Calculation of epidemic arrival time distributions using branching processes

Alastair Jamieson-Lane [*]

Institute for Chemistry and Biology of the Marine Environment (ICBM), University of Oldenburg, Oldenburg, Germany

Bernd Blasius [†]

Institute for Chemistry and Biology of the Marine Environment (ICBM), University of Oldenburg, Oldenburg, Germany
and Helmholtz Institute for Functional Marine Biodiversity (HIFMB), Oldenburg, Germany

The rise of the World Airline Network over the past century has led to sharp changes in our notions of "distance" and "closeness"—in terms of both trade and travel, but also (less desirably) with respect to the spread of disease. When novel pathogens are discovered, countries, cities, and hospitals are caught trying to predict how much time they have to prepare. In this paper, by considering the early stages of epidemic spread as a simple branching process, we derive the full probability distribution of arrival times. We are able to rederive a number of past arrival time results (in suitable limits) and demonstrate the robustness of our approach, both to parameter values far outside the traditionally considered regime and to errors in the parameter values used. The branching process approach provides some theoretical justification to the "effective distance" introduced by Brockmann and Helbing [Science **342**, 1337 (2013)]; however, we also observe that when compared to real-world data, the predictive power of all methods in this class is significantly lower than has been previously reported.

I. INTRODUCTION

It was said by Jules Verne in 1873 that "The world has grown smaller, since a man can now go round it ten times more quickly than a hundred years ago," that one could travel around the world in 80 days [1]. One can't help but wonder what Verne would think of the advances in technology that have occurred in the century and a half since, which now allow the circumnavigation of the globe in less than 80 hours. The World Aviation Network (WAN) has, over the past century, shrunk the globe to a fraction of its former size, allowing for vast increases in tourism, immigration, and trade. At the same time, diseases that might once have traveled at the speed of cart, ship or train now cross the world in a number of days or hours [2]. In 2003 Severe Acute Respiratory Syndrome (SARS) was able to spread from China to Vietnam, Hong Kong, and Canada within two weeks of being reported to the World Health Organization [3]. Similarly, the 2009 H1N1 flu pandemic first reported in Mexico was able to reach both Europe and Asia within a fortnight [4]. Now, in 2020, we see that coronavirus SARS-CoV-2 has spread quickly throughout China [5] and across the world [6,7], reaching all but a handful of countries within a few short months.

Modeling of disease spread and disease arrival time is critical, both to mitigate the effects of a disease and to determine where best to apply quarantine, screening, and transport restrictions [8] so as to minimize the spread of infection. Similar questions are also critical in the study of invasive species [9,10]. Modeling approaches range from direct SIR-like metapopulation models to highly detailed epidemic simulation models have been created, such as GLEAMviz [11]. Depending on how they are set up, these models can include such factors as vaccination, incubation times, multiple susceptibility classes, nonlinear responses, seasonal forcing, quarantine, and the stochastic movement of individual agents, providing a detailed and comprehensive modeling framework. Unfortunately, in many cases, such models quickly become black boxes in and of themselves—objects to be studied and analyzed only via costly simulation—allowing predictions of the future that can be observed, but seldom understood. This difficulty is further exacerbated by the difficulties of determining parameter values, often a delicate task during the initial stages of a disease outbreak when information is scarce.

Depending on the questions being asked, explicit modeling of such intricate details may be necessary; however, for other questions, such as the determination of epidemic arrival time (AT), it seems that simpler methods may suffice. For this reason, a number of authors [12–15] have proposed a variety of heuristics and metrics—artificial measures of distance based on flight data from the WAN. The goal of such metrics is to predict relative arrival times for an epidemic starting in a specified location, predicting, for example, that an epidemic starting in Vancouver will take twice as long to reach Istanbul as it will to reach Rio de Janeiro. Absolute AT will also depend on the parameters of the particular disease in question.

---

[*]Corresponding author: alastair.jamieson-lane@uni-oldenburg.de
[†]bernd.blasius@uni-oldenburg.de

A particularly elegant example of such a metric is Brockmann and Helbing's [12] "effective distance," in which they define the direct distance between a pair of locations as

$$d_{ij} = 1 - \log(P_{ij}), \quad (1)$$

where P_{ij} is the probability that a particular individual who is leaving location i is traveling to location j. The total effective distance between nodes a and b is then given as the shortest path between a given pair of nodes:

$$D_{ab} = \min \sum_{ij} d_{ij} = \min \sum_{ij} [1 - \log(P_{ij})], \quad (2)$$

where here we minimize over all possible paths from a to b, and d_{ij} denotes the effective distance along each step of this shortest path (and P_{ij} the corresponding probability).

Papers by Gautreau, Ianelli *et al.* provide more detailed and mathematically justified metrics, by first calculating the expected time for a growing epidemic to make its first "jump" [14], and summing over individual paths. Ianelli *et al.* extend this work, providing an algorithm for summing over *all* possible paths joining a pair of airports [13]. More recent work by Chen *et al.* [15] made use of linear spreading theory, and the matrix exponential, in order to attain similar accuracy at reduced computational cost.

Underlying each of the above efforts, either implicitly or explicitly, is an assumption of unbound growth, the idea that in the early stages of an epidemic, disease spread is not limited by population size. Such unbound growth can be modeled as smooth exponential growth, which may be viewed as either a linearization of the standard SIR-type model [14,15] or as a continuous approximation of discrete real-world populations. An alternative model for unbound growth, one often used when small populations, rare events, and probabilities are of interest, is the branching process [16]. The branching process is a classical tool in the study of extinction and evolution [17] and has been used to study epidemics [18,19], surnames and genealogies [20], and cancers [21,22]. In this paper we apply the branching process framework in order to calculate not only the mean but also the full distribution of ATs for arbitrary networks.

In Sec. II we introduce a basic "multicompartment branching process" model, and from this model derive a system of ODEs that precisely determine the probability of epidemic arrival by a given time. In Sec. III we determine the mean and variance in the AT and explore how these results both support and contrast with the results of past papers. Section IV explores the sensitivity of our predictions to perturbations to system parameters and network structure. In Sec. V we use data from real-world flight networks, and compare our predictions, along with the predictions of past authors to epidemic arrival times as observed for in the 2003 SARS epidemic and 2009 H1N1 influenza pandemic. In both cases we observe correlation between predicted and observed results but also significant noise. We are, unfortunately, unable to reproduce previously published results and instead observe correlation far weaker than has previously been reported.

Our goal in this paper is to provide a "canonical" approach to calculating arrival times, that is to say, an approach which relies on minimal assumptions and approximations, and can
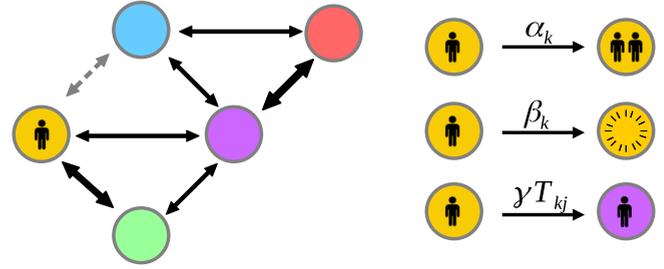

FIG. 1. (Left) Schematic diagram of a traveler in a network; each node representing a city or airport, and edges representing a flight link in the WAN. Edges will have different weights depending on the number of passengers flying. Colors are used purely to distinguish between different nodes. (Right) Each infected individual in location k can take one of three actions at any given moment of time: they may infected another individual (at infection rate α_k), they may recover, removing themselves from the infectious population (at recovery rate β_k), or they may travel to some other location j at transport rate γT_{kj}.

be trusted to provide accurate results across all of parameter space. To the extent that the method works, it can be seen as providing a foundation to past methods and heuristics. Where our predictions disagree with observed arrival times for real-world epidemics, this is suggestive of either flaws in the data, or gaps in the underlying model, gaps that will require not simply better mathematics, but instead better understanding of the system under study.

II. DIFFUSION ON A NETWORK AND THE BRANCHING PROCESS

Let us begin by concretely defining the model; we consider a network of N connected nodes, each node representing a location. In our case these nodes refer to particular airports in the WAN. Individuals travel from node i to node j at transport rate γT_{ij}. Here T_{ij} gives the relative transport rate along a given route, while γ is the global flight rate. While γ and T can be folded together as one parameter, maintaining this distinction will prove convenient later. We select T_{ii} to represent the rate at which individuals *leave* location i; $T_{ii} = -\sum_{j \neq i} T_{ij}$. Aside from this conservation condition, we make no assumptions of structure and symmetry. We delay the discussion of T's relationship to real-world data to Sec. V.

Each node in our network contains a population of $I_k(t)$ infected individuals, initially set to zero. This population evolves according to a continuous time branching process: in any given small time interval dt $I_k(t)$ will increase by one with probability $\alpha_k I_k(t)dt$ and decrease by one with probability $\beta_k I_k(t)dt$. This represents infection and recovery (α and β respectively). Parallel with this branching process, infected individuals may travel from one location, i, to a neighboring location, j, with probability $\gamma T_{ij} I_i(t)dt$, resulting in an increment at location j and a decrement at location i. We thus have what might be described as a "continuous time multitype branching process" [16]. An example network, along with our three epidemic processes are depicted in Fig. 1.

Now that we have defined our epidemic spreading process, we would like to determine the travel time from a to b. Stated formally, suppose a novel diseases originates at time $t = 0$,

at node a. We would like to know the time at which the first infectious individual arrives in location b, that is to say the earliest time such that $I_b(t) > 0$. We denote this epidemic arrival time (AT) as τ_a^b. If no infected individuals ever arrives at b we say that $\tau_a^b = \infty$. The distribution of τ_a^b will depend on the starting location a, the target location b, and the network parameters γ, T, α and β. We wish to understand this dependence and do so by considering the survival function $S_a^b(t) = \mathbb{P}(\tau_a^b > t)$. In the language of Markov chains and statistical mechanics, τ_a^b can be thought of as a First Passage Time-type problem [23]. Related problems have been consider in the study of signaling molecules finding their way to a DNA binding site [24], albeit in a continuous, rather than network-based geometry.

Consider "patient zero," initially placed in location a at time $t = 0$. Now consider the state of the system a short time later, at time $t = dt$. With overwhelming probability, nothing will have happened during this small time window, and the survival times will still be governed by $S_a^b(t)$. With probability $dt\gamma T_{a,k}$ patient zero will have moved to location k, and the survival time distribution will instead be governed by $S_k^b(t)$. In the case where $k = b$, survival is now impossible, hence $S_b^b(t) = 0$. Patient zero may also infect someone with probability $dt\alpha_a$. In this case, the probability of survival until time t is given by the probability that *neither* traveler arrives at b. As both travelers are independent and identical, this probability is given by $[S_a^b(t)]^2$. Finally, it is possible that the initially infected patient zero may simply recover. This happens with probability β_a. In this case, no infected individual will ever reach b, and survival probability is 1 for all time.

