

Research



Cite this article: Park SW, Champredon D, Dushoff J. 2020 Inferring generation-interval distributions from contact-tracing data. *J. R. Soc. Interface* **17**: 20190719. <http://dx.doi.org/10.1098/rsif.2019.0719>

Received: 17 October 2019
Accepted: 4 June 2020

Subject Category:
Life Sciences—Mathematics interface

Subject Areas:
biomathematics, computational biology

Keywords:
infectious disease modelling, generation interval, basic reproductive number, population structure, contact tracing

Authors for correspondence:
Sang Woo Park
e-mail: swp2@princeton.edu
Jonathan Dushoff
e-mail: dushoff@mcmaster.ca

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.5025770>.

Inferring generation-interval distributions from contact-tracing data

Sang Woo Park^{1,2}, David Champredon^{3,5} and Jonathan Dushoff^{3,4}

¹Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA

²Department of Mathematics and Statistics, ³Department of Biology, and ⁴Michael G. DeGroot Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada

⁵Department of Pathology and Laboratory Medicine, University of Western Ontario, London, ON, Canada

id SWP, 0000-0003-2202-3361; DC, 0000-0002-7090-8757; JD, 0000-0003-0506-4794

Generation intervals, defined as the time between when an individual is infected and when that individual infects another person, link two key quantities that describe an epidemic: the initial reproductive number, $\mathcal{R}_{\text{initial}}$, and the initial rate of exponential growth, r . Generation intervals can be measured through contact tracing by identifying who infected whom. We study how realized intervals differ from ‘intrinsic’ intervals that describe individual-level infectiousness and identify both spatial and temporal effects, including truncating (due to observation time), and the effects of susceptible depletion at various spatial scales. Early in an epidemic, we expect the variation in the realized generation intervals to be mainly driven by truncation and by the population structure near the source of disease spread; we predict that correcting realized intervals for the effect of temporal truncation but *not* for spatial effects will provide the initial forward generation-interval distribution, which is spatially informed and correctly links r and $\mathcal{R}_{\text{initial}}$. We develop and test statistical methods for temporal corrections of generation intervals, and confirm our prediction using individual-based simulations on an empirical network.

1. Introduction

An epidemic can be described by the exponential growth rate, r , and the reproductive number, \mathcal{R} . The reproductive number is defined as the average number of secondary cases arising from a primary case; its value in a fully susceptible population, also known as the basic reproductive number \mathcal{R}_0 , is of particular interest as it provides information about the final size of an epidemic [1,2] as well as the endemicity level [3–5]. However, estimating the reproductive number directly from disease life history requires detailed knowledge, which is not often available, particularly early in an outbreak [6]. Instead, the reproductive number is often indirectly estimated from the exponential growth rate, which can be estimated from incidence data [7–11]. These two quantities are linked by generation-interval distributions [12–16].

At the individual level, a generation *interval* is defined as the time between when a person becomes infected and when that person infects another person [13]. While this definition is widely used in the literature, it is not directly related to a population-level *distribution*. There are important distinctions to be made when defining generation-interval distributions at the population level. The *intrinsic* generation-interval distribution describes the expected time distribution of infectious contacts made by a primary case [17]. On the other hand, *realized* generation-interval distributions describe the time between actual infection events over the course of an epidemic. Since some infectious contacts will be made with non-susceptible people, and thus not result in infection, realized distributions can differ systematically from the intrinsic distribution.

The shape of the realized generation-interval distribution depends on the reference time and perspective [17–21]. When an epidemic is growing

