but regularly attempts to exchange configurations between independent simulations on different intervals, utilizing a suitable Metropolis criterion. Again, we refer to the literature for details. Note that we also applied the replica exchange approach to the entropic sampling to obtain the final pdf estimates.

V. SIMPLE SAMPLING SIMULATIONS

To choose points of interest in parameter space and a suitable length t_{max} of the outbreak simulations, we have performed some test simulations prior to the large-deviation simulations.

A. Critical transmission probability

We want to analyze the behavior of the model in the nonpandemic phase, in the pandemic phase, and close to the epidemic threshold. Since we work in discrete time, the parameters are not rates but probabilities. Thus, unlike
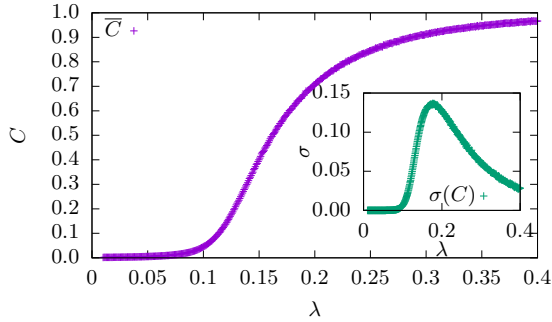
FIG. 2. Example for the average $\overline{C}$ as a function of the transmission probability λ for $N = 800$ and $\mu = 0.14$. The inset shows the variance. Error bars are smaller than symbol sizes.

FIG. 4. The duration Δt_{90} until the outbreaks are completely finished for 90% of the independent outbreak simulations as a function of λ, for different system sizes N. The largest and smallest value of N are labeled. In between the values behave monotonously. The used values are $N \in \{200, 600, 1200, 4800, 9600\}$. The dashed line indicates the value of λ_c.

to the continuous-time case, there is no natural or neutral timescale and we cannot set one of the probabilities to 1. Therefore, the recovery probability is chosen to be $\mu = 0.14$ as a working basis in all simulations. Since we are still free to choose λ, the general results should not depend much on the specific value of μ, unless it approaches 0 or 1. Thus, the task is to determine the critical transmission probability λ_c of the epidemic threshold. For this purpose, we investigate network sizes up to $N = 3200$. For each value of N, we generated 200 000 random networks and performed outbreak simulations for each network. Initially, only node 0 is infected, while all other nodes are susceptible. Here, all outbreaks are iterated until no infected node remains. We measure the average cumulative fraction $\overline{C}(\lambda)$ of infected nodes and also calculated the variance $\sigma(C)$ for each combination (N, λ). Errors are estimated with bootstrap resampling [76]. For a different number N of nodes the curves $C(\lambda)$ change so little, that it would be hardly visible. We therefore only show an example of this in Fig. 2.

We define the finite-size critical transmission $\lambda_c(N)$ as the peak of the variance $\sigma(C, N)$. To measure the peak of the critical transmission, we fit Gaussian-shaped functions around the maxima, respectively. We then apply standard finite-size scaling to calculate the critical transmission rate $\lambda_c(\infty)$ by fitting

$$\lambda_c(N) = \lambda_c(\infty) + aN^{-b} \quad (7)$$
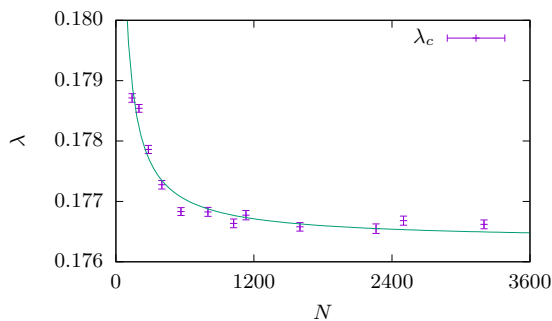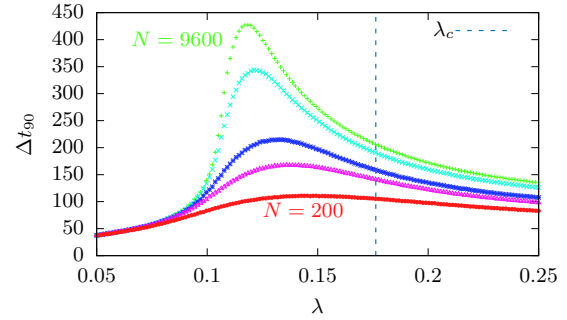
to the data, as shown in Fig. 3.

We obtain a value $\lambda_c(\infty) = 0.1763(2)$ for the critical transmission. The other fit parameter were $a = 0.24(24)$ and $b = 0.91(21)$, with rather large error bars, but these values are not of interest here. Note that fitting $\lambda_c(N) = \lambda_c(\infty) + a_2 \log(N)^{b_2}$ looks very similar and leads to a similar critical transmission $\lambda_c(\infty) = 0.1750(4)$.

B. Disease duration

For the large-deviation simulation, we cannot simply run each outbreak simulation until the disease dies out, because the MCMC scheme operates with a vector of random numbers which must be of fixed length. Thus, we have to find a suitable timescale for the duration of the outbreak simulations.

For this purpose, we performed simulations in the same manner as described in the previous section and measured the duration Δt it takes until no infected nodes remain, i.e., $i(\Delta t) = 0$, for each simulation. After this time, the state of the nodes will not change because the outbreak dynamics are finished. For each parameter set (N, λ), we measured 100 000 randomly generated networks.

For each considered parameter set (N, λ) the characteristic time Δt_p, which describes how long it takes until p% of the outbreak dynamics are finished, is calculated. As an example we show the measured curves for Δt_{90} in Fig. 4. The curves show the typical signs of a dynamical phase transition accompanied by critical slowing down [77]. Critical slowing down plays an important role for *early warning signals* of infectious disease transitions [78]. Interestingly, the outbreak takes longest well below the epidemic threshold λ_c, which fits with previous observations [79].

To investigate the worst-case scenario, we look at

$$\Delta t_p^{\max}(N) := \max_{\lambda}[\Delta t_p(N, \lambda)]. \quad (8)$$

The result for $\Delta t_{90}^{\max}(N)$ can be found in Fig. 5. We used this result to set up the length of the outbreak simulations within the large-deviation approach, see below. But beyond this technical aspect, it is also interesting to investigate the scaling behavior. Critical slowing down often leads to a power-law behavior of the correlation length [80,81]. Since $\Delta t_{90}^{\max}(N)$ can
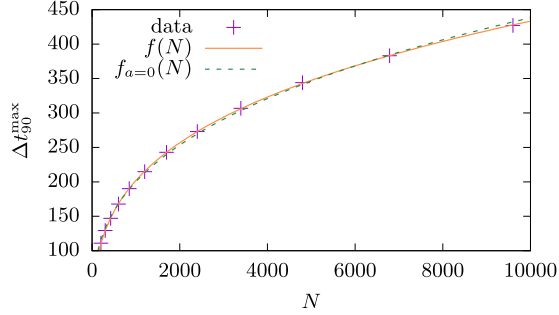

FIG. 3. Critical transmission λ_c for $\mu = 0.14$ as a function of the number N of nodes with fit to $\lambda_c(N)$ from Eq. (7).

FIG. 5. Maximum $\Delta t_{90}^{\max}$ over all values of λ of how long it takes until 90% of outbreaks are finished, as a function of N. The continuous line shows the fit to $f(N)$, while the dashed line shows the fit to $f_{a=0}(N)$.

be understood as correlation length we fitted a power law,

$$f(N) = a + bN^c, \tag{9}$$

to the $\Delta t_{90}^{\max}(N)$ data. This seems to describe the relation very well. The fit parameter are $a = -50(4)$, $b = 36(2)$, and $c = 0.2824(43)$. That means that the time it takes until 90% of the outbreaks are over scales roughly with the fourth root of the system size.

The fit still works well with a fixed $a = 0$, i.e., for

$$f_{a=0}(N) = bN^c, \tag{10}$$

which leads to $b = 19.1(6)$ and $c = 0.340(4)$. Also note that both fits work exceptionally well for all Δt_p except for p very close to $p = 100\%$ or $p = 0\%$. It is even possible to find simple functions for the fitting parameter, i.e., $a(p)$, etc. The functions which are obtained by using these functional parameters, e.g., $f(N|p) = a(p) + b(p)N^{c(p)}$, also fit the data reasonably well.

Still, as this is only preparation for our large deviation sampling, we do not pursue this any further.

It is also interesting to look at the duration right at the critical transmission. This can be found in Fig. 6. Here we set $a = 0$ for the fit because the errors become unreasonably large otherwise. The obtained parameters from the fit are $b = 51(3)$ and $c = 0.1555(8)$, which means that here the duration scales roughly only with the 7th root of the system size.
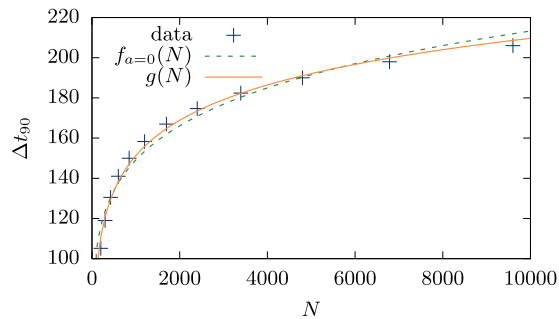

FIG. 6. The time Δt_{90} how long it takes until 90% of outbreaks dynamics are finished as a function of the network size N for the critical value $\lambda = 0.1763$. The lines show the results of fits to $f_{a=0}(N)$ and $g(N)$.