The probability of surviving for t units of time after dt is thus a superposition of the possible survival functions for the states the system could transition to. By definition the probability of surviving for t units of time after time dt is also precisely equal to $S_a^b(dt + t)$. Stated algebraically we thus have

$$S_a^b(dt + t) \approx \left[1 - dt\left(\alpha_a + \beta_a + \gamma \sum_{k \neq a} T_{a,k}\right)\right] S_a^b(t)$$
$$+ \gamma \sum_{k \neq a} dt\, T_{a,k} S_k^b(t) + \alpha_a dt\, [S_a^b(t)]^2 + \beta_a\, dt,$$

$$(3)$$

and rearranging and taking limits we find

$$\lim_{dt \to 0} \frac{S_a^b(dt + t) - S_a^b(t)}{dt} = \gamma \sum_{k \neq a} T_{a,k}\left[S_k^b(t) - S_a^b(t)\right]$$
$$+ \alpha_a\left\{[S_a^b(t)]^2 - S_a^b(t)\right\}$$
$$+ \beta_a\left[1 - S_a^b(t)\right]. \quad (4)$$

remembering that $T_{a,a}S_a^b(t) = -\sum_{k \neq a} T_{a,k}S_a^b(t)$, we can simplify to

$$\dot{S}_a^b(t) = \gamma \sum_k T_{a,k} S_k^b(t) + \left[\beta_a - \alpha_a S_a^b(t)\right]\left[1 - S_a^b(t)\right],$$

$$S_b^b(t) = 0, \quad S_{a \neq b}^b(0) = 1. \quad (5)$$

We refer to arrival times calculated using this method as "branching process arrival times" (BP AT). These equations
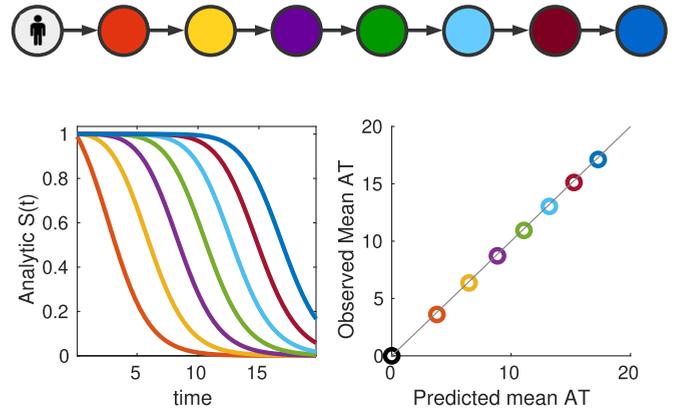

FIG. 2. Here we consider the arrival time for an epidemic propagating down a simple chain (top). (Left) Survival curves, calculated using Eq. (5) are approximately logistic, with the exception of the first step. (Right) Mean arrival times as predicted via Eq. (5) perfectly match mean arrival times as observed over the course of 5000 Gillespie simulations. Parameter values $\alpha = 0.5$, $\gamma = 0.1$. Nodes and survival curve are color coded so as to match the network diagram.

are equivalent to those given by Goldie and Coldman [21], who model the arrival of treatment resistant cancer cells via rare mutation (although in Goldie's case, a much smaller collection of "types" of individual are considered).

It is important to note that Eq. (5) *does not* model the internal state of the system. While normal differential equations might be constructed by modeling the internal "state" of the system at time t, and using this to model the system at $t + dt$, instead we merely observe that the survival function $S_a^b(dt + t)$ (whatever it might be), can be written as a superposition of the (as yet unknown) survival functions $S_k^b(t)$ and hence derive Eq. (3). In so doing, we have lent heavily upon the assumption that system parameters are constant in time, that the behavior and survival curve associated with a single traveler observed at time t is identical to the behavior and survival curve of a traveler observed at time 0. For a discussion of how a similar approach might be taken in the context of time varying parameters, see Appendix A.

A. Example networks

In order to test Eq. (5), and get something of an intuitive handle on the behavior of our system, let us consider three example networks of successively increasing complexity. For each network we solve Eq. (5) numerically using Matlab's ode45 [25] for each target node, and compare to survival times observed in exact agent based simulations, where we track individual infection, migration, and recovery events precisely. Where possible simulations are conducted using Gillespie's exact algorithm [26]. In cases where such an approach is computationally infeasible, we make use of τ leaping [27].

As our first example, consider a simple chain graph, in which infection begins at one end of the chain and is allowed to propagate from one node to the next (Fig. 2). Here we observe that arrival time scales with number of steps taken, as might be expected. Survival curves are sigmoidal, which, as we will later see, is a typical behavior for reasonable parameter values. Mean arrival times as predicted by survival curves
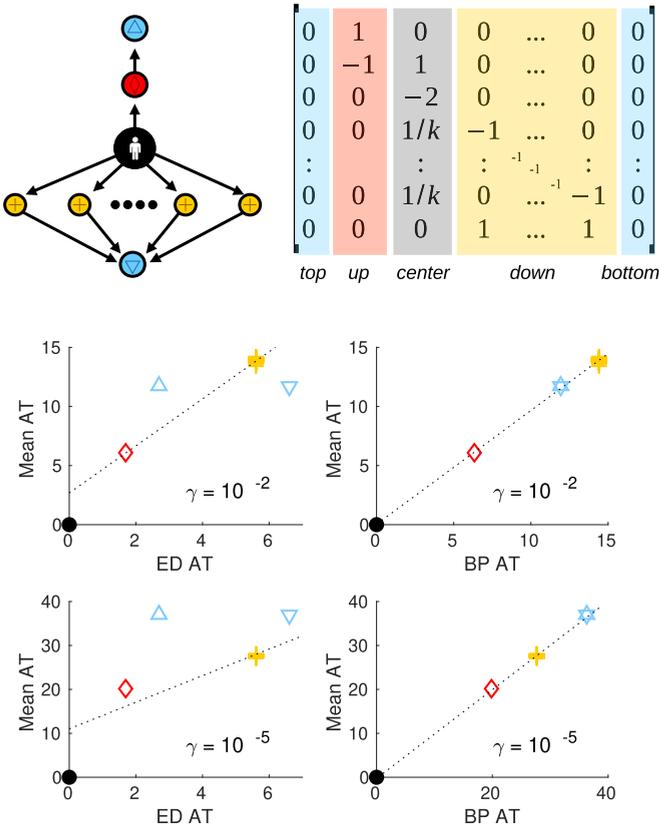
FIG. 3. (Top left) The spinning top graph consists of $k + 4$ nodes: a central starting node (black) with no incoming edges, the initial site of infection, a "top" and "bottom" node with no outgoing edges (blue, triangles), a single "up" node directly between the center node, and the top (red, diamond) and k "down" nodes between the center node and the bottom (yellow, plus symbols). Infected individuals in the central node will travel either up or down at a rate of γ, those traveling down select among the k "down" nodes uniformly at random. (Top right) Transition matrix T for the spinning top graph. (Bottom) Graphs of mean AT (averaged over 150 simulations) vs either effective distance (ED) or branching process arrival time (BP AT) for the spinning top graph. In all cases, we assume $\alpha = 0.5$, $\beta = 0$, $k = 50$, and γ as specified on the scatter plot. Comparison of ED and BP AT demonstrates that ED distinguishes between the "top" and "bottom" node of the graph, while BP correctly handles multiplicity of paths and assigns the pair identical ATs. We also see that BP AT is responsive to changes in γ and correctly predicts the change in expected order of arrival when γ is changed from 10^{-2} to 10^{-5}. Simulations for this figure are carried out via τ leaping.

and as observed over 5000 Gillespie simulations are found to agree.

We next consider a "spinning top" graph (Fig. 3); this graph is constructed in order to illustrate the effects of multiplicity of paths. Here one node of the graph is accessible through only a single possible path, while another node (at equal distance) can be accessed via a great multiplicity of paths, each with a correspondingly reduced probability.

Here we observe that the BP AT gives accurate predictions of arrival time, correctly predicting that the arrival time in the two most distant nodes is equal, regardless of the multiplicity of paths. Counterintuitively, it is also observed that when k (the number of paths) is greater than $1/\gamma$, the mean arrival

time in the outer blue nodes *precedes* the mean arrival time in the intermediate yellow nodes. This occurs because, while the epidemic must reach *some* intermediate node before passing to the outer blue node, it is also perfectly possible for a large number of intermediate nodes to be uninfected at this stage, resulting in a higher *mean* arrival time. In the case $\gamma = 10^{-2}$ this is precisely what occurs, and the ordering of mean arrival times inverts. The qualitative behavior of the system is thus sensitive to transport rate γ. All of this is correctly described and predicted by the BP AT; heuristic models that account only for a single shortest path [such as Eq. (1)] do not capture this effect (see Fig. 3).

As a third example, in order to illustrate the possible complexity of arrival time distributions, we construct an artificial network with a wide spread of transportation, infection and recovery rates, and compare numerical solutions of Eq. (5) with survival curves observed over the course of 25 000 Gillespie simulations [26]. In this case we observe that survival curves exhibit rich, complex behavior. Nonetheless, Eq. (5) correctly predicts the distribution of arrival times as observed in direct agent-based simulation (Fig. 4) over multiple timescales.

The branching process approach gives similar accuracy over a wide variety of parameter values and network structures. A gallery of such comparisons is provided in Appendix B.

III. DERIVATION OF MEAN AND VARIANCE IN ARRIVAL TIME

Equation (5), while accurate, is somewhat cumbersome and does not provide any closed form for values of interest such as expected arrival time and variance in arrival time. If we place no constraints on α, β, and γT, then it is easy to construct survival curves with complex topology such as Fig. 4, and it can be shown that *any* valid survival curve [that is any curve of the form $0 \leqslant f(t) \leqslant 1$, $f'(t) \leqslant 0$, $f(0) = 1$] can be well approximated, for suitably chosen network and parameter values (see Appendix C). While mathematically interesting, such results are of little use to epidemiologists. More useful results can be obtained by placing some mild constraints on parameter values based on real-world understanding.

The first assumption that we make is that we are not interested in epidemics that go extinct before spreading; we care only about those epidemics that reach international prevalence. For simplicity in the following discussion, this is achieved by assuming $\beta_k = 0$ for all k. A discussion of nonzero β_k can be found in Appendix D.

The second assumption that we make in what follows is that epidemic dynamics (infection and recovery) take place on on the time scale of days and weeks, while international travel is a rare occurrence, the vast majority of the human population taking international flights at most once or twice per year [28]. Stated algebraically we have $\gamma T \ll \alpha$; travel is rarer than infection. This mimics the "rare mutation" assumption made by Goldie and Coldman [21] when studying the the development of treatment-resistant cancer cells and rules out slow epidemics such as HIV.

We further assume that infection rate is equal across all locations: $\alpha_k = \alpha$. This greatly simplifies notation and has limited impact upon our final conclusions.
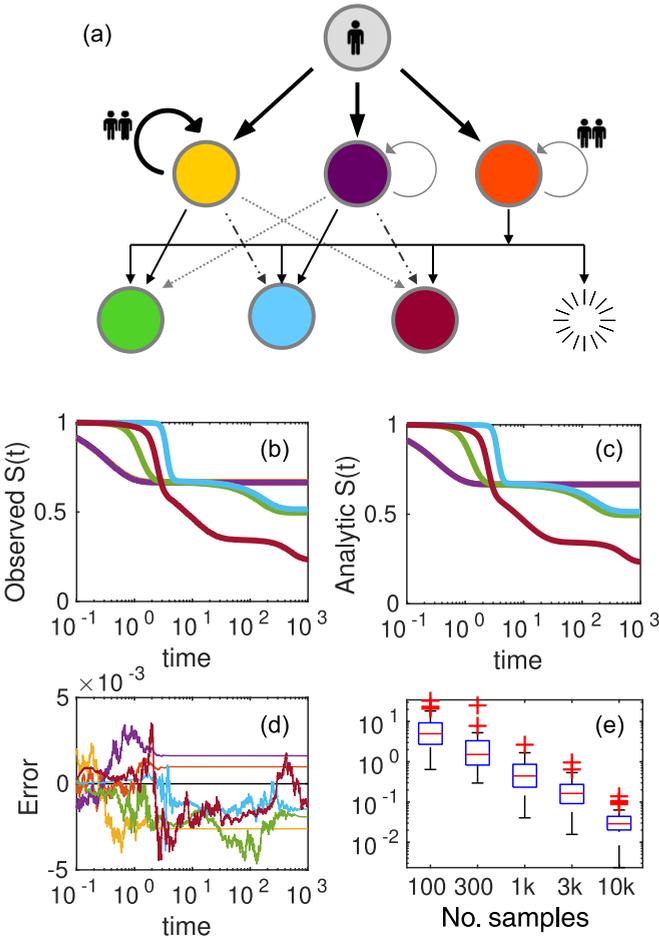
FIG. 4. (a) Here we construct a complex network, designed to highlight the potential complexity of the survival curve $S(t)$. Patient zero is placed at the top of the graph and can transition to one of three "intermediate" nodes. These intermediate nodes have either high infection rate $\alpha = 4$ or low infection rate $\alpha \approx 0.01$, and transition to the three "final" nodes at a variety of different rates (1, 10^{-2}, or 10^{-4}). The third intermediate node permits recovery from infection (transition to the empty circle) at rate 0.01. While this network is not reflective of real-world transportation networks, it does provide a useful illustration of the possible complexity of survival curves, and serves to demonstrate just how closely theory matches simulation even in these complex situations. (b) Survival curves based on 25 000 Gillespie simulations of the epidemic spreading process. Lines are color coded to match the nodes in the network. (c) We solve Eq. (5) and display the analytic survival curve $S_a^b(t)$ for each possible b, keeping a fixed. (d) We plot the difference between the observed and analytic survival curves over time. (e) We sample n of our 25 000 simulations, then calculate $\sum \int |\text{error}|^2 \, dt$ when comparing observed and predicted survival probability. Here we integrate over time and sum over possible arrival locations.