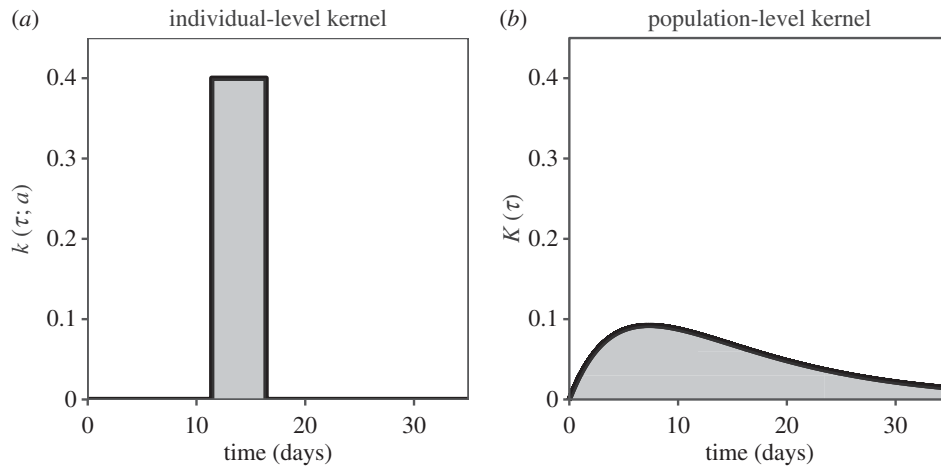


Figure 1. Comparison of individual- and population-level kernels. (a) An individual-level kernel of an infected individual with a latent period of 11.4 days followed by an infectious period of 5 days; this represents an individual realization of a random process. (b) A population-level kernel of infected individuals with latent and infectious periods exponentially distributed with means of 11.4 days and 5 days, respectively; this represents a population average of a random process. Shaded areas under the curves are equal to individual- and population-level reproductive numbers, both of which are set to 2 in this example. Parameters are chosen to reflect the West African Ebola outbreak [25].

exponentially, as often occurs near the beginning of an outbreak, the number of newly infected individuals will be large relative to the number infected earlier on. A susceptible individual is thus relatively more likely to be infected by a newly infected individual. Thus, ‘backward’ generation intervals, which look at a cohort of infectees and ask when their infectors were infected, will be shorter on average than intrinsic generation intervals—the converse is true when an epidemic is subsiding [17,19,21]. Likewise, we can define ‘forward’ generation intervals, which look at a cohort of infectors and ask when their infectees were infected. Mean forward generation intervals tend to decrease over the course of an epidemic as a result of susceptible depletion [17–20].

Realized generation intervals are also affected by spatial structure. In a population that does not mix homogeneously, susceptibility will tend to decrease more quickly in the neighbourhood of infected individuals than in the general population. This means that infectious contacts made late in an individual’s infection are more likely to be ineffective because of contacts that were made earlier (because the contactee may have been infected already). As a result, realized generation intervals (from the perspective of an infector) will typically have a shorter mean than the intrinsic generation-interval distribution in a non-homogeneous population. This perspective allows us to reinterpret the finding of [22] that, given an intrinsic generation interval and an observed growth rate, the reproductive number on various network structures is always smaller than would be predicted from homogeneous mixing.

In practice, realized generation intervals are often difficult to measure for many diseases, because it is difficult to observe when individuals become infected; in most cases, observable events are clinical (e.g. onset of symptoms). There are some exceptions: for example, generation intervals are commonly measured directly through contact tracing for canine rabies, where infection events are bites [23]. Intervals between observed disease progression events (commonly, onset of signs or symptoms) are called serial intervals [13]. Serial intervals are in many ways similar to generation intervals, but there are also complexities in their use [21]. We will not address these complexities here.

While an epidemic is ongoing, realized generation intervals, at least in theory, can be measured by identifying who

infected whom and when, and aggregated to form a single distribution. We typically want to try to make inference based on this aggregated distribution—that is, on all available data that have been gathered since the beginning of an epidemic. These aggregated measurements are ‘truncated’ because we do not know what happens after the time of last observation. The distribution of these truncated intervals is similar to backward intervals during the exponential growth phase: there is a bias towards over-sampling shorter intervals, which are more likely to have concluded in time to be observed. We therefore predict that removing the truncation bias from aggregated generation intervals early in an epidemic will yield the initial forward generation-interval distribution, which contains information about the population structure and allows us to correctly infer the initial reproductive number from the initial exponential growth rate.

In this study, we explore spatio-temporal variation in realized generation intervals. We extend previous frameworks to investigate how aggregated generation-interval distributions change over time. We classify spatial effects on realized generation intervals into three levels (egocentric, local and global) and discuss how these affect realized generation-interval distributions. Finally, we compare two methods for accounting for temporal bias and test our prediction using individual-based simulations.