TABLE I. Parameters for the simulations: transmission probability λ, number of nodes N, the approach used, the number $\#N$ of independent network realizations, the number $\#I$ of intervals used in the WL or REWL sampling, and the number B of exchanges performed per MCMC attempt for the arrays ξ_λ and ξ_μ of random numbers.

λ	N	Approach	$\#N$	$\#I$	B
0.1763	200	WL	15	24	75
0.1763	400	WL	7	24	150
0.1763	800	WL	4	24	166
0.1763	1600	WL	1	24	900
0.1763	3200	WL	1	24	2048
0.1763	6400	WL	1	48	3072
0.1	3200	WL	1	24	1024
0.4	3200	REWL	1	27	256

Note that usually a power law with a small exponent cannot be well distinguished from a logarithmic behavior. Therefore, we also fitted a logarithmic function,

$$g(N) = \alpha \log(N\beta). \tag{11}$$

The quality of this fit is not good for the $\Delta t_{90}^{\max}$ data, but works well for specific values of λ, e.g., $\lambda = \lambda_c$ and is therefore included in Fig. 6. In the latter case the obtained fit parameters are $\alpha = 25.3(6)$ and $\beta = 0.39(6)$.

VI. CUMULATIVE FRACTION C OF INFECTIONS

Using the large-deviation approach, we now present the result for the distribution of the fraction C of cumulative infected nodes. Note that the study of the large deviation refers to the dynamics on a given network. Since in a real-world situation the contact network is given, we do not study rare-events with respect to rare networks here. First, we present results for the pdf of C. They are obtained by using Wang-Landau, plus afterward refining the result with entropic sampling. If necessary, i.e., in case we observed nonconvergence, then we applied REWL instead.

The parameters we use for the simulations are presented in Table I, for the different networks sizes N and values of the transmission probability λ. At the critical point $\lambda = \lambda_c$, we study also finite-size scaling by considering different network sizes. For the representative values smaller and larger than the critical transition, $\lambda = 0.1$ and $\lambda = 0.4$, respectively, we perform simulations only for a rather large system size of $N = 3200$.

Note that we use a recovery probability of $\mu = 0.14$ everywhere, which, as a side note, is in the range of recovery probabilities used to model the current corona virus pandemic [6,12], although there is a wide variablilty of models leading to other parameter values [82–85].

For small sizes, we perform the full large-deviation sampling for a small number $\#N$ of few independently drawn networks, while for the largest sizes, where we assume some kind of self-averaging, we study only $\#N = 1$ generated network. The latter case also corresponds somehow to the real-life simulation, where only one contact network is given, but the dynamics evolves randomly.
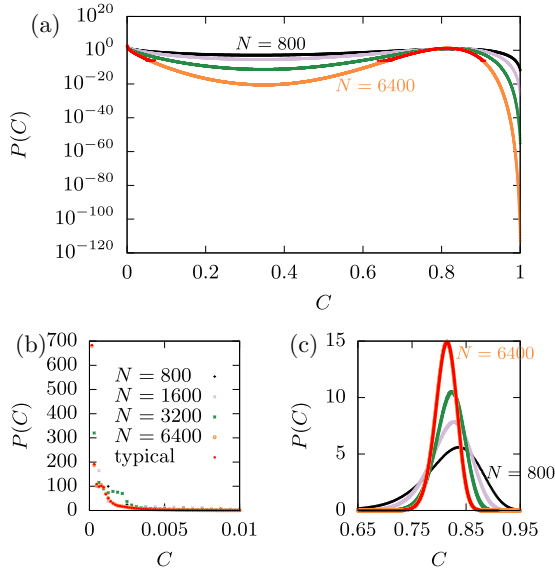
FIG. 7. Probability density of total infections C for $\mu = 0.14$ and $\lambda = 0.1763$. Panel (a) shows the full distribution in logarithmic scale, while panels (b) and (c) highlight the two peak regions in linear scale. For panels (a) and (c) the largest and smallest values of N are labeled, in between the values behave monotonically, except for a small area around $C = 0.8$, where the order is reversed, see panel (c). The used values were $N \in \{800, 1600, 3200, 6400\}$. We also included the typical-event sampling results for $N = 6400$ in red.

For the large-deviation simulation, we need to choose a length of the vector of random numbers, which determines the maximum time duration $t_{\max}$ of an outbreak that can be covered. In theory, arbitrary long outbreaks are possible, so one has to choose a cutoff time anyway. We have chosen as maximum outbreak time of $t_{\max} = 3\Delta t_{90}^{\max}(N)$, the latter one as determined in Sec. V B. To verify whether this is long enough, we keep track, during the large-deviation sampling, of how often the outbreak was *un*finished after the given time. When considering all the different network sizes, transmission probabilities and intervals of $E \equiv C$, the highest, i.e., worst-case frequency $f_{\neq}$ of observing a nonfinished outbreak occurred for $N = 3200$ in the interval of $E \in [0.47, 0.56]$ with $f_{\neq} \approx 1.4 \times 10^{-5}$. Typically, the frequency was much lower, e.g., the worst interval for $N = 6400$ exhibited $f_{\neq} \approx 1.7 \times 10^{-8}$. Since an unfinished outbreak constitutes only a few infected nodes anyway, this shows that to observe extremely unlikely events in terms of C, and clearly M anyway, one does not have to cover extreme unlikely long durations of outbreaks and the choice of $t_{\max}$ is sufficient.

In Fig. 7 the probability density $P(C)$ is plotted for different system sizes N. Note that here and in the following the pdfs $P(E)$, where E can be either of C and M, are always normalized such that $\int_E P(E) = 1$. Note also that we sample the histograms with the highest possible resolution of one bin per possible value of C. Whenever we average over different networks, we calculate the pdfs for each of them and then merge them by averaging the logarithmic probabilities and normalizing again.

We are able to measure the probability density over the whole range of its support, extending over up to 115 decades
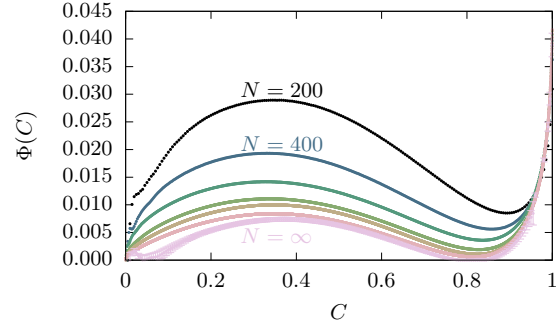

FIG. 8. Rate function Φ as a function of total infections C for different values of N. The smallest two and the biggest value of N are labeled. In between, the values behave monotonically. The values $N \in \{200, 400, 800, 1600, 3200, 6400, \infty\}$ are used. $N = \infty$ is gained by extrapolation and all other values are measured.

in probability. To put that into perspective, for $N = 6400$ we calculated C about 3.1×10^9 times, once per MCMC attempt, during entropic sampling and WL combined. That means, if we use typical-event sampling to create a histogram and estimate the probability density function with the same numerical effort, we are only able to resolve probabilities with a resolution of about 10^{-9}. This is also shown in Fig. 7, where we also see that the typical-event sampling results and the large deviation results agree very well. To resolve the whole density function with typical-event sampling, one would need about 3×10^{105} times as much computational power as used here. The computational advantage would grow even higher for larger values of N.

Having the whole probability density function is interesting from an insurance perspective. Let us assume there is a cost function $\mathrm{Cost}(C)$. Using the pdf we measured one can easily calculate the exact expected value of the cost function. Even the very improbable cases will be relevant here because they will likely be associated with very high costs, e.g., the financial loss due to a pandemic.

As visible by the two peaks in the pdfs, the disease either dies out very quickly, corresponding to the peak near $C = 1/N \approx 0$, or about 80% will contract the disease over the evolution of the outbreak. Intuitively this makes sense, as only one node is infected in the beginning, thus the disease dies out if that node recovers before infecting anyone. If the disease does not die out quickly, however, then it will persist until a good fraction of the network is immune. The observed behavior becomes more pronounced for larger networks, as visible by a decrease of $P(C)$ for intermediate values of C.

To relate to mathematical large-deviation theory, we also study the empirical rate functions, defined as

$$\Phi(C, N) := -\frac{\ln(P(C))}{N} + \Phi_0, \qquad (12)$$

where $\Phi_0 = \mathrm{const} = \min_C \left(-\frac{\ln(P(C))}{N} \right)$, such that the minimum of the rate function occurs at $\Phi = 0$. The calculated rate functions are displayed in Fig. 8.