Taken together, these three assumptions ($\beta_k = 0, \alpha_k = \alpha, \gamma T_{ij} \ll \alpha$) are sufficient to avoid the more complex survival curve dynamics, as observed in Fig. 4. Instead, we find $\dot{S}_a^b(t) \approx -\alpha S_a^b(t)[1 - S_a^b(t)]$, and the arrival time distribution is well approximated by a logistic function:

$$S_a^b(t) \approx \frac{1}{1 + e^{\alpha(t - \mu_a^b)}} + O(\gamma). \qquad (6)$$

Here μ_a^b is the mean arrival time for an epidemic starting in node a and spreading to b. Determination of μ_a^b is of some considerable interest. This value is dependent upon the dynamics of the system when $[1 - S_a^b(t)] = O(\gamma)$; that is to say, the region where the transport $\gamma T S_a^b(t)$ is no longer overwhelmed by logistic decay in probability.

A. Mean arrival time

Determination of μ_a^b can be seen as equivalent to the question of determining when the exponential term of Eq. (6) is equal to 1. For the sake of convenience, we label this term $Q_a^b(t)$ and make use of the change of variables $S_a^b(t) = 1/[1 + Q_a^b(t)]$. This leads to the following differential equation:

$$\dot{Q}_a^b(t) \approx \alpha Q_a^b(t) - \left[1 + Q_a^b(t)\right]^2 \gamma \sum_k T_{a,k} \frac{1}{1 + Q_k^b(t)}. \qquad (7)$$

Given Eq. (6), the mean arrival time μ_a^b satisfies $Q_a^b(\mu_a^b) = e^{\alpha(\mu_a^b - \mu_a^b)} = 1$. Because $Q_a^b(t) < 1$ for $t < \mu_a$, and $\gamma T \ll \alpha$, we can safely assume $\alpha Q_a^b(t) \gg \gamma \{2 Q_a^b(t) + [Q_a^b(t)]^2\}$ in our region of interest, hence allowing the approximation $[1 + Q_a^b(t)]^2 \approx 1$. Because $\sum_j T_{ij} = 0$, we can restate Eq. (7) as

$$\dot{Q}_a^b(t) \approx \gamma \sum_k T_{a,k} \frac{Q_k^b(t)}{1 + Q_k^b(t)} + \alpha Q_a^b(t). \qquad (8)$$

At this stage it may be tempting to make the further approximation $\frac{Q_k^b(t)}{1 + Q_k^b(t)} \approx Q_k^b(t)$. Switching to matrix-vector notation and setting $\mathbf{Q}^b(t)$ as a vector whose entries are $Q_a^b(t)$, we find

$$\dot{\mathbf{Q}}^b(t) \approx (\gamma T + \alpha I) \mathbf{Q}^b(t), \qquad (9)$$

and using the matrix exponential $e^M = I + M + M^2/2! + \cdots$ we solve to find

$$\mathbf{Q}^b(t) \approx e^{(\gamma T + \alpha I)t} \mathbf{Q}^b(0). \qquad (10)$$

Here we observe that Eqs. (9) and (10) bear a striking resemblance to the calculation of the expected number of infections at b, given an initial epidemic starting at a, as studied by Chen *et al.* [15]. Under this interpretation of $Q_a^b(t)$ we are effectively approximating our arrival time as the time at which the *expected* number of infected individuals in our target node, b, reaches one. In cases where a single path of length d dominates the spread of our epidemic from a to b we can make the following further approximation:

$$Q_a^b(t) = 1 = e^{\alpha t} \left[I + \frac{\gamma T t}{1!} + \frac{\gamma^2 T^2 t^2}{2!} + \cdots \right]_{a,b} \qquad (11)$$

$$\approx e^{\alpha t} [T^d]_{a,b} \frac{\gamma^d t^d}{d!} \qquad (12)$$

$$\alpha t \approx -\log \left(\frac{\gamma^d T_{a,b}^d t^d}{d!} \right). \qquad (13)$$

Choosing to ignore a number of small terms, we find

$$\alpha t \approx -d \log \gamma - \log \left(\prod T_{i,j} \right) \qquad (14)$$

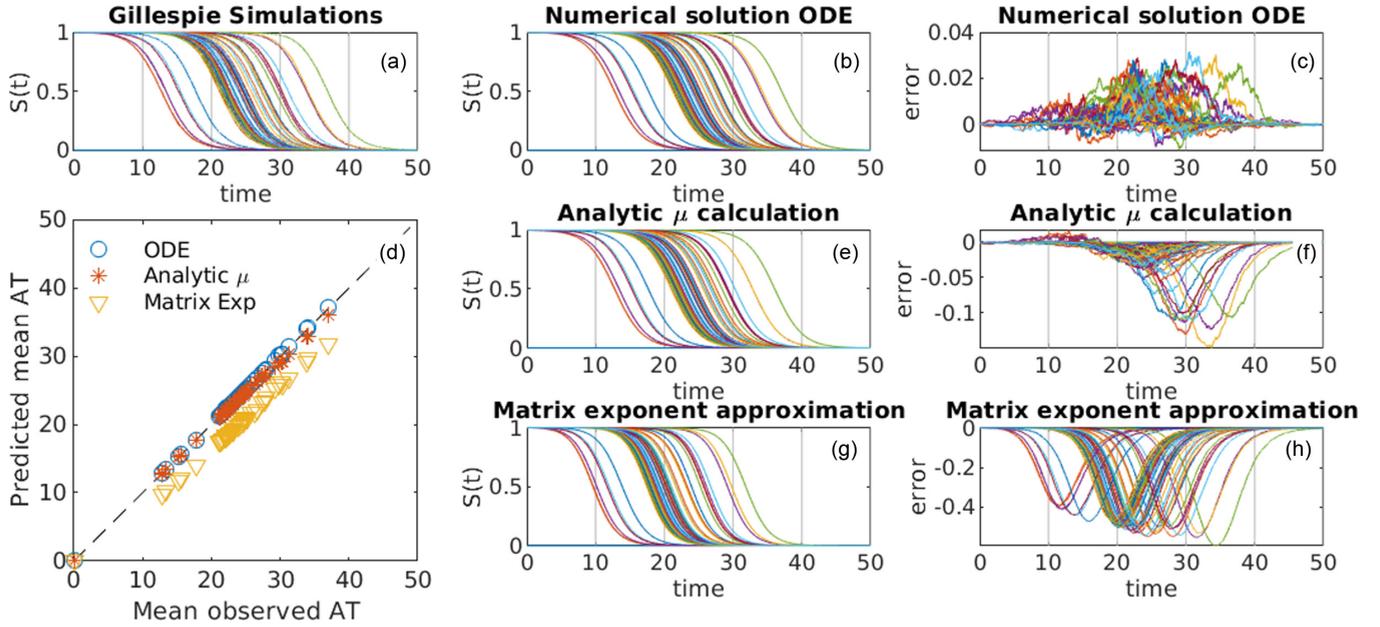$$= -\sum (\log \gamma + \log T_{i,j}), \qquad (15)$$

FIG. 5. Comparison of arrival times as observed in 5000 direct Gillespie simulations of epidemic spread on a randomly generated network with $N = 155$ nodes (a), to survival times as predicted using either numerical solutions [Eq. (5), (b)], logistic curves centered according to μ [Eq. (18), (e)], or logistic curves with $Q(t)$ as defined by simple matrix exponentiation [Eq. (10), (g)]. In the right-hand column we give the difference between predicted and observed survival probabilities for each method. Note the varying scales of the y axis. (d) Comparison of mean AT as observed in 5000 simulations with the mean AT as predicted by the three numeric methods shown in the adjoining panes.

where here $T_{i,j}$ are the transport rates along each link of the most probable path epidemic path. We thus reconstruct the results of shortest path methodology similar to Gautreau's original study [14] and the work of Brockmann and Helbing [12]. More detailed discussion of this matrix exponential approach and its relation to previous work is provided in Appendix E. Survival curves as predicted by Eq. (10) are given in Fig. 5.

While the approximation $Q/(1 + Q) \approx Q$ is convenient in that it allows us to linearize the equation, more accurate results can be obtained by retaining $1/(1 + Q)$ and following on from Eq. (8) directly. Applying integrating factors we find

$$e^{-\alpha t} Q_a^b(t) \approx \int_0^t e^{-\alpha \tau} \gamma \sum_k T_{a,k} \frac{Q_k^b(\tau)}{1 + Q_k^b(\tau)} \, d\tau. \qquad (16)$$

Substituting the ansatz $Q_k^b(\tau) = e^{\alpha(\tau - \mu_k^b)}$ for $k \neq a, b$, along with $\frac{Q_b^b(t)}{1 + Q_b^b(t)} = 1 - S_b^b(t) = 1$, we find

$$e^{-\alpha t} Q_a^b(t) \approx \gamma T_{a,b} \int_0^t e^{-\alpha \tau} \, d\tau$$

$$+ \sum_{k \neq b} \gamma T_{a,k} \int_0^t e^{-\alpha \tau} \frac{e^{\alpha(\tau - \mu_k^b)}}{1 + e^{\alpha(\tau - \mu_k^b)}} \, d\tau. \qquad (17)$$

Multiplying through by $e^{\alpha t}$ and integrating provides

$$Q_a^b(t) \approx \gamma T_{a,b} \frac{e^{\alpha t} - 1}{\alpha}$$

$$+ \sum_{k \neq b} \gamma T_{a,k} \frac{e^{\alpha(t - \mu_k^b)}}{-\alpha} \log\left(\frac{e^{-\alpha t} + e^{-\alpha \mu_k^b}}{1 + e^{-\alpha \mu_k^b}}\right). \qquad (18)$$

In effect, Eq. (18) solves exactly for one particular $Q_a^b(t)$ under the assumption that all other survival curves are logistic. If we then demand μ_a^b such that $Q_a^b(\mu_a^b) = 1$, Eq. (18) gives a system of $N - 1$ equations and $N - 1$ unknowns [there is no equation associated with $Q_b^b(t)$]. This system of equations can be solved using Newton's method, allowing accurate approximation of $Q_a^b(t)$ and hence $S_a^b(t)$ without resorting to computationally expensive numerical ODE solvers. This approximation is particularly helpful in regions of parameter space where γ is very small, and $Q_a^b(t)$ spends large amounts of time in the sensitive region near 0, rendering more direct numerical methods unstable.

A comparison of the various approaches described in this section is given in Fig. 5. Numeric solutions of Eq. (5) give the best results (when using high-accuracy ODE solvers), while predictions based on simple matrix exponentiation [Eq. (10)] are the least accurate. All three approaches are highly correlated with mean arrival times as observed in full agent-based simulations, however. When we are interested in relative rather than absolute arrival time, any of the three approaches will suffice, and matrix exponential-based approaches are the fastest (see Appendix E for details on efficient computation). When more accuracy, or knowledge of full probability distributions, is required, we recommend using either Eq. (5) or (18).