2. Intrinsic generation-interval distributions

Generation-interval distributions are often considered as population averages, but we can distinguish population-level distributions from individual-level distributions [13,24]; making this distinction clear will be particularly useful when we discuss spatial components later (figure 1). An individual-level intrinsic infection kernel $k(\tau, a)$ describes the rate at which an infected individual with ‘aspect’ a makes ‘infectious contacts’ (contacts which will cause infection if the contactee is susceptible). Individual aspects may represent variation in the course of infection (e.g. duration of latent and infectious periods) and the level of infectiousness, which can depend both on biological infectiousness and on contact patterns. Hereafter, we use t, s to represent calendar time and τ, x to represent time since infection.

Assuming that the individual properties are independent of risk of infection, the population-level kernel is given by integrating over these individual variations,

$$K(\tau) = \int k(\tau, a)f(a) da, \quad (2.1)$$

where $f(a)$ represents a probability density over a (possibly multi-dimensional) aspect space. The population-level kernel describes the rate at which infectious contacts are made by an infected individual, on *average*.

Assuming that a population mixes homogeneously, we can write

$$K(\tau) = \mathcal{R}_0 g(\tau), \quad (2.2)$$

where $\mathcal{R}_0 = \int K(\tau) d\tau$ is the basic reproductive number (the expected number of secondary cases caused by a randomly chosen infectious individual in a fully susceptible population [1]) and $g(\tau)$ is the expected time distribution of infectious contacts made by a primary case (the intrinsic generation-interval distribution [17]). If the proportion of susceptibles contacted is not changing (e.g. in a homogeneously mixed population at the endemic equilibrium, or when the number of cases is vanishingly small), $g(\tau)$ also describes the realized (forward) generation intervals.

In a homogeneously mixing population, current disease incidence at time t , $i(t)$, is the product of the current infectiousness of individuals infected in the past and the current proportion of the population susceptible, $S(t)$,

$$i(t) = S(t) \int K(\tau)i(t-\tau) d\tau = \mathcal{R}_0 S(t) \int g(\tau)i(t-\tau) d\tau. \quad (2.3)$$

This model, referred to as the renewal equation, can describe a wide range of epidemic models [14,15,26–30]. Over a period of time where the susceptible proportion remains approximately constant ($S(t) \approx S(0)$), we would expect approximately exponential growth in incidence $i(t)$; assuming $i(t) = i(0) \exp(rt)$ yields the Euler–Lotka equation [31], which provides a direct link between the initial exponential growth rate r and the initial reproductive number $\mathcal{R}_{\text{initial}} = \mathcal{R}_0 S(0)$,

$$\frac{1}{\mathcal{R}_{\text{initial}}} = \int g(\tau) \exp(-r\tau) d\tau. \quad (2.4)$$

Under the homogeneous mixing assumption, the intrinsic generation-interval distribution $g(\tau)$ provides the correct link between r and $\mathcal{R}_{\text{initial}}$.

3. Realized generation-interval distributions across time

Realized generation intervals can be measured either forward (from the perspective of a cohort of infectors) or backward (from the perspective of a cohort of infectees) in time [17,21]. The forward generation-interval distribution $f_i(\tau)$ describes the infection time of *infectees* caused by a cohort of infectors who were infected at time t . Similarly, the backward generation-interval distribution $b_i(\tau)$ describes the infection time of *infectors* for a cohort of infectees who were infected at time t . For a single infector–infectee pair, both backward and forward *measurements* should give the identical generation interval. Therefore, the density of new infections occurring at time $t+\tau$ caused by individuals infected at time t can be expressed in terms of both the

forward and backward generation-interval distributions,

$$\mathcal{R}_c(t)i(t)f_i(\tau) = i(t+\tau)b_{t+\tau}(\tau). \quad (3.1)$$

Here, $\mathcal{R}_c(t)$ represents the case reproductive number, which is defined as the average number of secondary cases caused by a primary case infected at time t over the course of their infection [32].