Since an apparent convergence is visible when increasing network size N, we also estimate the rate function for $N = \infty$ with finite-size scaling, in a similar way to the disease duration
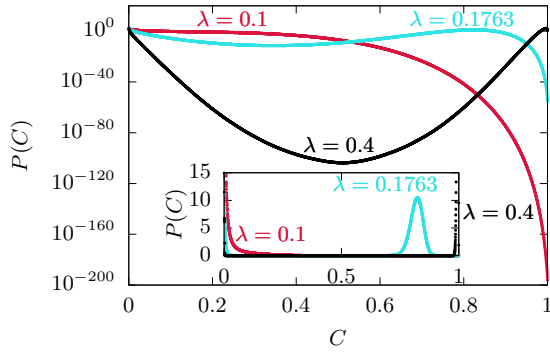
FIG. 9. Probability density of total infections C for $\mu = 0.14$ and $N = 3200$ for different values of λ, below, at, and above the critical threshold. Linear scale in inset.

in Sec. V B. For this purpose we fit the function

$$k(N) := \Phi_\infty + \eta N^{-\kappa} \tag{13}$$

for each value of C.

The errors of the fit are used as error bars and the result is included in Fig. 8. The results show that the numerically obtained rate function seems to converge well. This means that the mathematical *large-deviation principle holds*, i.e., the size dependence on N is in leading order given by $P(C) \sim \exp[-N\Phi(C) + o(N)]$. As a consequence of this "well behaving," analytical progress regarding $P(C)$ might be feasible, e.g., through application of the Gärtner-Ellis theorem [41–44].

So far, we have considered only the critical point $\lambda \approx \lambda_c$. A comparison with the other values of the transmission probability is shown in Fig. 9. As one would expect, higher values of λ lead to an increased probability for larger values of C. Correspondingly, lower values of C become far more likely for the lower transmission probabilities. Note that for all three considered values of λ, we observe rather high probabilities for $C \approx 0$, since only one node is infected in the beginning. But this is only an effect emerging from the initial condition and can be ignored when discussing the main part of $P(C)$. Even if one started with a larger number of initially infected nodes, this will only affect the height of the peak for small values of C and the overall weight of the part for $C \gg 0$, but not the shape.

A. Correlations

We want to further analyze the properties of typical and atypical outbreaks to obtain insight into their structure and maybe even identify possible causes for extreme events. For this purpose, we store during the entropic sampling for each WL interval 200 000 time evolutions of outbreaks, at steps evenly spaced out in the entropic sampling Monte Carlo time. We store for these time evolutions the fraction i of infected, the fraction s of susceptible and the fraction c of the total fraction of so-far infected nodes during the time evolution, see Sec. II. Below, T will denote either of these quantities, and we call

$$T = (T[0], \dots, T[t_{\max} - 1]) \tag{14}$$
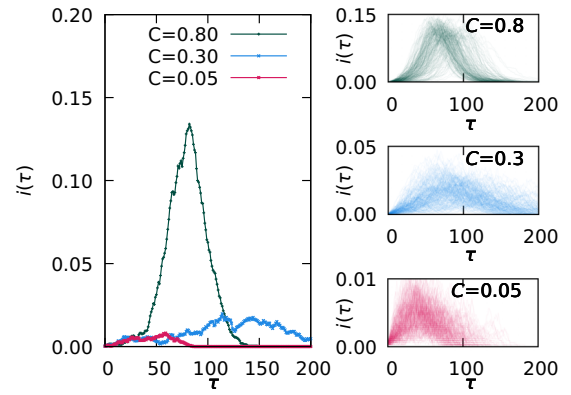
a *time series*.

FIG. 10. Fraction of infected nodes i as function of time τ for three different values of C. The plot on the left shows a single time series for each C value as examples, whereas the plots on the right each show 250 time series for their respective C value.

Each time series is binned according to its energy E, i.e., C here, or M in Sec. VII. We denote by b_E the set of a number B_E of time series collected for histogram bin E, i.e.,

$$b_E = \left\{ T_0^E, \dots, T_{B_E-1}^E \right\}. \tag{15}$$

As an example, in Fig. 10 we show a collection of time series $i(\tau)$ for three different values of the cumulative fraction C of infected individuals, for $N = 3200$ and $\lambda = \lambda_c$. One can observe that the infection can last way longer for medium values of C and dies out quicker for very large or low values of C.

In the next subsection, we use heat maps to investigate how similar the time series are when comparing them pairwise. Afterward, we calculate other measurable quantities of the time series to relate them to the values of C they exhibit, respectively.

1. Disparity heat maps

To measure how similar the time series are to each other, we first normalize each series by dividing through its maximal value encountered during the outbreak. This way we can better compare the shape of the time series and are not comparing their magnitudes.

We define a distance d for two normalized time series T, T' as

$$d(T, T') := t_{\max}^{-1} \sum_{\tau=0}^{t_{\max}-1} |T[\tau] - T'[\tau]|. \tag{16}$$

We define the *disparity* $V_T(E, E')$ between time series from E and E' as the averaged distance $d(T_\alpha^E, T_\beta^{E'})$ for pairs of time series taken from the bins E and E', respectively. Here, we used 500 time series per bin, drawn randomly from all saved time series that were collected for the respective bin. Hence a total of up to 1 600 000 time series are used in each analysis.

In Fig. 11 we show the disparity V_i, for the fraction of infected, for $\lambda \approx \lambda_c$ color-coded, i.e., in form of a heat map. Note that we are able to show the disparities over the full range of possible values for C, which is only possible because we applied the large-deviation approach. When using simple
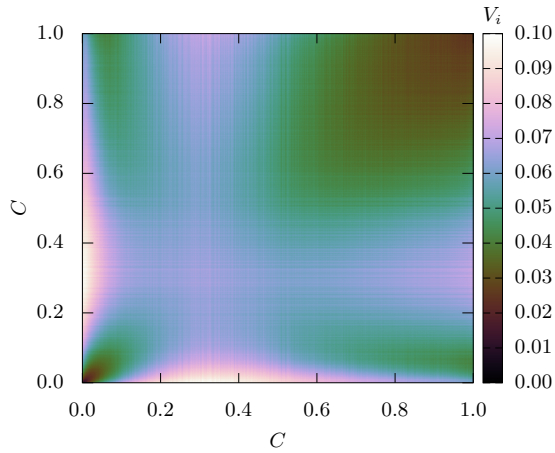
FIG. 11. Disparity V_i of the time series $i(\tau)$ of the fraction of infected individuals for pairs of time series binned with respect to their total fraction C of infections, for $N = 3200$, $\mu = 0.14$, and $\lambda = 0.1763$.

FIG. 12. Disparity V_c of the time series $c(\tau)$ of the fraction of cumulative infections for pairs of time series binned with respect to their total fraction C of infections for $N = 3200$, $\mu = 0.14$, and $\lambda = 0.1763$.

sampling instead, only a very small range of values near $C \approx 0$ and near $C \approx 0.8$ would be accessible. Here we are able instead to identify three different regions. The first is located in the range $0 \leqslant C \leqslant 0.1$, the second is $0.1 < C \leqslant 0.5$, and the third for $C > 0.5$. Region one and three seem to consist of time series where the shapes are similar to each other, within the region, visible by the dark color around the diagonal in the heat map. But they are quite different to other regions. Region two seems to consist of different time series that are not even that similar to time series from its own region, i.e., here we observe strong fluctuations from time series to time series. Note that in Fig. 10 the time series are selected from these regions, and thus illustrate their behaviors. Fig. 10 tells us that time series from the first region exhibit only small fractions of infected and the outbreak dies out quickly. For the second region, we observe medium strong outbreaks, but they may take very long and the shapes and durations fluctuate strongly. In the third region, many individuals get infected during the outbreak, leading to an even larger fraction of individuals infected at the same time, but the outbreak finished more quickly than in region two.

For $\lambda = 0.1$ the heatmap looks similar (not shown), though now region 2 is shifted toward larger values of C, i.e., the region is $0.15 < C \leqslant 0.8$, and the other regions change accordingly.

For $\lambda = 0.4$ the heatmap also looks similar (not shown). Here the second region is much smaller, i.e., $0.075 < C \leqslant 0.21$, and the other regions change accordingly.

In Fig. 12 we show the disparity heat map for the time series of cumulative infections, i.e., V_c. This heat map adumbrates the three regions as well, though they are much less pronounced. Thus, to compare the dynamics of infections, the current fraction of infections allows for a better insight compared to the cumulative fraction of infections. For $\lambda = 0.4$ we see the same, though the region borders changed, as mentioned previously. For $\lambda = 0.1$ the heat map also hints at the regions previously mentioned; however, region 2 is very faint and barely visible.
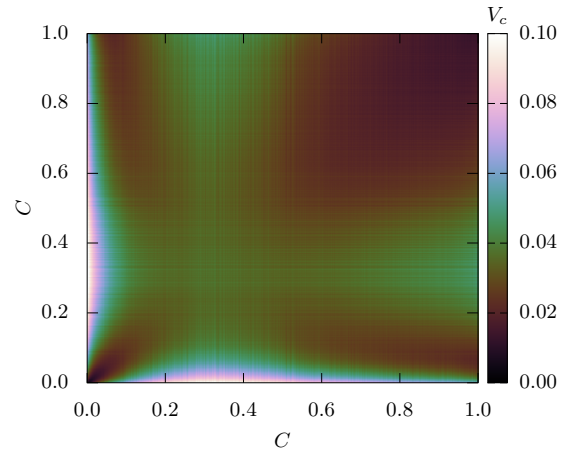
2. Conditional density

To study the relation of other measurable properties Q of the time series to their values of C, we study conditional densities. Again we bin each time series according to its energy E, here $E = C$ [see Eq. (15)] and then obtain a normalized histogram $\rho_T(Q|E)$ of Q given E. Again T will be either i, s, c, or omitted, if suitable.