B. Variance

Another value of practical interest is the variability in arrival time; recognizing the difference between 52 ± 3 days and 52 ± 30 days allows us to understand how *precise* we expect arrival predictions to be. When $\gamma T \ll \alpha$, we can use standard logistic distribution results [29] to determine

$\operatorname{var}(\tau_a^b) = \frac{\pi^2 s^2}{3} = \frac{\pi^2}{3\alpha^2}$, where s is the "scale parameter" of the logistic distribution, in our case equal to $1/\alpha$. In plain language, we find that variance is high for slowly growing epidemics, and low for fast growing epidemics, where the infected population spends shorter periods of time at intermediate levels.

This determination of variance leads naturally to the question of covariance; if our epidemic arrives three days earlier than expected in Paris, will it also inevitably arrive three days early in Istanbul or São Paulo? Or are the two arrival times relatively independent? This question is of particular importance because, generally speaking, we do not *know* when an epidemic has started, and hence can only ever compare arrival times in different cities relative to one another.

Unfortunately, by its nature $S_a^b(t)$ contains minimal information about the state of the system; calculation of such state information using probability generating function techniques [30] would require us to take infinitely many derivatives of $S_a^b(t)$ with respect to our initial conditions $S_a^b(0)$. It would appear that full covariance and correlation information is thus inaccessible.

One avenue that we *can* take is to examine the variance in arrival time given a larger starting population, for example, $p = 50$, with all individuals starting at location a. Such an assumption effectively removes the large levels of variance associated with the epidemics "initial take-off" time (the time taken to grow from $p = 1$ to $p = 50$), and leaves only the variation associated with the spreading process throughout our network (and subsequent take off time in the various locations the epidemic emigrates to). Given that we typically do not get to observe the true epidemic start data in practice, such an assumption is likely to better reflect real-world data.

The survival time distribution for a starting population of p is simply $[S_a^b(t)]^p$. It can be shown (see Appendix F) that for an initial population p, the arrival time has variance

$$\operatorname{var}(\tau) = \frac{\pi^2}{3\alpha^2} - \sum_{k=1}^{p-1} \frac{1}{\alpha^2 k^2}. \tag{19}$$

In the limit $\sum_{k=1}^{\infty} (\alpha k)^{-2} = \pi^2/6\alpha^2$.

Of further note, we also observe that as p increases the pth power of the logistic curve $[(1 + e^{(t-\mu)})^{-p}]$ approaches the Gumbel distribution (Fig. 6) as observed by Gautreau *et al.* [14] in their original study of the epidemic arrival time process. The Gumbel distribution can be loosely thought of as an exponential waiting time with exponentially increasing rate parameter (corresponding to the growing population) and has cumulative distribution function of the form $\exp[e^{-(t-\mu)\beta}]$. The continuous population assumption implicit in this result is valid when $p \geqslant 100$ but causes noticeable discrepancies for $p \leqslant 5$.

IV. ROBUSTNESS AND LIMITATION

So far we have explored the branching process model in what may be considered close to optimal conditions; we have assumed that T, α, and γ are exactly known, and that the susceptible population of each node is large enough so as not to limit epidemic growth and spread. Each of the above assumptions may be violated in one context or another, and
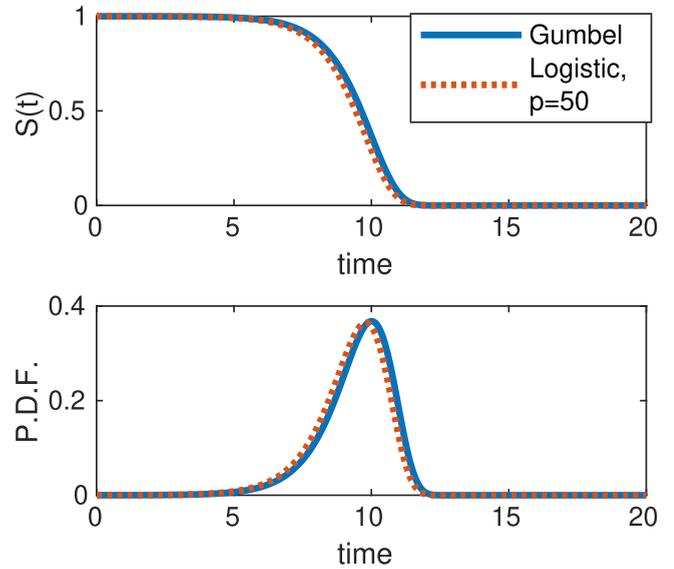

FIG. 6. Here we compare the Gumbel distribution (the survival curve identified by Gautreau *et al.* when modeling population as continuous [14]), and the exact survival curve assuming a initial population $p = 50$, as discussed in Sec. III B. (Top) Survival curve for Gumbel distribution ($\exp[-e^{(t-\mu)}]$) or logistic ($[1 + e^{(t-\mu)}]^{-p}$) like arrival times with $p = 50$. (Bottom) Probability density function for the same. Here we compare only shape, and hence, in this example the curves have different μ parameter; this acts only to slide the mean value so as to make the curves more comparable.

for this reason it is critical to understand how far each can be stretched. Such understanding sheds light not only on the BP AT itself, but also on past distance metrics, which we have shown to rely on a similar theoretical foundations.

The first, and most likely assumption to be violated in practice is the notion that we know α and γ. While γ can be determined to some reasonable level of accuracy via commercial flight data, α is a number that will vary from illness to illness and may be known only to a limited degree of accuracy, particularly in the early stages of a pandemic. Fortunately the model turns out to be robust against variations in both α and γ—even when varying these parameters by an order of magnitude compared to the "true" values used in simulations, predicted ATs remain highly correlated with those observed in simulation (see Fig. 8). When using incorrect α the constant of proportionality between τ_i and observed ATs is no longer equal to one, suggesting that the BP model robustly estimates $\alpha \tau_i$, such that the *relative* time of arrival at any two locations is largely independent of α and γ, but that incorrect estimates of α will lead to corresponding inaccuracy in the *absolute* values of τ_i; if α is estimated a factor of two too high, then τ_i will be a factor of two too low.

Another systematic source of error is inaccuracies in flight network data. These can be produced due to uncounted or unregistered flights (or alternative forms of transportation), out of date data, or as a direct result of changes to individual flight plans as a result of the epidemic itself. In order to test robustness to noise in available flight data, we compare the results of full Gillespie simulation using flight network T, to ATs predicted using the perturbed matrix $\hat{F}$, where
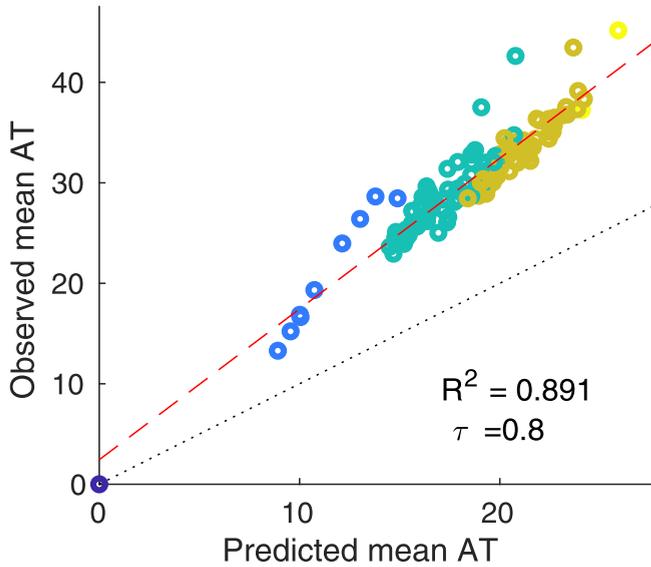
FIG. 7. When populations are very small, the exponential growth assumption that the BP AT relies upon no longer gives accurate predictions. Here we plot expected arrival time according to the BP model, against the average results of 5000 Gillespie simulations on a random graph of 135 nodes, where the susceptible population in each node is 18.2 ± 7.5. We assume infection rate $\alpha = 0.5$, and recovery rate $\beta = 0$ (for small populations, $\beta > 0$ generally leads to extinction). Dots are color coded based on the minimum *number* of steps required to get from the source to target location; bluer (darker) nodes being closer to the source, and yellow (lighter) for nodes which are many steps away. For each graph we give both the R^2 and Kendall-τ correlation statistics. The dotted black line gives the "1:1" line that would be expected of accurate predictions. The dashed red line gives the line of best fit, as used to calculate R^2.

$\hat{T}_{ij} = T_{ij}e^{N(0,\xi)}$. Here ξ is a constant representing noise level. As can be seen in Fig. 9, the resulting errors are modest.

While we have thus far considered arrival times in the context of the global aviation network, effective distance measures have also been considered on a much smaller scale in order to study the spread of antibiotic resistance between wards within a hospital [31]. In this case, assuming unbound growth becomes problematic, as the susceptible population in each node is rather small; a single ward might contain (for example) 10–30 beds, a far cry from the tens or hundreds of thousands found when each node represents an entire city. When the local susceptible population is small, the epidemic is frequently no longer in an exponential growth phase by the time it spreads to a neighboring node; "branching random walks" are no longer independent because it is impossible for person A to infect someone who has already been infected by person B. Population saturation inevitably leads to less accurate predictions; see Fig. 7.

In practice, for the SIR-type models the parameter window where this concern is relevant is relatively narrow; infections which grow quickly will infect the entire local population and drive themselves to extinction before spreading between nodes. For SIS- and SI-type models, or any infection which is expected to reach a local endemic equilibrium, it is important to determine whether exponential growth is a good approx-

imation before making use of the BP AT or any distance metric that relies upon the same underlying unbound growth assumption.

V. REAL-WORLD DATA

Finally, while comparison to simulations provides an effective test bed, useful in terms of repeatability and certainty of data, we are more interested in the performance of models as they apply to the real world. For this purpose, we make use of flight data provided by Dirk Brockmann (private communication) to construct T, and compare observed and predicted AT for both SARS and H1N1 outbreaks. In order to do this, we must consider one major and one minor complication.

Firstly, in order to calculate the BP AT distribution, we would prefer to have access to the full transport matrix T_{ij}, the probability that an individual in location i will travel to j on any given day. Instead, we have access to F_{ij}; a matrix containing the total *number* of passengers traveling from i to j during a given time period. This matrix is approximately symmetric, reflecting the fact that most flights are round trips. In order to obtain T from F, we would need to know the total population, p_i, serviced by each airport; this would give $T_{ij} = F_{ij}/p_i$. Unfortunately, the population serviced by any given airport is unknown and may well be ill defined.

In order to approximate T_{ij}, we assume that the total number of flights entering and exiting an airport is proportional to the local population p_i. In this case, T is found by dividing each column of F by the total number of flights in that column and is equivalent to Brockmann *et al.*'s P_{ij}, the probability to fly from i to j, *conditioned* on the assumption that we board some plane in airport i [12]. The parameter γ can then be interpreted as the probability that a randomly selected person boards a plane on a given day, regardless of their destination. In the real world, we might assume γ to vary significantly from place to place, based on economic factors, prevalence of tourism, and the size of the population that any given airport is servicing. For the time being we approximate γ as a constant and rely on the model's low sensitivity to γT to prevent difficulties.

A second consideration, this time far more specific to the particular data available, is the fact that epidemic arrival times are given on a country by country basis, while flight data is provided on a city-by-city basis. Coarse graining can be achieved in two possible ways: it is possible to either coarse grain our transport matrix, calculating the total number of flights between each country and forming T from there, or we can use the fine-grained city by city matrix, and then set the arrival time in a given country as the *minimal* arrival time in any of that countries airports.