In a homogeneously mixing population, the forward and backward generation-interval distributions can be calculated exactly. The density of new infections occurring at time $t+\tau$ caused by infectors who were infected at time t is given by

$$i_t(t+\tau) = \mathcal{R}_0 i(t)g(\tau)S(t+\tau). \quad (3.2)$$

As shown in [17], the forward generation-interval distribution, $f_i(\tau)$, is proportional to $i_t(t+\tau)$,

$$f_i(\tau) = \frac{g(\tau)S(t+\tau)}{\int_0^\infty g(x)S(t+x) dx}. \quad (3.3)$$

In this case, the initial forward generation-interval distribution $f_0(\tau)$ during the exponential growth phase (when $S(t) \approx S(0)$) is equivalent to the intrinsic generation-interval distribution $g(\tau)$ and, therefore, provides the correct link between r and $\mathcal{R}_{\text{initial}}$. Likewise, the density of new infections occurring at time t caused by infectors who were infected at time $t-\tau$ is given by

$$i_{t-\tau}(t) = \mathcal{R}_0 i(t-\tau)g(\tau)S(t). \quad (3.4)$$

The backward generation-interval distribution, $b_i(\tau)$, is proportional to $i_{t-\tau}(t)$,

$$b_i(\tau) = \frac{i(t-\tau)g(\tau)}{\int_0^\infty i(t-x)g(x) dx}. \quad (3.5)$$

Substituting $\mathcal{R}_c(t) = \mathcal{R}_0 \int_0^\infty g(\tau)S(t+\tau) d\tau$ confirms that equation (3.1) holds for this model.

During an ongoing epidemic, generation intervals cannot be measured for infection events that have not happened yet. This effect is called ‘right truncation’. Therefore, even if we aggregate all realized generation intervals by identifying who infected whom through contact tracing (assuming that infection events are observable), their mean will be shorter than the mean intrinsic generation interval. The aggregated generation-interval distribution, by definition, is a weighted average of backward generation-interval distributions (weighted by incidence) up until calendar time t ,

$$a_i(\tau) \propto \int_{-\infty}^t i(s)b_s(\tau) ds. \quad (3.6)$$

The aggregated generation-interval distribution can also be expressed equivalently in terms of forward generation-interval distributions (weighted by incidence and case reproductive number),

$$a_i(\tau) \propto \int_{-\infty}^{t-\tau} \mathcal{R}_c(s)i(s)f_s(\tau) ds. \quad (3.7)$$

Equation (3.1) confirms that both expressions are identical.

For a single outbreak, the mean aggregated generation interval will always be shorter than the mean intrinsic generation interval (figure 2). There are two reasons for this phenomenon. First, longer generation intervals are more likely than short intervals to be missed because of right truncation. In particular, if we assume that the initial forward generation-interval distribution remains constant ($f_i(\tau) \approx f_0(\tau)$) when an epidemic is growing exponentially ($i(t) \approx i(0)$)