For the measurable quantities Q we considered

(i) M as defined above, i.e., the maximum of the fraction i of currently infected nodes during an outbreak;

(ii) the time steps $\tau_{\max}$ until the maximal value of the fraction i of infected is reached. This measures the timescale it takes for an outbreak to reach its maximum activity. This is interesting for practical purposes, as it translates to the time where the maximal healthcare capacity is required;

(iii) the time steps $\tau_{\min}$ until the minimal value of the fraction s of susceptible nodes is reached. This means that, after this time, no additional nodes obtained an infection, although the recovery of the remaining infected nodes still takes some time. This quantity is a measure for the outbreak duration;

(iv) the number of time steps τ_{10}^{90} it took such that the fractions i or c here, raised from 10% to 90% of its maximal value, respectively. In the few cases where, for analyzing i, this occurred several times, we only consider the duration of the first occurrence. These timescales quantify how long the outbreak is very active;

(v) the fraction f_{sw} of how many neighbors of an infected node are infected through a long-range edge, i.e., along those edges which were rewired during graph generation. Hence,

$$f_{\text{sw}} = \frac{1}{N} \sum_{i, I_i > 0} \frac{L_i}{I_i}, \qquad (17)$$

where I_i is the number of neighbors of node i infected by node i, and L_i is the number of neighbors of node i infected through long-ranging edges;

In Fig. 13 we show the distribution $p_s(\tau_{\min}|E)$ for the time steps $\tau_{\min}$ it took until the minimum of susceptible nodes was reached, conditioned to the value of C. As one would expect,
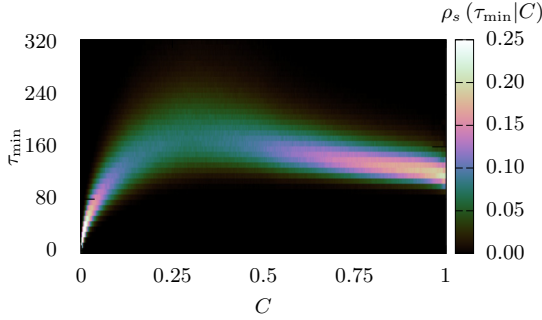
FIG. 13. Conditional density $\rho_s(\tau_{\min}|C)$, which shows the probability of $\tau_{\min}$, i.e., how many time steps it takes until the last node got infected during an outbreak, for any given C. The system size is $N = 3200$, the recover probability $\mu = 0.14$ and the transmission probability $\lambda = 0.1763$.
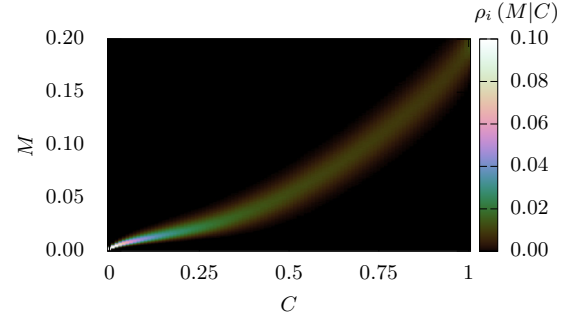

FIG. 15. Conditional density $\rho_i(M|C)$, which shows the probability of M, i.e., the maximum of the time series $i(\tau)$, for any given value of C. The system size is $N = 3200$, the recover probability $\mu = 0.14$ and the transmission probability $\lambda = 0.1763$.

this distribution is centered at small times for low values of C. For $0.25 \leqslant C \leqslant 0.5$ the disease survives the longest and exhibits the largest spread in timescales. When increasing C further, the life time of the disease decreases again. These results support the insight gained for the different regions from looking at the sample time series in Fig. 10.

For $\lambda = 0.1$ (not shown) the shape is similar, though flatter and the range where the disease survives the longest stretches now from $0.25 \leqslant C \leqslant 0.75$. Also the disease generally survives longer, because the pool of susceptible nodes decreases more slowly. For $\lambda = 0.4$ (not shown) the shape looks even more similar to the one shown in Fig. 13, though the maximum is now more pronounced and around $C \approx 0.13$ and the disease dies out even faster, because it rushes more quickly through the population.

In Fig. 14 the conditional distribution $p_i(\tau_{\max}|E)$ for the time of the peak infection is shown. As one can see, the shape is similar to the one from Fig. 13. From this similarity one can conclude $\tau_{\min} \sim \tau_{\max}$, i.e., the longer the disease lasts, the later the peak of the fraction of infected occurs. We find the same results for $\lambda = 0.1$ and $\lambda = 0.4$ (not shown).

Looking back at the diagonals of Figs. 11 and 12 we notice that we observe the largest value of $\tau_{\max}$ as well as the largest variance of $\tau_{\max}$ for the range where the *disparity* V was the

highest, i.e., where the respective time series exhibited the most variation. This makes sense intuitively and reminds one of the behavior of systems near critical slowing down.

In Fig. 15 the conditional distribution $p_i(M|C)$ for the maximum of nodes, which are in the infected state at the same time, is shown. Here, the three regions, which where visible in Figs. 13 and 14 are not apparent. Instead one observes a generally monotonous relation between C and the center of the distribution of M. Still, this is compatible with the outbreak examples shown in Fig. 10. The same behavior can be observed for $\lambda = 0.1$ (not shown), though the peak value M is generally lower and for $\lambda = 0.4$ (not shown) the peak value is higher.

In Fig. 16 we show the conditional density $\rho_i(\tau_{10}^{90}|C)$ for the duration of the most-active phase of the outbreak. As one can see, the spread of the duration times is the largest for region two, where the time series also looked more chaotic (see Fig. 10 for $T = i$). If we look at the τ_{10}^{90} for $T = c$ (not shown), the plot looks quite similar, though the durations are a bit longer overall. For $\lambda = 0.1$ the heat map looks similar, but more like a half moon, and the values scatter most around $C \approx 0.5$. Also the values are a bit higher overall. For $\lambda = 0.4$ the basic shape also looks similar, but the slope is steeper at the beginning. It also flattens out sharply at about $C \approx 0.5$,
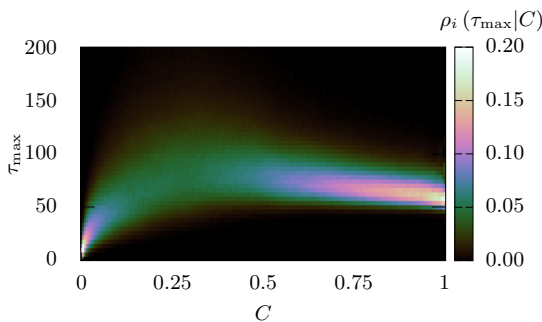

FIG. 14. Conditional density $\rho_i(\tau_{\max}|C)$, which shows the probability of $\tau_{\max}$, i.e., how many time steps it takes to reach the peak of the infection time series $i(\tau)$, for any given value of C. The system size is $N = 3200$, the recover probability $\mu = 0.14$ and the transmission probability $\lambda = 0.1763$.
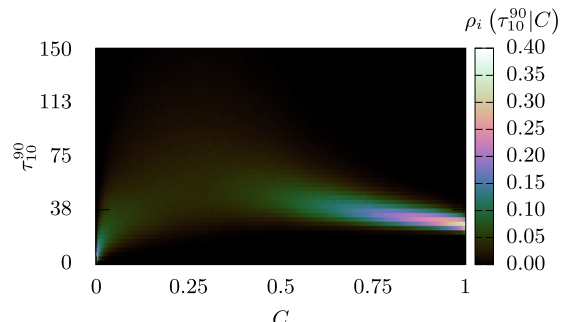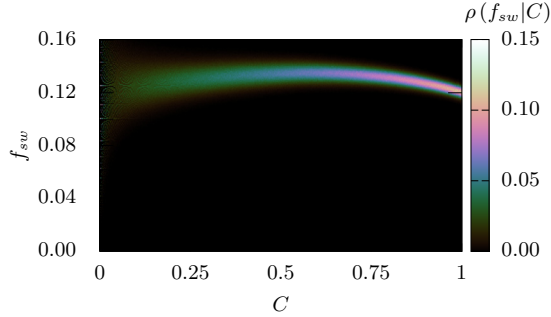

FIG. 16. Conditional density $\rho_i(\tau_{10}^{90}|C)$, which shows the probability of τ_{10}^{90}, i.e., the duration between reaching 10% and 90% of the maximum of $i(\tau)$, for any given value of C. The system size is $N = 3200$, the recover probability $\mu = 0.14$ and the transmission probability $\lambda = 0.1763$.