With these two details taken care of, we are now ready to compare the observed arrival times of H1N1 and SARS to predictions made using either the branching process arrival time [Eq. (5)] or the effective distance metric [Eq. (2) [12]]; see Fig. 10. In order to best compare to the previously published results of Brockmann *et al.*, we make use of the coarse grained country-to-country T matrix. Qualitatively similar results are observed in all cases when using the fine grained city-by-city T matrix.
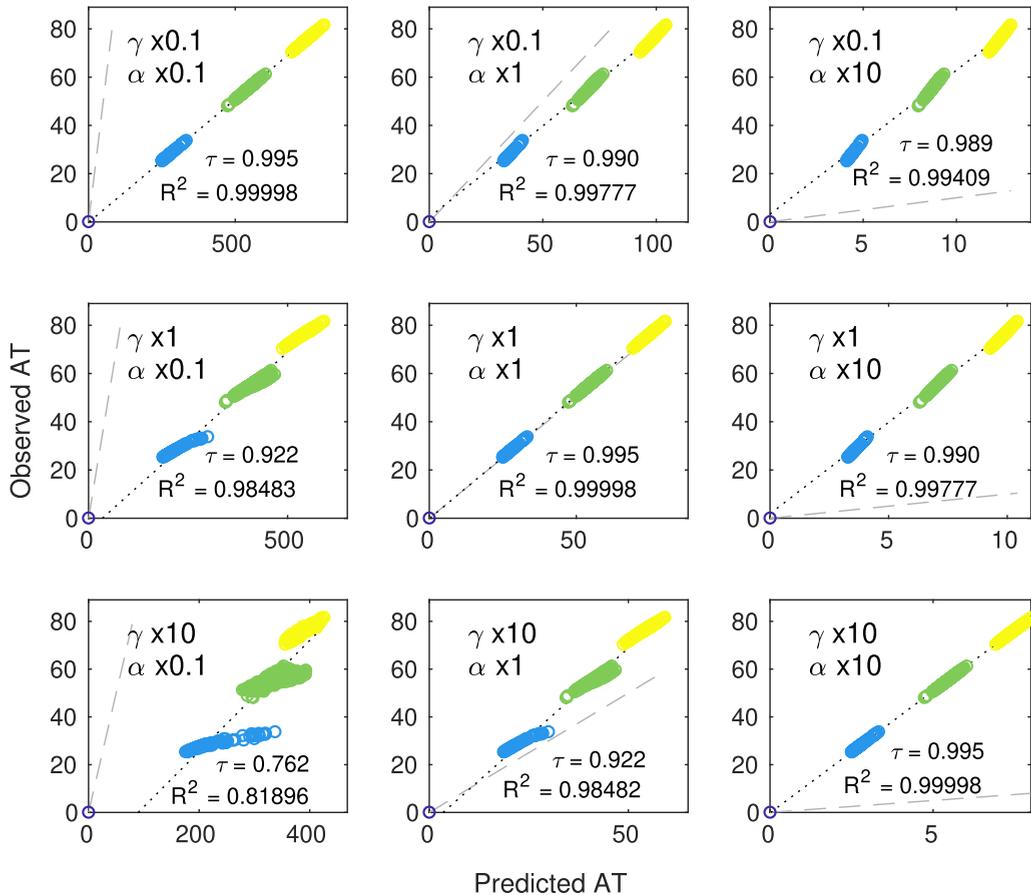
FIG. 8. Calculating the BP AT using incorrect growth and transport parameters leads to suboptimal results. Here we compare the predicted mean AT [according to Eq. (5)] with mean observed AT (averaged over 5000 agent-based Gillespie simulations) in the case of using α and γ which are either a factor of 10 too high or to low compared to their true values ($\alpha = 0.5$, $\gamma = 10^{-3}$). Here we consider a network with $N = 135$ nodes, where each each node is linked to each other node with probability 8/135. For each graph we give both the R^2 and Kendall-τ correlation statistics. Data are colored according to the minimum number of flights from origin to target [blue (dark) for one flight to yellow (light) for three]. The 1:1 (perfect agreement) line is given in gray. The black dotted line gives the best linear fit used to calculate R^2. While errors change the magnitude of predicted AT (note varying axis scales), the effect on *relative* AT predictions is minor in eight out of the nine cases considered.
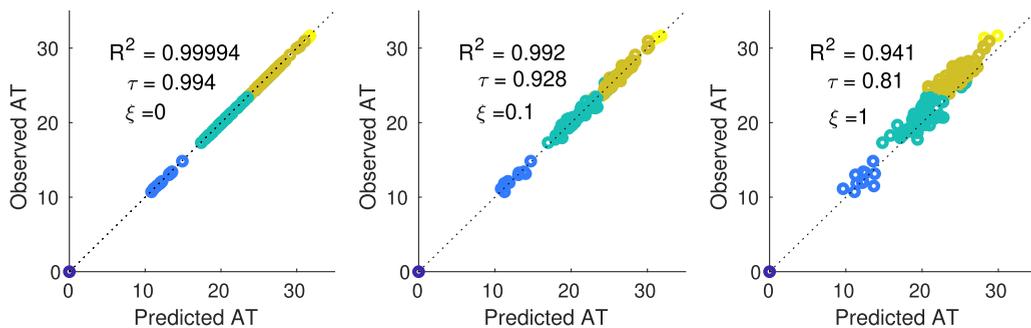

FIG. 9. Comparing the results of Gillespie simulations with the predictions made using a perturbed version of the original random network. We consider a network with $N = 135$ nodes, where each each node is linked to each other node with probability 8/135. We are still able to accurately predict AT, even in the case of significant perturbations to T, indicating that BP AT is robust even when link weights are increased or decreased by an order of magnitude. For each graph we give both the R^2 and Kendall-τ correlation statistics. Nodes are color coded based on the number of steps to the initial infection site, with blue nodes (lower left) accessible with only a single flight, and yellow nodes (top right) requiring four. The 1:1 line, representing perfect prediction, is given by the dotted black line. Note that here, after noise is applied to calculate $\hat{T}$, we demand symmetry and recalculate diagonal entries so as to preserve mass. Similar results are observed for scale-free networks.

We find that both BP AT and effective distance arrival time ("ED AT") methods give qualitatively similar results (see Fig. 10), and that these results would appear in many

cases plausible: for example, in the case of SARS, we observe that predicted arrival times in both South Korea and Hong Kong are low, while the arrival time in the United States
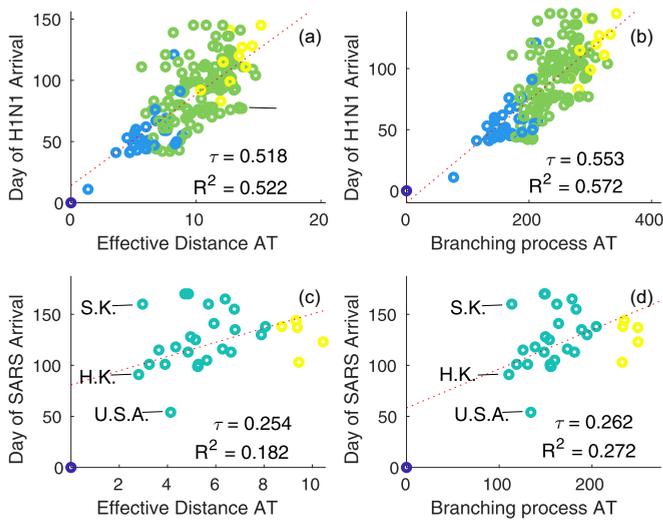
FIG. 10. Comparison of distance metrics to arrival times in real-world epidemics for the 2009 H1N1 epidemic (top) and the 2003 SARS epidemic (bottom). As in previous figures, color coding of nodes represents the minimum number of flights from one location to another, with yellow (pale) nodes requiring more flights and blue (darker) nodes fewer. Unlike previous plots, here each node represents a country. For each graph we give both the R^2 and Kendall-τ correlation statistics. (a) We attempt to reconstruct Fig. 2D of Brockmann and Helbing (2013) [12]; we find that numerous details of the general *shape* are reconstructed (for example the "branch" indicated by the dash, and the cloud of separated points along the top of the panel), but that the figure as presented in the original paper is slanted significantly closer to the diagonal. (b) For suitably chosen α and γ values, the BP method improves upon R^2 but is still limited by the accuracy of the underlying network, and the inherently noisy details of real-world epidemiology. (c), (d) In the case of the SARS epidemic we were not able to reproduce the extremely tight correlation previously reported [12]. Once again, use of the BP metric yields improvements in R^2 compared to our reconstruction the ED metric. The three marked nodes in the bottom images corresponding to (from top to bottom) South Korea, Hong Kong, and the United States. As might be expected for distance measure based on flight data, both Korea and Hong Kong are "close" to China according to both metrics, while the United States is farther away. This would appear to be in contrast to previously published results [12], which appear to indicate that the United States is closest to China, while Korea is rather distant. Data provided by Dirk Brockmann (private communication).

is predicted to be marginally higher. This result is entirely plausible given airline traffic between the respective countries. In practice, predictions for both BP AT and ED AT do not match observed arrival times: SARS is reported to have arrived in the United States 54 days after the first reports in China (16 November 2002), and in South Korea a full 160 days after these first reports. For SARS, arrival times as reported from real-world data correlate at best weakly with predictions [note the small $R^2 < 0.3$ in Figs. 10(c) and 10(d)], indicating either inaccuracy in the data or alternatively some complexity in the real-world process not accounted for in current models. R^2 values are somewhat better when predicting H1N1 ($R^2 \approx 0.55$).

We find no method (including ED) is able to reproduce the very high R^2 values reported in Brockmann and Helbing's original paper, and that we are unable to recreate their figures [Figs. 2(d) and 2(e) in the original paper], instead observing a far broader scatter. It is unclear if the observed discrepancies is due to differences in implementation of the algorithm, differences in the underlying data, or some other cause. While it is possible that inaccuracies in WAN data are to blame for these discrepancies, this explanation seems unlikely; as demonstrated in Fig. 9, arrival time predictions are generally stable, even to substantial changes in F. If we consider the effective distance metric [Eq. (2)], increasing $D_{i,j}$ by 5, as is needed to correctly predict SARS late arrival time in South Korea, would require a $P_{i,j}$ two orders of magnitude smaller than our current value. Such drastic changes appear implausible.

VI. CONCLUSIONS

Better understanding the spread of epidemics through the WAN allows for both real-time forecasting (as has been used during the current COVID-19 pandemic) and the possibility of network design, making changes to the WAN or local transport networks so as to slow epidemic spread [8].

In studying the question of epidemic arrival time, a variety of models have been used, from the the most intricate agent based simulations [11] to the intuitively appealing "distance"-based models [12] and a number of analytic approaches in between [13–15]. While each of these models approaches the question of epidemic arrival time from a different view point, one common thread is the assumption of exponential growth, often justified as a linearization of more classical SIR-type models. Spread is governed by unbound epidemic growth, and rare transportation events. Population saturation, detailed viral dynamics, and a travelers tendency to return to their port of origin are ignored in all but the most detailed of models.

Given this basic premise, we were able to formulate the problem of epidemic arrival times in term of "branching processes" [16,21] and calculate the full probability distribution of possible arrival times explicitly. These predictions match perfectly to corresponding simulations. If we further assume that air travel is rare and infection and recovery are common, it is possible to rederive many of the results of past papers and in some cases improve upon them. We are also able to give predictions in regions of parameter space where past methods are known to fail, and show that theoretical predictions provide a reasonable match to simulations, even when parameter values are altered significantly.