FIG. 17. Conditional density $\rho(f_{sw}|C)$, which shows the probability of f_{sw} for any given C. The system size is $N = 3200$, the recover probability $\mu = 0.14$, and the transmission probability $\lambda = 0.1763$.
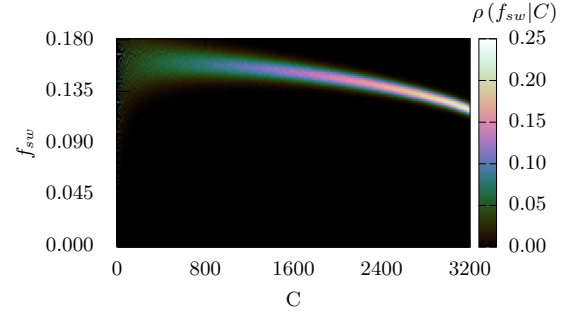

FIG. 18. Conditional density $\rho(f_{sw}|C)$, which shows the probability of f_{sw} for any given C. The system size is $N = 3200$, the recover probability $\mu = 0.14$, and the transmission probability $\lambda = 0.1$.

i.e., at a relatively small value. The values are lower overall compared to the other two values of λ.

We are also interested in the effect of the long-ranging connections. For this we measure the fraction of long-ranging edges that caused an infection.

In Fig. 17, we show the conditional distribution $\rho(f_{sw}|C)$ of the fraction of infections through long-range edges, i.e., those edges which are responsible for the small-world behavior. For small values of C, the values of f_{sw} scatter strongly, because here the disease dies out very quickly and thus f_{sw} is obtained by averaging over only very few contagions. Overall we see a weak correlation of C and f_{sw}, where f_{sw} increases slightly for larger C until $C \approx 0.75$. Thus, we see a weak effect that to see a global pandemic, the spread has to go to a slightly larger extend through long-range connections and spreads a little bit less locally. This, even within such a simple model, supports the often used real-word strategy to suppress with higher priority long-distance traveling. Interestingly, because this correlation is seen also for rather large values of C, this would also help at least a bit even if a pandemic has broken out already, not only in an early stage to prevent pandemic outbreaks. However, to actually reach all individuals, i.e., for even larger values of C, many local edges have to be involved since there are still many nodes which only have local neighbors. This explains the decrease visible for large values of C. We first suspected that the overall weak correlation was due to our high probability $p = 0.1$ of rewiring, because of which one third of the nodes are adjacent to at least one long-range edge.

The plots for $\lambda = 0.1$ in Fig. 18 and $\lambda = 0.4$ in Fig. 19, however, paint a different picture. For the transmission probability ($\lambda = 0.1$) below the critical value, we observe an anti correlation between f_{sw} and C, whereas we see a correlation for the transmission probability ($\lambda = 0.4$) above the critical value. The latter is what we expected, since the transmission via long-ranging edges infects more distant nodes, which can start new infection clusters.

VII. MAXIMUM FRACTION M OF CURRENTLY INFECTED

Next, we study the large-deviation properties with respect to the maximum fraction M of simultaneously infected nodes,

in a similar way as we have done for C. Although we have seen a strong relationship between C and M in Fig. 15, we will show below that not all results obtained for $E = C$ transfer directly to the case $E = M$. Note that for obtaining these results we had to perform completely independent large-scale simulations with energy $E = M$ to access also the outbreaks which have a rare behavior with respect to M. For our largest system size we evaluated M about 1.5×10^{10} times during entropic sampling and Wang Landau combined, which is thus the total number of local MC attempts.

The parameters we use for the simulations are presented in Table II, for the different networks sizes N and values of the transmission probability λ.

For the noncritical values of λ, namely, $\lambda = 0.1$ and $\lambda = 0.4$, we perform simulations only for a system size of $N = 1600$. We originally wanted to use a system size of $N = 3200$ here, though we encountered problems during the sampling of $\lambda = 0.4$. Note that we sample $N = 6400$ ($\lambda = \lambda_c$) with a combination of WL and REWL.

In Fig. 20 the probability density $P(M)$ is shown for different system sizes N and $\lambda = \lambda_c$. Here, we never encountered an outbreak that lasted longer than our simulation time t_{max}. This means, we have chosen t_{max} large enough to understand the behavior of M over the full range of possible values sufficiently. Using the large deviation approach we are able to measure probabilities ranging over 2500 decades. Thus, if
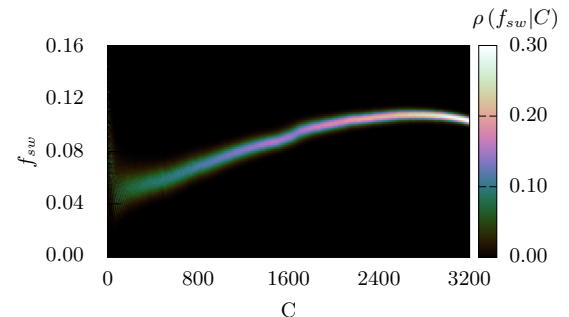

FIG. 19. Conditional density $\rho(f_{sw}|C)$, which shows the probability of f_{sw} for any given C. The system size is $N = 3200$, the recover probability $\mu = 0.14$, and the transmission probability $\lambda = 0.4$.

TABLE II. Parameters for the simulations: transmission probability λ, number of nodes N, the approach used, the number #I of intervals used in the WL or REWL sampling, and the number B of exchanges performed per MCMC attempt for the arrays ξ_λ and ξ_μ of random numbers.

λ	N	Approach	#I	B
0.1763	400	WL	8	66
0.1763	800	WL	20	166
0.1763	1600	REWL	24	512
0.1763	2400	WL	24	1024
0.1763	3200	WL	24	2048
0.1763	6400	WL	42	3072
0.1763	6400	REWL	6	4096
0.1	1600	REWL	24	512
0.4	1600	REWL	24	512

FIG. 21. Probability density of the maximum fraction M of simultaneously infected for $\mu = 0.14$ and $N = 1600$ and different λ. Linear scale in inset.

one were to sample the same distribution using only a typical-event sampling approach, one would need at least 10^{2490} times as much computational power as we used, which is clearly infeasible.

We also include the typical-event sampling results for $N = 6400$ where we use 10^{10} samples. Looking at the logarithmic probabilities the large deviation results and typical-event sampling agree very well. However, at the peaks in the linear scale, seem to be discrepancies. These come
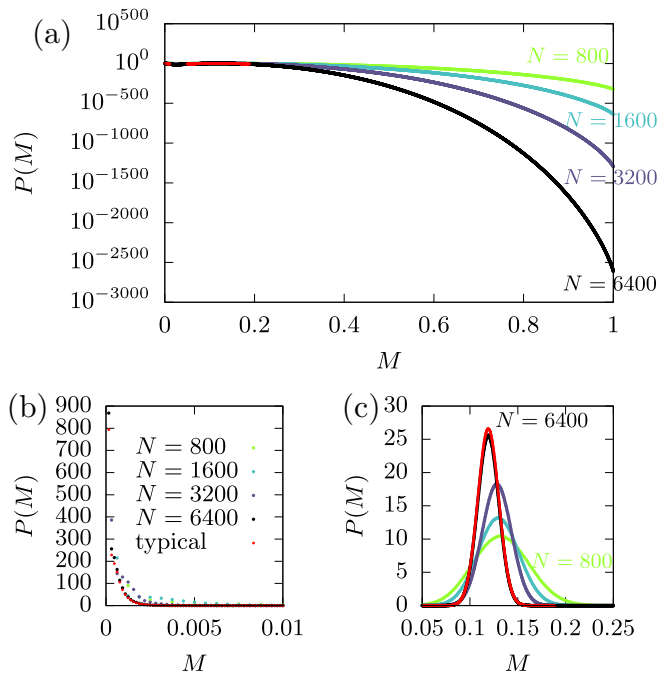

FIG. 20. Probability density of maximum fraction M of simultaneously infected for $\mu = 0.14$ and $\lambda = 0.1763$. Panel (a) shows the full distribution in logarithmic scale, while panels (b) and (c) highlight the two peak regions in linear scale. For panels (a) and (c) the largest and smallest values of N are labeled, in between the values behave monotonically, except for a small area around $M = 0.125$, where the order is reversed, see panel (c). The used values were $N \in \{800, 1600, 3200, 6400\}$. We also included the typical-event sampling results for $N = 6400$ in red.
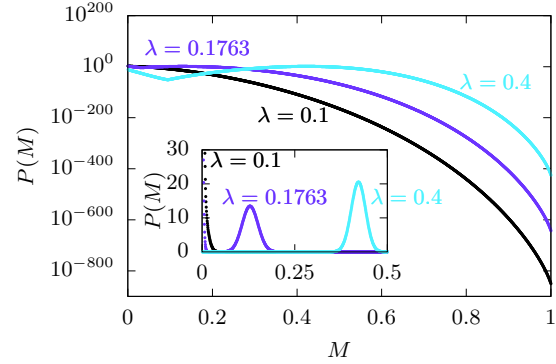
from errors that accumulate during the gluing of the large deviation intervals, which come inevitably into play because of the sampling fluctuations. But the deviations are small: the first peak would align perfectly if we add 0.04 to our typical-event sampling data in the logarithmic range, while the second peak would align perfectly if we subtract 0.015 from the typical-event sampling data in the logarithmic range. Looking at the number of magnitudes this measurements covers in the probability density range, these errors are really rather small, though they become noticeable in the linear range.