Unfortunately, when comparing to real-world data we observe that the predictive powers of ED metrics may be significantly lower than previously reported [12]. When compared to real-world data, ED and BP like methods predict roughly 50% of variance in the cases of H1N1, and only 20% of variance when compared to SARS-2003, significantly lower than would be predicted given the modest intrinsic variance of our models. While the Branching Process Arrival Time makes very few assumptions and gives exact results given the underlying model assumed, these findings are suggestive of a gap in our knowledge—either in the data available for epidemic arrival times or (more likely) in the model
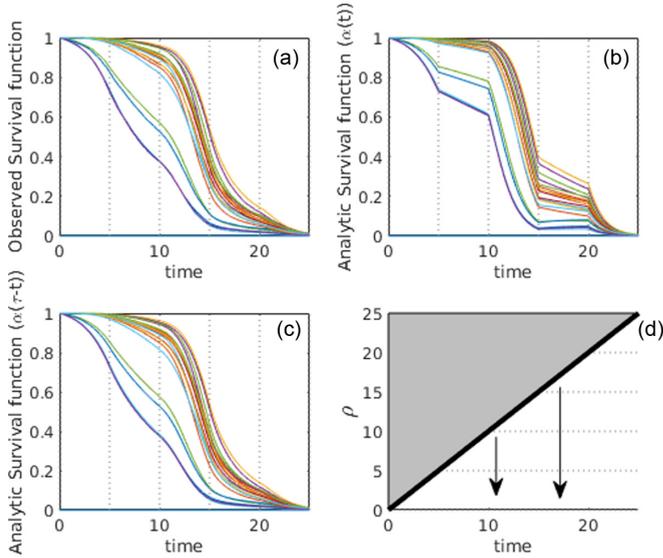
FIG. 11. (a) Survival curves observed over the course of 12,000 simulations, for piecewise constant α_k. The values of α_k change when $t = 5, 10, 15 \ldots$, indicated by the dashed gray lines. $\beta_k = 0$, $\gamma = 0.063$. (b) $S_a^b(t)$ as calculated using Eq. (5) by simply substituting in the time varying $\alpha_k(t)$ in for α_k. This approach violates our Markov assumption and gives nonphysical results. (c) Survival curves $S_a^b(t, 0)$, calculated using Eq. (A1). By correctly accounting for time variation in α_k we are able to match observed survival curves exactly. (d) Schematic diagram showing the flow of information in Eq. (A1); solution curves trace back from our initial conditions at $t = \rho$ to our survival curve of interest at $\rho = 0$. The invalid region $t < \rho$ is greyed over.

itself. It seems likely that future efforts would be best directed towards determining what real-world factors are currently unaccounted for, and which of these are most critical.

Full code for simulation based figures is available on github [32].

ACKNOWLEDGMENT

APPENDIX A: TIME-VARYING PARAMETERS AND FLIGHT NETWORKS

Construction of Eq. (5) relies heavily upon the assumption that the process is memoryless and does not keep track of time.

In order to see the importance of this Markov assumptions, suppose we take Eq. (5) and replace α with some piecewise constant function $\alpha(t)$: both in our simulation and in the ODE. As can be seen in Fig. 11 (top panels), simply replacing α with $\alpha(t)$ leads a rapid divergence of $S_a^b(t)$ and our observed survival curve. Under such naive assumptions, our calculated $S_a^b(t)$ is no longer monotone decreasing, and hence can no longer be considered a survival curve in any sense of the word. This breakdown comes about because even though $S_a^b(t)$ does
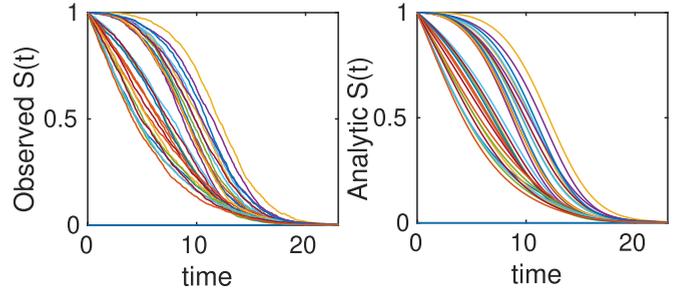

FIG. 12. Here we simulate a scale-free network as produced by Mathew George's SFNG function [33]. Here we set $N = 30$ nodes, infection rate $\alpha = 0.5$ and recovery rate $\beta = 0.1$. (Left) Observed survival curves given 1500 simulations. Simulation done using the exact Gillespie algorithm [26]. (Right) Survival curves as calculated using Eq. (5).

not track possible states of the population directly, it must (inevitably) encode these states implicitly. When α changes midway through the process, this encoding is rendered invalid, and $S_a^b(t)$ no longer behaves in a sensible manner. Without the Markov assumption, $S_a^b(dt + t)$ is no longer a superposition of $S_k^b(t)$.

Calculation of survival curves when dealing with time varying parameters is possible however. To do this, we expand the one-dimensional $S_a^b(t)$ to the two-dimensional $S_a^b(t, \rho)$. We define $S_a^b(t, \rho)$ to be the probability that an infected traveler, initially observed at position a at time ρ, has no decedents arriving at b before time t. By definition $S_a^b(t, \rho) = 1$ whenever $t \leqslant \rho$, and $S_b^b(t, t) = 0$. Using similar arguments to Eq. (3) we find that $S_a^b(t, \rho)$ is a superposition of $S_a^b(t, \rho + d\rho)$. Taking limits and converting into a differential equation we find

$$-\frac{\partial S_a^b(t, \rho)}{\partial \rho} = T(\rho)S_a^b(t, \rho) + \left[\beta(\rho) - \alpha(\rho)S_a^b(t, \rho)\right]$$
$$\times \left[1 - S_a^b(t, \rho)\right]. \tag{A1}$$

Combining Eq. (A1) with the initial conditions given at $t = \rho$, it is possible to determine $S_a^b(t, 0)$ (the survival curve of interest) for any given time t by solving a simple ODE in the ρ direction (see Fig. 11, left panels). This approach accounts for time-varying parameters whenever α, β, γ, and T vary independently from the epidemic course. In the case where parameters are dependent on epidemic spread (for example, border closures in response to observed spread of disease), more complicated methods are needed. Because Eq. (A1) must be solved independently for each t this approach is more computationally costly than the approach taken for Markov systems.

APPENDIX B: GALLERY

Here we give a variety of figures, exploring the possible behavior of arrival time distributions in a variety of networks and parameter regimes. In Figs. 12 and 13 we examine a variety of scale free networks, varying network size, and parameter values. In Fig. 14 we demonstrate the robustness of Eq. (5) to unusual parameter values by consider the (unrealistic) case of a fast-moving traveler who is noninfectious. We observing tight agreement between analytic predictions and observed
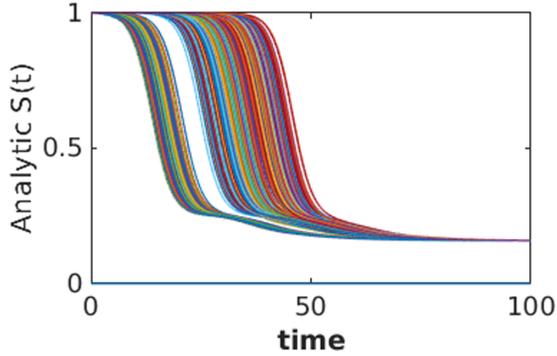
FIG. 13. Survival curves as calculated using Eq. (5), for a larger scale-free network. $N = 3000$, $\gamma = 0.001$, $\alpha = 0.5 \pm 0.11$, $\beta = 0.1 \pm 0.03$. Here we present only analytic results, as simulations are prohibitively expensive to run.

survival time in simulations. Finally, in Fig. 15 we consider arrival times for travelers on a 30-by-30 mesh.

In all cases where simulation is feasible, we observe strong agreement in arrival probabilities between analytic and simulation based results, for all t values sampled.

APPENDIX C: CONSTRUCTION OF ARBITRARY SURVIVAL CURVES

One of the more concerning results of Eq. (5) is the implication that it is possible to construct networks with arbitrary survival curves. In this section we sketch a method for constructing such a network. The purpose of this exercise is to demonstrate that for arbitrary networks, the probability distribution of arrival times may have arbitrary complexity, and hence that any analytically tractable results require further assumptions to be made on α, β, and T. More detail would be needed in order to make the argument rigorous, the work here is only intended as a proof of concept.
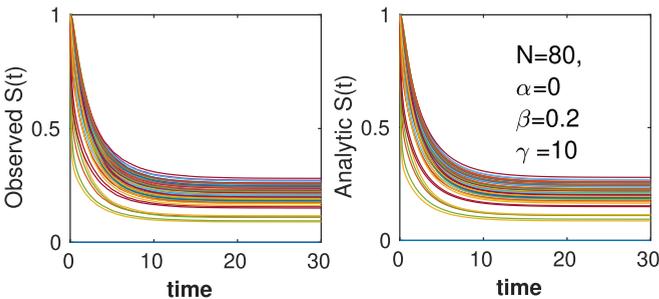

FIG. 14. The analytic survival curve gives valid predictions even far from the usual parameter regime; here we consider a traveler on a random network with, $\gamma = 10$, $\beta = 0.2$, $\alpha = 0$; a fast moving traveler that eventually decays, but does not duplicate. Arrival time distributions as observed in 15 000 simulations perfectly match survival probabilities as calculated numerically using Eq. (5). Unlike the $\gamma \ll 1$ domain, where survival probabilities decay approximately logistically, here we observe roughly exponential decay. Also in contrast to the $\gamma \ll 1$ case, different locations plateau at different levels, representing different probabilities that the traveler will *ever* arrive.
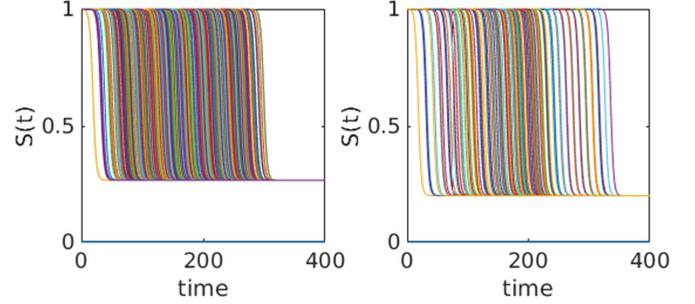

FIG. 15. (Left) Analytic survival time curves, for a 30×30 lattice, with periodic boundary conditions. $N = 900$, $\gamma = 0.001$, and epidemic parameters vary by node: $\alpha = 0.5 \pm 0.11$, $\beta = 0.1 \pm 0.03$. (Right) Analytic survival curves for 20×20 lattice, $N = 400$, $\gamma = 0.001$, epidemic parameters constant across all nodes, $\alpha = 0.5$, $\beta = 0.1$. In both cases we observe "logistic" decay from 1. Mean arrival time scales with distance from the initial site of infection. When epidemic parameters are homogeneous, symmetry results in many survival curves overlapping perfectly (the survival curve for "one step north and one step east" is the same as "one step south, one step west"). For variable parameter values, this symmetry is broken, and we see a smooth distribution of arrival times.

Suppose we are given some function $f(t)$ such that $f(0) = 1$, $f'(t) \leqslant 0$, $f(t) \geqslant 0$. We assume that $f'(t)$ is well defined for all time. Our goal is to create a network such that $S_0^N(t)$ approximates $f(t)$ as closely as possible. Consider a simple "diamond" network as depicted in Fig. 16. In this network, node zero has a directed edge connecting it to nodes 1 to $N - 1$, and nodes 1 to $N - 1$ have a single outgoing edge, directed towards the "final" node N. We assume $\beta_k = 0$ for all intermediary nodes. We take a travel rate $\gamma = 1$.

For these "intermediate" nodes, the survival time is governed by the equation

$$\dot{S}_k^N(t) = -T_{k,N} S_k^N(t) - \alpha_k S_k^N(t) + \alpha_k \left[S_k^N(t)\right]^2, \quad (C1)$$

which permits solutions of the form

$$S_k^N(t) = \frac{\alpha_k + T_{k,N}}{\alpha_k + (\alpha_k + 2T_{k,N})e^{(\alpha_k + T_{k,N})(t - \mu_k)}}. \quad (C2)$$

Here μ_k is selected such that $(\alpha_k + 2T_{k,N})e^{-\mu_k(\alpha_k + T_{k,N})} = T_{k,N}$. Because both α_k and $T_{k,N}$ are free parameters, it is possible to select them so as to construct a logistic function with arbitrary mean and scale parameters. This allows us to approximate arbitrary step functions.

We now wish to select the transition rates $T_{0,k}$ such that $S_0^N(t) \approx f(t)$. We select $\alpha_0 = 0$ such that the initial infection at node zero does not replicate, and instead simply 'jumps' to one of our $N - 1$ intermediary nodes, or recovers. $S_0^N(t)$ is governed by the equation

$$\dot{S}_0^N(t) = \sum T_{0,k}\left[S_k^N(t) - S_0^N(t)\right]\beta_k\left[1 - S_0^N(t)\right]. \quad (C3)$$

If we select $T_{0,k}$ and β_k very large, then our initial infection will linger in the starting node for only a short time, and

$$S_0^N(t) \approx \frac{\sum T_{0,k} S_k^N(t) + \beta_k}{\sum T_{0,k} + \beta_k}, \quad (C4)$$
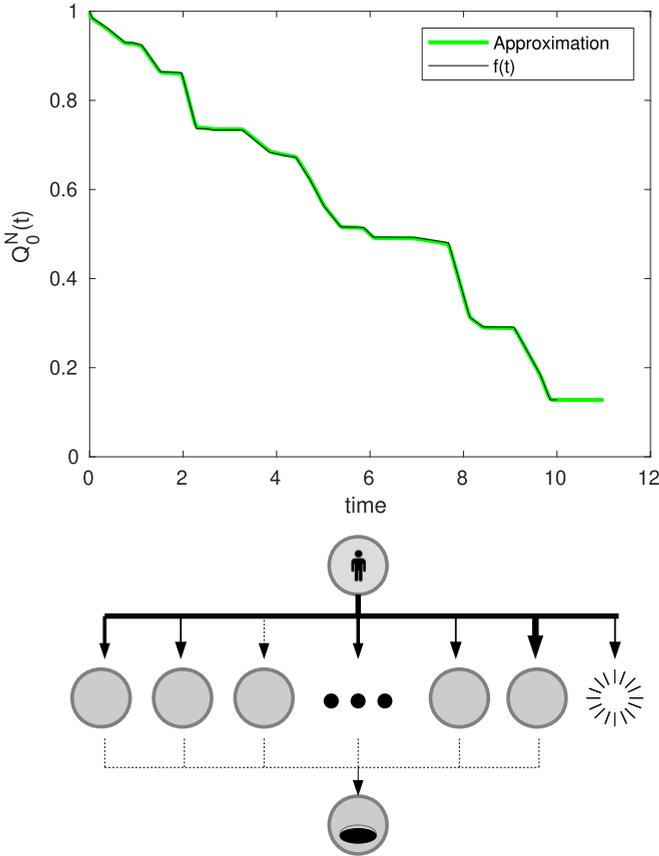
FIG. 16. (Top) We generate a random piecewise linear decreasing function, and then approximate it using the survival curve of a network. Here we use *analytic* expressions for our stepwise function as given in Eq. (C2). Direct integration of Eq. (5) runs into numerical difficulties due to the exceptionally small $T_{k,N}$ involved. (Bottom) A schematic diagram of the diamond network used to match our randomly generated survival curve. Infection rate in our central layer, along with transport rate to the "target" node is selected such that $Q_k^N(t)$ for each of our intermediary nodes approximates a step function, with a wide array of "step times." Transport rates from our starting node to intermediate nodes are selected such that $Q_0^N(t)$ is (approximately) a superposition of these many step functions, with appropriate weight given to each such that $|Q_0^N(t) - f(t)|$ is small.

that is to say, $S_0^N(t)$ is a sum of step functions, and a constant. Transition parameters are selected such that $\beta_k/(\sum T_{0,k} + \beta_k)$ is equal to $f(\infty)$ and $T_{0,k}/(\sum T_{0,k} + \beta_k) = f[(\mu_k + \mu_{k+1})/2] - f[(\mu_k + \mu_{k-1})/2]$. In this way $f(t)$ can be approximated using a series of (approximate) step functions. A demonstration of this is given in Fig. 16. More sophisticated algorithms would presumably be able to approximate f using fewer nodes, but for our purposes this simple approach will suffice.

APPENDIX D: SURVIVAL CURVES FOR NONZERO β

Throughout the main text, we frequently made use of the simplifying assumption $\beta = 0$; an unrealistic assumption in real-world context. Suppose we wish to consider nonzero β, in the context of global epidemic spread. We assume, as previously, that travel is rare, and also that we interest ourselves

only in diseases such that $R_0 > 1$, hence $\gamma \ll \beta < \alpha$. This leads to the equation

$$\dot{S}_a^b(t) \approx [\beta_a - \alpha_a S_a^b(t)][1 - S_a^b(t)], \tag{D1}$$

and hence,

$$S_a^b(t) \approx \frac{1 + \beta_a e^{(\alpha_a - \beta_a)(t - \mu_a^b)}}{1 + \alpha_a e^{(\alpha_a - \beta_a)(t - \mu_a^b)}}. \tag{D2}$$

Hence, in the limit of small γ, the arrival time distribution can be approximated using a logistic function. As in the main text μ_a^b is some mean arrival time constant determined by a, b, and γT. In the limit $t \to \infty$ $S_a^b(t) \to \beta_a/\alpha_a$; the probability that $\tau_a^b = \infty$ is precisely the probability of early epidemic extinction in our branching process. Unfortunately, this nonzero probability of "infinite" arrival time leads to $E(\tau_a^b) = \infty$, not an especially informative result. It is thus useful to instead consider the arrival time *conditional* on $\tau_a^b < \infty$:

$$S_a^b(t) = \frac{S_a^b(t) - S_a^b(\infty)}{S_a^b(0) - S_a^b(\infty)}, \tag{D3}$$

$$\approx \frac{1}{1 + e^{(\alpha_a - \beta_a)(t - \mu_b)}}. \tag{D4}$$

At this stage we have an "ideal" logistic distribution [29], with mean value μ_a^b (currently unknown) and scale parameter $1/(\alpha_a - \beta_a)$. Hence, when making predictions concerning real-world epidemics that have not gone extinct in their early stages, it is useful to remap our variables such that we imagine $\beta_k = 0$ and α_k is the *net* grow rate of our disease (formerly labeled $\alpha_k - \beta_k$).

APPENDIX E: COMPARISON TO AND IMPROVEMENT ON PAST EXPONENTIAL GROWTH METHODS

In this Appendix we will discuss the Matrix Exponential method, as used by Chen *et al.* [15] (referred to by them as "linear spreading theory"). We begin by introducing the method itself, and its connection to the Branching Process approach studied in the main text. We then explain some of the computational details necessary for fast calculation, and the circumstances under which the approximation is and is not accurate.

Let us begin by describing the method itself. Suppose, at time $t = 0$, we have a population of infected individuals $\mathbf{P}(0)$ with $P_k(0)$ infected individuals located at node k and none elsewhere. These individuals travel between nodes at transport rate γT, and infect others at some rate α. This rate is assumed to be constant; we assume that the total population is large enough such that population constraints are not relevant on the timescale we are interested in. Under this unbound growth assumption, the *expected* number of infectious individuals if governed by the linear equation

$$\dot{\mathbf{P}}(t) = \gamma T^\top \mathbf{P}(t) + \alpha \mathbf{P}(t). \tag{E1}$$

This equation permits solutions of the form

$$\mathbf{P}(t) = \exp[(\gamma T^\top + \alpha I)t]\mathbf{P}(0). \tag{E2}$$

Here, we make use of the matrix exponential; that is to say $\exp M = I + M + M^2/2! + M^3/3! + \cdots$. Here we use the transposed transition matrix, $T^\top$, because here we consider

the flow of infected individuals *forward* through the network, as opposed to the flow of arrival probability backwards through the network (as we do in the main text).

Suppose we start with an initial population of 1 in node a. If we wish calculate the deterministic time when the *expected* population in node b reaches 1, we must solve $\mathbf{P}_b(t) = 1$. Written slightly differently,

$$1 = \delta_b^\top \exp[(\gamma T^\top + \alpha I)t]\delta_a, \qquad (E3)$$

where δ_a, δ_b are indicator vectors, equal to 1 at index a and b respectively, and equal to zero elsewhere. This equation is equivalent to Eq. (7) of Chen *et al.* [15]. Here we have reached the equation via a direct appeal to unbound population growth, while Chen *et al.* instead describe their formalism as a linearization of a full SIR model.

Equation (E3) can also be reached via an appeal to branching processes. By making the change of variables $S \to Q$ as we do in the main text, suitable approximations lead to

$$Q_a^b(t) \approx e^{(\gamma T + \alpha I)t}Q_a^b(0), \qquad (E4)$$

i.e., Eq. (10). The expected arrival time at node a is precisely the time when $Q_a^b(t) = 1$, that is,

$$1 = \delta_a^\top e^{(\gamma T + \alpha I)t}\delta_b. \qquad (E5)$$

The two methods give the same results (under the influence of suitably powerful approximations).

Before moving on to discuss the accuracy of these results, let us first discuss some computational difficulties. Equation (E3) is transcendental and can thus only be solved approximately, using some suitable numerical scheme. In their paper, Chen *et al.* approach this difficulty by identifying the minimal number of steps d required to get from a to b, and truncating the polynomial expansion of their matrix exponential after this many steps:

$$1 = \delta_b^\top e^{(\gamma T^\top + \alpha I)t}\delta_a = e^{\alpha t}\sum_{k=0}^{\infty}\delta_b^\top \frac{\gamma^k T^{\top k}t^k}{k!}\delta_a \qquad (E6)$$

$$1 \approx e^{\alpha t}\delta_b^\top \frac{\gamma^d T^{\top d}t^d}{d!}\delta_a. \qquad (E7)$$

All terms before the dth term are zero, all terms after are higher powers of γ and thus assumed to be small. Chen *et al.* solve Eq. (E7) using the Lambert-W function [34,35], a nonlinear function originally constructed to solve equations of the form $ye^y = x$. In cases where we have a single "most probable" path from a to b, we can make the approximation

$$\alpha t \approx -\log\left(\delta_b^\top \frac{\gamma^d T^{\top d}t^d}{d!}\delta_a\right) \qquad (E8)$$

$$\approx -d\log\gamma - \log\left(\prod T_{i,j}\right) \qquad (E9)$$

$$\approx -\sum\log\gamma + \log T_{i,j}. \qquad (E10)$$

Here $T_{i,j}$ are the transition rates along the steps in our shortest path. Such an estimate neglects terms of magnitude $d\log(t/d)$, along with contributions from all paths except the shortest. This result mirrors those of both Gautreau *et al.* [14] as well as Brockmann *et al.* [12] (with the notable difference

being that $-\log\gamma$ is replaced by 1 when using the effective distance metric).

An alternative method to solve Eq. (E3) (not used by Chen *et al.*) is the iteration scheme:

$$1 = \delta_a^\top e^{\gamma T t_n}e^{\alpha t_{n+1}}\delta_b, \qquad (E11)$$

$$\alpha t_{n+1} = -\log\left(\delta_a^\top e^{\gamma T t_n}\delta_b\right). \qquad (E12)$$

Here we make use of the fact that αt_{n+1} varies faster than $\gamma T t_n$; hence, if we have even a moderately accurate approximation t_n, $e^{\gamma T t_n}$ will be close to the correct value $e^{\gamma T t}$. In contrast $e^{\alpha t_{n+1}}$ varies quickly: demanding t_{n+1} to solve Eq. (E12), we quickly approach true solutions of Eq. (E5).