The probability density function exhibits a peak at $M = \frac{1}{N} \approx 0$ for the same reasons as there is a peak at $C \approx 0$ in $P(C)$. We also observe a peak at about $M = 0.125$. This means, if the infection survives the first few steps one can assume that typically about 12.5% of the network will be infected at the same time at some point of the outbreak. This determines the capacity of the health care system required to cope with typical outbreaks. If one wants to be prepared for large atypical outbreaks, then the tails become important. Nevertheless, for substantially larger values of M the probability density becomes very small. Unsurprisingly, the least likely case is that the entire network is infected at the same time at some point of the outbreak. Clearly, these extreme tails of the distribution are not relevant for practical applications, but from a fundamental and scientific viewpoint, it is pleasing to be able to calculate the distribution over its full support.

In Fig. 21 the pdf is shown for different transmission probabilities λ. Note the kink for $\lambda = 0.4$. We are only able to resolve this interesting point where the distribution seems to be not differentiable by using the replica exchange Wang-Landau algorithm here. Unsurprisingly, larger transmission probabilities λ lead to an increase of the probability to observe larger values of M.

In Fig. 22 we show the rate function as measured for different system sizes. Clearly, the rate functions all agree very well. Basically, no finite-size effects are visible in contrast to the case of $P(C)$. This means one can use the rate function to predict the pdf for any system size N. We verified that by using the rate function calculated for $N = 3200$ to accurately predict the pdf for $N = 400$ and $N = 2400$. Thus, it is not necessary to perform any extrapolation of the rate function. This means that for $P(M)$ our numerical results also indicate
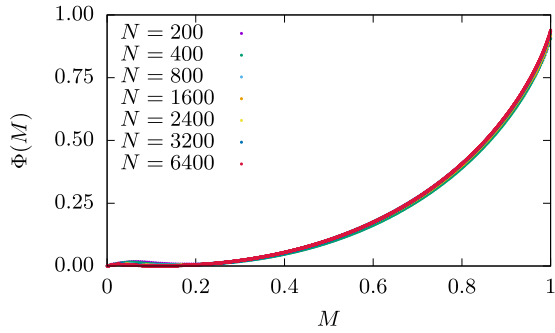
FIG. 22. Empirical rate function $\Phi(M)$ for multiple system sizes N and $\mu = 0.14$ and $\lambda = 0.1763$.

that the large-deviation principle is fulfilled. Nevertheless, the kink visible for $\lambda = 0.4$ hints that in the pandemic phase for larger values of λ, the mathematical properties of the rate function could pose some problems to an analytic treatment.

A. Correlations

1. Disparity heat maps

Again we study the disparity of outbreaks, now for pairs of times series of outbreaks classified according their values of M, respectively. In Fig. 23 the disparity V_i for the time series of the fraction of infected is shown, see Sec. VI A 1. As in Fig. 11, we again see the three regions, first for very small values $0 \leqslant M < 0.035$, the second for $0.035 \leqslant M < 0.08$ and the third for $M \geqslant 0.08$. Hence, in contrast to the case of the cumulative fraction C of infections, two of the three regions are visible on a much smaller range of values. For $\lambda = 0.1$ the heat map (not shown) looks quite similar, though the second region is shifted toward even lower M. Note that for $\lambda = 0.4$ we are only able to ensure convergence of the distribution for $N = 1600$ and therefore are comparing different system sizes. Still, the differences of the disparity plots between such rather large system sizes are very small, as we observe in general during
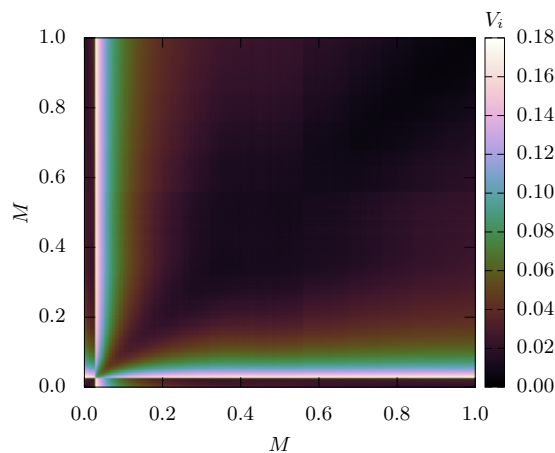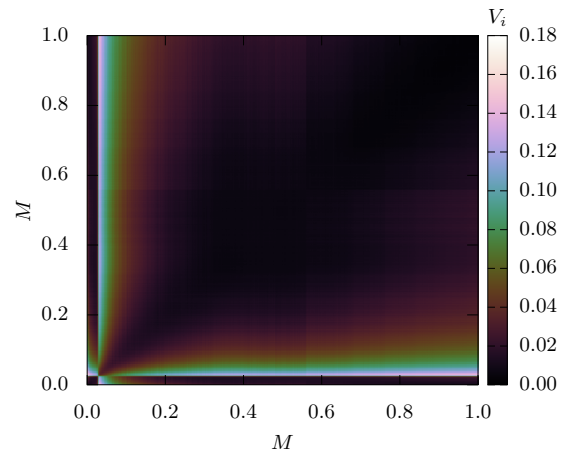

FIG. 24. Disparity V_c of the time series $c(\tau)$ of the fraction of cumulative infections for pairs of time series binned with respect to their total fraction C of infections for $N = 3200$, $\mu = 0.14$, and $\lambda = 0.1763$.

our study. The heat map (not shown) also looks rather similar, though the second region is shifted toward larger M.

Figure 24 shows the disparity when comparing the time series of the cumulative infections. It looks quite similar to Fig. 23. This is also in contrast to the case of classifying the outbreaks according C, where the two disparity heat maps V_i and V_c appeared more different. For $\lambda \in \{0.1, 0.4\}$ the heat maps V_i and V_c (not shown) also look alike.

2. Conditional density

Although we have already studied $\rho(M|C)$, we show in Fig. 25 the distribution $\rho(C|M)$ of C conditioned to the value of M. See Sec. VI A 2 for details. Note that the nonzero values must be located above the diagonal, because for every outbreak $C \geqslant M$ holds by definition.

In general, C and M still are monotonously related. But one can see, there is a sharp increase of C right between regions two and three, near $M \approx 0.08$. In this small interval, outbreaks with relatively small and relatively large values of C lead to the same observed maxima M. Thus, the change from region two to three coincides with a strong change and

FIG. 23. Disparity V_i of the time series $i(\tau)$ of the fraction of infected individuals for pairs of time series binned with respect to their total fraction C of infections for $N = 3200$, $\mu = 0.14$ and $\lambda = 0.1763$.
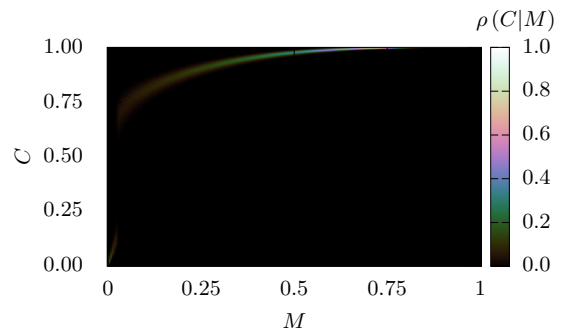

FIG. 25. Conditional density $\rho(C|M)$, which shows the probability of C, i.e., the total fraction of infections, for any given M. The system size is $N = 3200$, the recover probability $\mu = 0.14$, and the transmission probability $\lambda = 0.1763$
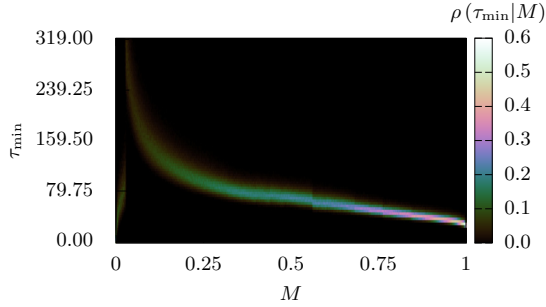
FIG. 26. Conditional density $\rho_s(\tau_{\min}|M)$, which shows the probability of $\tau_{\min}$, i.e., how many time steps it takes until the last node is infected during an outbreak. The system size is $N = 3200$, the recover probability $\mu = 0.14$, and the transmission probability $\lambda = 0.1763$.