This iteration scheme requires us to calculate large matrix exponentials repeatedly, an operation which is generically computationally expensive, but can be made significantly less costly by first computing the eigenvector decomposition $T = VDV^{-1}$. Here V is a matrix containing the eigenvectors of M, while D is a diagonal matrix containing the corresponding eigenvalues. This allows us to instead compute the matrix exponential via elementwise exponentials of the diagonal elements:

$$\alpha t_{n+1} = -\log\left[(\delta_a^\top V)e^{\gamma D t_n}(V^{-1}\delta_b)\right]. \qquad (E13)$$

Matlab code to implement this iteration scheme, approximating transport times *to* node b from all starting points is given by the following:

```
[V,D_original] =eig(T);
N=size(T,1);
b=7;
delta_b= zeros(N,1);
delta_b(b)=1;

D= gamma * diag(D_original);
Right= V\delta_b;

Tk=-ones(N,1);

for(aaa=1:N)
guessT= (-log(gamma))/alpha
deltaT=5;
Left= V(aaa,:);
while(abs(deltaT)>10^-3)
expGammaT=Left*(exp(D*guessT).*Right);
deltaT= -log(expGammaT)/alpha
end
Tk(aaa)=guessT;
end
```

The computational cost of the above code is overwhelmingly dominated by `eig(M)`, and hence the runtime scales like $O(n^3)$, similar to the Floyd-Warshall algorithm as might be used to identify shortest paths when calculating the effective distance. It is possible to calculate the time taken to arrive in each location *from* a fixed starting location a by using the transposed transition matrix, $T^\top$, rather than T.

Now, in all uses of such methods, it is important to note that the deterministic time when the *expected* population reaches 1 (as calculated using Chen *et al.*'s method), and the *expected* time when the stochastically varying population reaches one

are *not* equivalent concepts. This is best illustrated by considering the extreme case $\alpha = \beta = 0$, where we observe that $\delta_a^\top e^{\gamma T t} e^{\alpha t} \delta_b < 1$ for *all* t. Taken at face value this would erroneously imply that the arrival time has an expected value of infinity; in practice this merely illustrates the difficulties of using such an approximation for diseases with low α (such as HIV). In general, Eq. (10) is observed to be a good approximation when $\alpha \gg \gamma T$ and performs poorly when $\alpha \ll 1$.

APPENDIX F: VARIANCE IN ARRIVAL TIME

Here we provide the algebraic details for calculating variances in arrival time, assuming a starting population greater than 1.

Recall that τ_a^b denotes the arrival time at b, assuming an epidemic starts at a, and has mean μ_a^b. The survival function of $S_a^b(t) = P(\tau_a^b > t)$. In what follows, we suppress subscripts and superscripts, considering some generic τ:

$$E(\tau) = \int_0^\infty S \, dt = \mu. \tag{F1}$$

Assuming the mean value $\mu \gg 0$,

$$\dot{S} = \alpha S(1 - S), \tag{F2}$$

$$\int_0^\infty S^k - S^{k+1} \, dt = \int_0^\infty S^{k-1} \dot{S}/\alpha \, dt, \tag{F3}$$

$$\int_1^0 \hat{S}^{k-1}/\alpha \, dS = \left[\frac{S^k}{\alpha k} \right]_1^0 = -1/\alpha k, \tag{F4}$$

$$\int_0^\infty S^{k+1} \, dt = \int_0^\infty \hat{S}^k \, dt - \frac{1}{\alpha k}, \tag{F5}$$

$$\int_0^\infty S^{N+1} \, dt = \mu - \sum_{k=1}^N \frac{1}{\alpha k}. \tag{F6}$$

Hence, if the expected arrival given an initial population of 1 is μ, then the arrival time assuming $K + 1$ individuals is $\mu - \sum_1^N 1/\alpha k$. This makes sense, given that the expected time of transition form k to $k + 1$ individuals, for a branching rate of α is precisely equal to $1/\alpha k$.

Next, we need to determine the expected value of $E(\tau^2)$ for a starting population of N. For a starting population of 1, we have $E(\tau^2) = \mu^2 + \pi^2/3\alpha^2$ [29]. For larger populations:

$$\int_0^\infty 2t(S^k - S^{k+1}) \, dt = \int_0^\infty 2t S^{k-1} \dot{S}/\alpha \, dt,$$

$$S = 1/\{1 + \exp[\alpha(t - \mu)]\},$$

$$t = \log\left(\frac{1 - S}{S}\right)/\alpha + \mu,$$

$$\int_0^\infty 2t(S^k - S^{k+1}) \, dt = \frac{2}{\alpha^2} \int_1^0 \log\left(\frac{1 - S}{S}\right) S^{k-1} \, dS$$

$$= \frac{2}{\alpha^2} \frac{H_{k-1}}{k},$$

where H_k is the kth harmonic number $\sum_{i=1}^k \frac{1}{i}$ and $H_0 = 0$:

$$\frac{2}{\alpha} \int_0^\infty t S^{k-1} \, dS = \frac{2}{\alpha^2} \frac{H_{k-1}}{k} - \frac{2\mu}{\alpha k},$$

$$\int_0^\infty 2t S^{N+1} \, dt = \mu^2 + \frac{\pi^2}{3\alpha^2} + \sum_{k=1}^N \left[\frac{2}{\alpha^2} \frac{H_{k-1}}{k} - \frac{2\mu}{\alpha k} \right].$$

The variance is thus given as

$$E(\tau^2) - E(\tau)^2 = \mu^2 + \frac{\pi^2}{3s^2} + \sum_{k=1}^N \left[\frac{2}{s^2} \frac{H_{k-1}}{k} - \frac{2\mu}{sk} \right]$$

$$- \left(\mu_a^b - \sum_1^N 1/sk \right)^2 \tag{F7}$$

$$= \frac{\pi^2}{3s^2} + \sum_{k=1}^N \left[\frac{2}{s^2} \frac{H_{k-1}}{k} \right] - \left(\sum_1^N 1/sk \right)^2 \tag{F8}$$

$$= \frac{\pi^2}{3s^2} + \frac{1}{s^2} \sum_{k=1}^N \left[\frac{2H_{k-1} - H_N}{k} \right] \tag{F9}$$

$$= \frac{\pi^2}{3s^2} - \sum \frac{1}{s^2 k^2}. \tag{F10}$$

[1] J. Verne, *Around the World in Eighty Days* (Porter & Coates, Paris, 1873).

[2] L. Hufnagel, D. Brockmann, and T. Geisel, Forecast and control of epidemics in a globalized world, Proc. Natl. Acad. Sci. USA **101**, 15124 (2004).

[3] WHO, WHO|Situation updates—SARS (2003), https://www.who.int/csr/sars/archive/en/.

[4] WHO, WHO|Situation updates—Avian influenza (2009), http://www.who.int/influenza/human_animal_interface/avian_influenza/archive/en/.

[5] B. Gross, Z. Zheng, S. Liu, X. Chen, A. Sela, J. Li, D. Li, and S. Havlin, Spatio-temporal propagation of COVID-19 pandemics, arXiv:2003.08382 [physics, q-bio].

[6] WHO, Novel Coronavirus (2019-nCoV) Situation Reports (2020).

[7] B. Blasius, Power-law distribution in the number of confirmed COVID-19 cases, Chaos **30**, 093123 (2020).

[8] F. Schlosser, B. F. Maier, D. Hinrichs, A. Zachariae, and D. Brockmann, COVID-19 lockdown induces structural changes in mobility networks—Implication for mitigating disease dynamics, arXiv:2007.01583 [physics, q-bio].

[9] P. Kaluza, A. Kölzsch, M. T. Gastner, and B. Blasius, The complex network of global cargo ship movements, J. R. Soc., Interface **7**, 1093 (2010)

[10] H. Seebens, M. T. Gastner, and B. Blasius, The risk of marine bioinvasion caused by global shipping, Ecol. Lett. **16**, 782 (2013).

[11] W. V. d. Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, The GLEaMviz computational tool, a publicly available software to explore realistic epidemic

spreading scenarios at the global scale, BMC Infect. Dis. **11**, 37 (2011).

[12] D. Brockmann and D. Helbing, The hidden geometry of complex, network-driven contagion phenomena, Science **342**, 1337 (2013).

[13] F. Iannelli, A. Koher, D. Brockmann, P. Hövel, and I. M. Sokolov, Effective distances for epidemics spreading on complex networks, Phys. Rev. E **95**, 012313 (2017).

[14] A. Gautreau, A. Barrat, and M. Barthelemy, Global disease spread: Statistics and estimation of arrival times, J. Theor. Biol. **251**, 509 (2008).

[15] L. M. Chen, M. Holzer, and A. Shapiro, Estimating epidemic arrival times using linear spreading theory, Chaos (Woodbury, NY) **28**, 013105 (2018).

[16] M. Kimmel and D. Axelrod, *Branching Processes in Biology*, 2nd ed. Interdisciplinary Applied Mathematics (Springer-Verlag, New York, 2015), Vol. 19.

[17] H. E. Schaffer, Survival of mutant genes as a branching process, in *Mathematical Topics in Population Genetics*, edited by K.-I. Kojima, Biomathematics, Vol. 1 (Springer, Berlin, 1970), pp. 317–336.

[18] M. V. González, I. M. Puerto, R. Martínez, M. Molina, M. Mota, and A. Ramos, editors, *Workshop on Branching Processes and Their Applications*, Lecture Notes in Statistics—Proceedings, Vol. 197 (Springer, Berlin, 2010).

[19] V. Colizza and A. Vespignani, Invasion Threshold in Heterogeneous Metapopulation Networks, Phys. Rev. Lett. **99**, 148701 (2007).

[20] H. W. Watson and F. Galton, On the probability of the extinction of families, J. Anthropol. Inst. G. B. Irel. **4**, 138 (1875).

[21] J. H. Goldie and A. J. Coldman, A mathematic model for relating the drug sensitivity of tumors to their spontaneous mutation rate, Cancer Treat. Rep. **63**, 1727 (1979).

[22] A. J. Coldman and J. H. Goldie, A stochastic model for the origin and treatment of tumors containing drug-resistant cells, Bull. Math. Biol. **48**, 279 (1986).

[23] L. Giuggioli, Exact spatiotemporal dynamics of confined lattice random walks in arbitrary dimensions: A century after Smoluchowski and Pólya, Phys. Rev. X **10**, 021045 (2020).

[24] A. Godec and R. Metzler, First passage time distribution in heterogeneity controlled kinetics: Going beyond the mean first passage time, Sci. Rep. **6**, 20349 (2016).

[25] L. F. Shampine and M. W. Reichelt, The MATLAB ODE suite, SIAM J. Sci. Comput. **18**, 1 (1997).

[26] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions, J. Phys. Chem. **81**, 2340 (1977).

[27] D. T. Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems, J. Chem. Phys. **115**, 1716 (2001).

[28] With some individuals traveling significantly more often, and many others never flying at all.

[29] N. Balakrishnan, *Handbook of the Logistic Distribution* (CRC Press, Boca Raton, FL, 2013).

[30] J. C. Miller, A primer on the use of probability generating functions in infectious disease modeling, Infect. Dis. Model. **3**, 192 (2018).

[31] F. Coerts, Network analysis as a tool to improve infection prevention & outbreak management, Master's thesis, University of Groningen, Groningen (2018).

[32] A. Jamieson-Lane, alastair-JL/EpidemicSpread (2020), original-date 2020-06-22T23:17:54Z, https://github.com/alastair-JL/EpidemicSpread.

[33] M. George, *B-A Scale-Free Network Generation and Visualization* (2020), https://www.mathworks.com/matlabcentral/fileexchange/11947-b-a-scale-free-network-generation-and-visualization.

[34] L. Euler, De serie Lambertina Plurimisque eius insignibus proprietatibus, Acta Acad. Scient. Petropol. **1779**, 29 (1783), https://scholarlycommons.pacific.edu/euler-works/532/.

[35] E. W. Weisstein, *Lambert W-function* (Wolfram Research, Champaign, IL, 2020).