FIG. 28. Conditional density $\rho(f_{sw}|M)$, which shows the probability of f_{sw} for any given M. The system size is $N = 3200$, the recover probability $\mu = 0.14$, and the transmission probability $\lambda = 0.1763$.

large fluctuations, similar to the behavior of physical phase transitions. This difference shows that when biasing with respect to M one analyses the behavior in a different way than when biasing with respect to C as in the previous section. The reason is that in fact there is an underlying joint distribution $P(C, M)$, but for sampling this one even down to the tails, one would have to apply a kind of two-dimensional rare-event sampling approach which is currently out of reach for the present problem and the considered graph sizes.

The observed behavior is even more pronounced for $\lambda = 0.4$ (not shown). Here this behavior also corresponds to the position of the kink in the pdfs from Fig. 21. With respect to our algorithmic approaches, we believe that these large fluctuations are a reason why our initially applied standard Wang-Landau approach did not converge and we had to use the replica exchange algorithm.

However, for $\lambda = 0.1$ we do not see such a jump and strong fluctuations in the conditional density (plot not shown). Thus, in the nonpandemic phase, the behavior seems to be simpler, even when including the large-deviation behavior in the analysis.

In Fig. 26 we show for $\lambda = \lambda_c$ the conditional distribution $\rho_s(\tau_{\min}|M)$ of the timescale $\tau_{\min}$ it takes until the outbreak stops to grow, see Sec. VI A 2. Here we see a sharp peak at the position M, where the jump occurred for Fig. 25. This
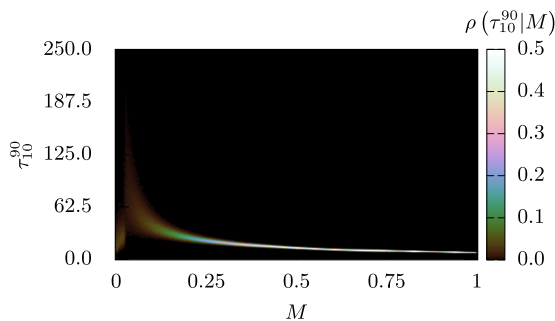
is consistent with the above observations, where region two was also associated with the longest outbreak durations. Also, reaching heavy outbreaks in terms of the fraction M of infections occurring at the same time, does not at all coincide with long-lasting outbreaks, since most of the figure exhibits a negative correlation. For $\lambda = 0.4$ the result (not shown) looks quite similar, though the outbreak dies down quicker overall and the peak is shifted a small bit to larger values of M. For $\lambda = 0.1$ the plot (not shown) looks quit similar as well, though the peak is shifted slightly toward smaller values of M.

In Fig. 27 we show the conditional density $\rho_i(\tau_{10}^{90}|M)$ for the duration τ_{10}^{90} of the highest-activity outbreak phase, see Sec. VI A 2. We can also see a peak which corresponds to region two. Beyond the peak, τ_{10}^{90} is negatively correlated, which appears meaningful, since the larger the peak of the infections in the epidemic phase, the less time it takes for the outbreak to evolve. For $\lambda = 0.4$ the plot (not shown) looks quite similar, though, as one would expect, the durations are shorter over all. Also the peak is shifted toward larger M again. For $\lambda = 0.1$ the plot (not shown) looks almost identical to the plot for $\lambda = 0.1763$.

The conditional density $\rho(f_{sw}|M)$ can be found in Fig. 28. In contrast to the case when conditioning to C, we see a clear monotonous correlation here: the higher the maximum M of the $i(\tau)$ time series, the higher the average fraction of infections which proceeded through long-range edges. This makes intuitively sense: As we have seen in Figs. 27 and 26 a higher peak of $i(\tau)$, i.e., higher value of M, is correlated with faster outbreaks. An infection via long-range edges should accelerate global spread and thus the infection process, leading to larger values of M. The same is observed for $\lambda = 0.4$ (not shown). For $\lambda = 0.1$, however, we again see the anti correlation we also observed when conditioning to C. Note that we use the exact same graph for M and C for all three λ values.

FIG. 27. Conditional density $\rho_i(\tau_{10}^{90}|M)$, which shows the probability of τ_{10}^{90}, i.e., the duration between reaching 10% and 90% of the maximum of $i(\tau)$, for any given value of M. The system size is $N = 3200$, the recover probability $\mu = 0.14$, and the transmission probability $\lambda = 0.1763$.
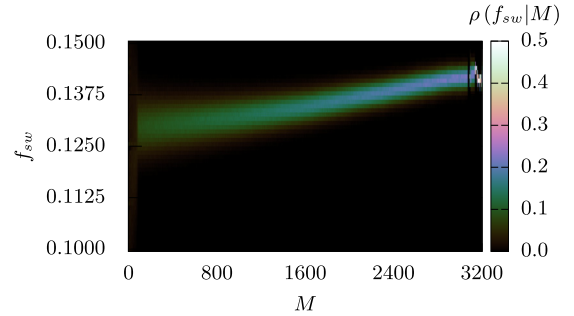
VIII. SUMMARY AND OUTLOOK

We investigated outbreak dynamics for diseases described by the standard SIR model. Our intention was to investigate typical, extremely mild, and extremely severe outbreaks in principle for arbitrary choices of the transmission probability λ and recovery probability μ. Here we considered a

fixed value of μ and three representative values of λ in the local-outbreak phase, in the pandemic phase, and near the pandemic threshold λ_c, respectively. To achieve this, we used large-deviation algorithms, in particular a suitably adapted Wang-Landau approach. We were able to numerically measure, by separate sets of large-scale simulations, the pdfs of the fraction C of cumulative infected individuals and the peak value M of the fraction of infected individuals over the whole range of their support for multiple system sizes. This allowed us to obtain results with probability densities as small as 10^{-2500}. Furthermore, we were able to estimate the rate functions for the distributions of these quantities, showing that the results are compatible with the mathematical large-deviation property. This means that the SIR process belongs to a mild or standard class with respect to the large deviations, such that mathematical tools like the Gärtner-Ellis theorem might be utilized to obtain analytical progress.

More specifically, we studied networks from the small-world ensemble for various system sizes up to $N = 6400$ nodes. To gauge our simulation requirements and parameter-space, we first performed simulation of standard SIR dynamics, i.e., without the large-deviation approach, to obtain the critical transmission probability λ_c and investigated the disease duration Δt_{90}, i.e., how long it takes until 90% of the outbreaks were finished.

Beyond obtaining the pdfs, by comparing the time series that are characteristic for different regions of the pdfs, we were able to see three distinct types of outbreaks: Very mild outbreaks (first region) as well as very severe pandemic outbreaks (third region), with respect to C or M or both, which also evolve very quickly. However, outbreaks in the second region, for intermediate values of C and M, behave somehow chaotic, and here we observe the largest times until they die down.

In this study, we have investigated the most simple case for the SIR model, with the intention to provide a case study proving the feasibility of using large-deviation techniques for epidemic simulations. Clearly, the approach is not limited to the standard case. In the future we plan, e.g., to investigate the effect of disease preventing measures, like lock downs or government orders to wear masks. This can be achieved technically by changing the transmission probability dynamically during an outbreak simulation. The time of change can be static or depend on the outbreak dynamics. Any change of the outbreak behavior will be visible in the measured pdfs, not only in the typical part, but also in the tails, the structure of the different outbreak dynamics and the measured correlations. In a similar way, the effect of vaccinations can in principle easily be measured.

Clearly, the large-deviation approach is also feasible for extensions of the SIR model, e.g., when other states are introduced, like *infected but not infectious* or *in quarantine*, or for spatial models, where the mobility plays a role. Within a longer perspective, this approach can also be used to study the rare jump of, e.g., a virus between populations. This can be achieved by studying two networks simultaneously, i.e., a multilevel network. One network represents an animal population, while the second one represents a human population. Usually the probability of an animal infecting a human is substantially smaller than the probability for human-human and animal-animal infections. Thus, such a transfer leading to a pandemic is a rare event for each single disease. Hence, this case is ideally suited to be target by a large-deviation approach.

[1] H. W. Hethcote, SIAM Rev. **42**, 599 (2000).

[2] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis* (Springer-Verlag, New York, 2000).

[3] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).

[4] C. E. Walters, M. M. Meslé, and I. M. Hall, Epidemics **25**, 1 (2018).

[5] L. Tang, Y. Zhou, L. Wang, S. Purkayastha, L. Zhang, J. He, F. Wang, and P. X.-K. Song, Int. Stat. Rev. **88**, 462 (2020).

[6] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann, Science **369**, 9789 (2020).

[7] W. C. Roda, M. B. Varughese, D. Han, and M. Y. Li, Infect. Disease Model. **5**, 271 (2020).

[8] M. Mandal, S. Jana, S. K. Nandi, A. Khatua, S. Adak, and T. Kar, Chaos, Solitons & Fractals **136**, 109889 (2020).

[9] F. Liu, X. Li, and G. Zhu, Sci. Bull. **65**, 1297 (2020).

[10] V. K. R. Chimmula and L. Zhang, Chaos, Solitons & Fractals **135**, 109864 (2020).

[11] G. E. Weissman, A. Crane-Droesch, C. Chivers, T. Luong, A. Hanish, M. Z. Levy, J. Lubken, M. Becker, M. E. Draugelis, G. L. Anesi, P. J. Brennan, J. D. Christie, C. W. Hanson, III M. E. Mikkelsen, and S. D. Halpern, Annals Intern. Med. **173**, 21 (2020).

[12] A. Karaivanov, PLoS One **15**, e0240878 (2020).

[13] W. O. Kermack and A. G. McKendrick, Proc. R. Soc. London A **115**, 700 (1927).

[14] S. Eubank, H. Guclu, V. S. Anil Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, Nature (London) **429**, 180 (2004).

[15] N. M. Ferguson, D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke, Nature (London) **437**, 209 (2005).

[16] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis* (CRC Press, Boca Raton, FL, 2013).

[17] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

[18] P. Guo, T. Liu, Q. Zhang, L. Wang, J. Xiao, Q. Zhang, G. Luo, Z. Li, J. He, Y. Zhang, and W. Ma, PLoS Neglected Trop. Dis. **11**, e0005973 (2017).

[19] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. Varkonyi-Koczy, U. Reuter, T. Rabczuk, and P. M. Atkinson, Algorithms **13**, 249 (2020).

[20] J. Kim and I. Ahn, Sci. Rep. **11**, 4413 (2021).

[21] W. Wang, M. Tang, H. E. Stanley, and L. A. Braunstein, Rep. Prog. Phys. **80**, 036603 (2017).

[22] W. Wang, Q.-H. Liu, J. Liang, Y. Hu, and T. Zhou, Phys. Rep. **820**, 1 (2019).

[23] Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, and D. Zhao, Phys. Rep. **664**, 1 (2016).

[24] E. Pardoux and B. Samegni-Kepgnou, J. Appl. Probab. **54**, 905 (2017).

[25] R. G. Dolgoarshinnykh (2009), http://www.stat.columbia.edu/~regina/research/sirs.pdf.

[26] Q.-H. Liu, M. Ajelli, A. Aleta, S. Merler, Y. Moreno, and A. Vespignani, Proc. Natl. Acad. Sci. USA **115**, 12680 (2018).

[27] M. E. J. Newman, Phys. Rev. E **66**, 016128 (2002).

[28] M. Boguñá, R. Pastor-Satorras, and A. Vespignani, Phys. Rev. Lett. **90**, 028701 (2003).

[29] J. Verdasca, M. M. Telo da Gama, A. Nunes, N. R. Bernardino, J. M. Pacheco, and M. C. Gomes, J. Theor. Biol. **233**, 553 (2005).

[30] T. Gross, Carlos J. Dommar D'Lima, and B. Blasius, Phys. Rev. Lett. **96**, 208701 (2006).

[31] S. Gómez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno, Europhys. Lett. **89**, 38009 (2010).

[32] W. Wang, Q.-H. Liu, L.-F. Zhong, M. Tang, H. Gao, and H. E. Stanley, Sci. Rep. **6**, 24676 (2016).

[33] J. Li, J. Wang, and Z. Jin, J. Math. Biol. **77**, 1117 (2018).

[34] A. K. Hartmann, *Big Practical Guide to Computer Simulations* (World Scientific, Singapore, 2015).

[35] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[36] J. A. Bucklew, *Introduction to Rare Event Simulation* (Springer-Verlag, New York, 2004).

[37] A. K. Hartmann, Phys. Rev. E **65**, 056102 (2002).

[38] A. K. Hartmann, Phys. Rev. E **89**, 052103 (2014).

[39] F. Wang and D. P. Landau, Phys. Rev. Lett. **86**, 2050 (2001).

[40] J. Lee, Phys. Rev. Lett. **71**, 211 (1993).

[41] F. den Hollander, *Large Deviations* (American Mathematical Society, Providence, RI, 2000).

[42] H. Touchette, Phys. Rep. **478**, 1 (2009).

[43] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications* (Springer, Berlin, 2010).

[44] H. Touchette, in *Modern Computational Science 11: Lecture Notes from the 3rd International Oldenburg Summer School*, edited by R. Leidl and A. K. Hartmann (BIS-Verlag, Oldenburg, 2011).

[45] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, Proc. Natl. Acad. Sci. USA **97**, 11149 (2000).

[46] A. Barrat and M. Weigt, Eur. Phys. J. B **13**, 547 (2000).

[47] D. J. Watts, Annu. Rev. Sociol. **30**, 243 (2004).

[48] Y. Feld and A. K. Hartmann, Chaos **29**, 113103 (2019).

[49] G. E. Crooks and D. Chandler, Phys. Rev. E **64**, 026109 (2001).

[50] A. K. Hartmann, in *New Optimization Algorithms in Physics*, edited by A. K. Hartmann and H. Rieger (Whiley-VCH, Weinheim, 2004), p. 253.

[51] S. Wolfsheimer, B. Burghardt, and A. K. Hartmann, Algor. Mol. Biol. **2**, 9 (2007).

[52] D. A. Adams, R. M. Ziff, and L. M. Sander, J. Chem. Phys. **133**, 174107 (2010).

[53] A. Engel, R. Monasson, and A. K. Hartmann, J. Stat. Phys. **117**, 387 (2004).

[54] A. K. Hartmann, Eur. Phys. J. B **84**, 627 (2011).

[55] A. K. Hartmann and M. Mézard, Phys. Rev. E **97**, 032128 (2018).

[56] C. Giardinà, J. Kurchan, and L. Peliti, Phys. Rev. Lett. **96**, 120603 (2006).

[57] V. Lecomte and J. Tailleur, JSTAT **2007**, P03004 (2007).

[58] W. Staffeldt and A. K. Hartmann, Phys. Rev. E **100**, 062301 (2019).

[59] A. K. Hartmann, Phys. Rev. Lett. **94**, 050601 (2005).

[60] Y. S. Elmatad, R. L. Jack, D. Chandler, and J. P. Garrahan, Proc. Nat. Acad. Sci. U.S.A. **107**, 12793 (2010).

[61] R. L. Jack and P. Sollich, Prog. Theor. Phys. Supp. **184**, 304 (2010).

[62] O. Melchert and A. K. Hartmann, Eur. Phys. J. B **86**, 323 (2013).

[63] P. Werner and A. K. Hartmann, Phys. Rev. E **104**, 034407 (2021).

[64] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[65] W. K. Hastings, Biometrika **57**, 97 (1970).

[66] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon Press, Oxford, 1999).

[67] Q. Yan and J. J. de Pablo, Phys. Rev. Lett. **90**, 035701 (2003).

[68] R. E. Belardinelli and V. D. Pereyra, Phys. Rev. E **75**, 046701 (2007).

[69] R. E. Belardinelli, S. Manzi, and V. D. Pereyra, Phys. Rev. E **78**, 067701 (2008).

[70] R. E. Belardinelli and V. D. Pereyra, Phys. Rev. E **93**, 053306 (2016).

[71] B. J. Schulz, K. Binder, M. Müller, and D. P. Landau, Phys. Rev. E **67**, 067102 (2003).

[72] D. P. Landau, S. Tsai, and M. Exler, Am. J. Phys. **72**, 1294 (2004).

[73] T. Vogel, Y. W. Li, T. Wüst, and D. P. Landau, J. Phys.: Conf. Ser. **487**, 012001 (2014).

[74] T. Vogel, Y. W. Li, T. Wüst, and D. P. Landau, Phys. Rev. E **90**, 023302 (2014).

[75] Y. W. Li, T. Vogel, T. Wüst, and D. P. Landau, J. Phys.: Conf. Ser. **510**, 012012 (2014).

[76] B. Efron, Ann. Statist. **7**, 1 (1979).

[77] F. Nazarimehr, S. Jafari, M. Perc, and J. C. Sprott, Europhys. Lett. **132**, 18001 (2020).

[78] E. Southall, T. S. Brett, M. J. Tildesley, and L. Dyson, J. R. Soc. Interface **18**, 20210555 (2021).

[79] P. Holme, PLoS One **8**, e84429 (2013).

[80] D. S. Fisher, Phys. Rev. Lett. **56**, 416 (1986).

[81] J.-X. Le and Z. R. Yang, Phys. Rev. E **71**, 016105 (2005).

[82] Q.-H. Liu, A. I. Bento, K. Yang, H. Zhang, X. Yang, S. Merler, A. Vespignani, J. Lv, H. Yu, W. Zhang, T. Zhou, and M. Ajelli, PLoS Comput. Biol. **16**, e1008467 (2020).

[83] Q.-H. Liu, J. Zhang, C. Peng, M. Litvinova, S. Huang, P. Poletti, F. Trentini, G. Guzzetta, V. Marziano, T. Zhou, C. Viboud, A. I.

Bento, J. Lv, A. Vespignani, S. Merler, H. Yu, and M. Ajelli, Nat. Commun. **13**, 322 (2022).

[84] S. K. Pani, N.-H. Lin, and S. R. Babu, Sci. Total Environ. **740**, 140112 (2020).

[85] D. Parker and O. Pianykh, Am. J. Epidemiol. **190**, 1081 (2021).