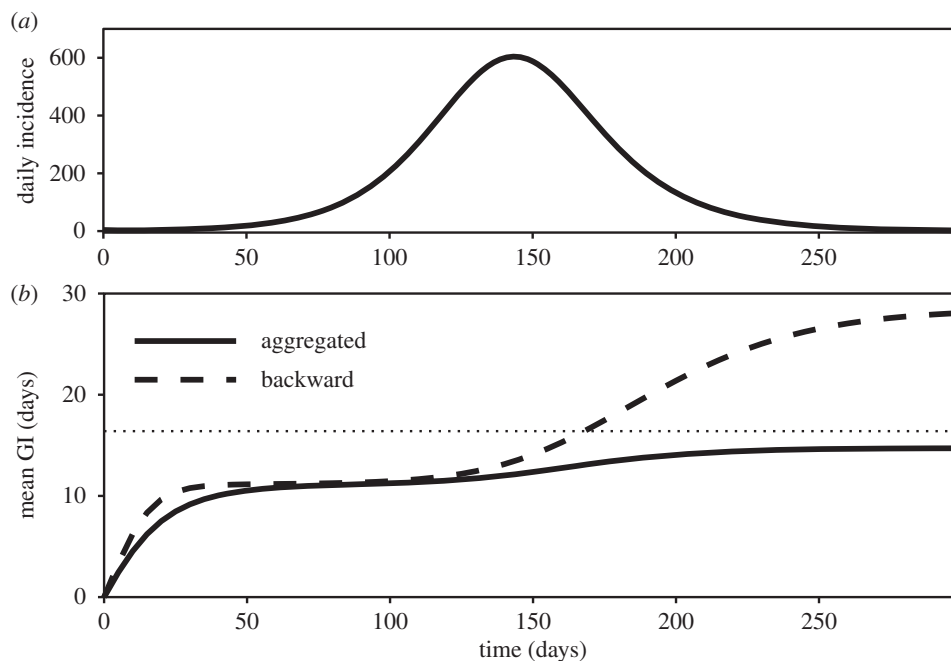


Figure 2. Temporal variation in the mean backward and aggregated generation interval. A deterministic susceptible–exposed–infectious–recovered (SEIR) model is simulated using Ebola-like parameters [25]: mean latent period $1/\sigma = 11.4$ days, mean infectious period $1/\gamma = 5$ days and the basic reproductive number $\mathcal{R}_0 = 2$. The mean backward and aggregated generation intervals are calculated over the course of an epidemic. The dotted horizontal line represents the mean intrinsic generation interval.

$\exp(rt)$ and $\mathcal{R}_c(t) \approx \mathcal{R}_c(0)$, the initial aggregated (or backward) generation-interval distribution is just the initial forward generation-interval distribution discounted by the rate of exponential growth [21],

$$a_0(\tau) = b_0(\tau) \propto f_0(\tau) \exp(-r\tau). \quad (3.8)$$

A deterministic simulation confirms that the aggregated generation-interval distribution has the same mean as the backward generation-interval distribution during this period (figure 2). Second, the decreasing number of susceptibles over the course of an epidemic makes long infectious contacts less likely to result in infection [17]. Overall, we therefore expect naively using the aggregated generation-interval distribution to underestimate the initial reproductive number.

4. Realized generation-interval distributions across space

The effects of spatial structure on realized generation intervals can be understood in terms of effect of multiple contacts. Infected individuals may contact the same susceptible individual multiple times, but only the first infectious contact gives rise to infection in a given individual (after this, they are no longer susceptible). Therefore, we expect realized generation intervals from an individual in a spatially structured population to have a smaller mean than their mean intrinsic generation interval. To explore the effects of spatial structure on realized generation intervals, we relax our assumption that the population is homogeneous. Instead, we assume that a disease spreads on a network; infected individuals contact their ‘acquaintances’ at random, but ‘acquaintanceships’ are predetermined by the network structure before the beginning of an epidemic [22].

We first consider the infection process from an ‘egocentric’ point of view, taking into account infectious

contacts made by a single infector. We define the egocentric kernel as the rate at which secondary infections are realized by a single primary case with aspect a in the absence of other infectors,

$$\hat{k}(\tau; a) = k(\tau; a) \exp\left(-\delta(a) \int_0^\tau k(x; a) dx\right), \quad (4.1)$$

where $k(\tau; a)$ is the individual-level intrinsic kernel and $e^{-\delta(a) \int_0^\tau k(x; a) dx}$ is the probability that a susceptible acquaintance has not yet been contacted by the focal individual. The dilution term, $\delta(a)$, models how contacts are distributed among the acquaintances.

Throughout this paper, we assume that there is a constant per-pair contact rate λ [22]. In this case, the intrinsic infectiousness of an individual $\mathcal{R}(a) = \int k(\tau; a) d\tau$ is the product of the number of acquaintances $N(a)$, which can vary among individuals, the contact rate λ and the duration of infectious period; the dilution term is equal to the reciprocal of the number of acquaintances: $\delta(a) = 1/N(a)$. This assumption can be relaxed by allowing for asymmetry [22] or heterogeneity [33,34] in contact rates; for simplicity, we do not pursue these directions here.

The population-level egocentric kernel is found by integrating the individual-level kernel over individual variations,

$$\hat{K}(\tau) = \int \hat{k}(\tau; a) f(a) da, \quad (4.2)$$

where $f(a)$ represents a probability density over a (possibly multi-dimensional) aspect space. Essentially, the population-level egocentric kernel accounts for the probability that a susceptible individual has not been infected by the focal individual. Trapman *et al.* [22] used this same kernel (also assuming a constant per-pair contact rate) to study the effect of network structure on the estimate of the basic reproductive number. The population-level egocentric generation-

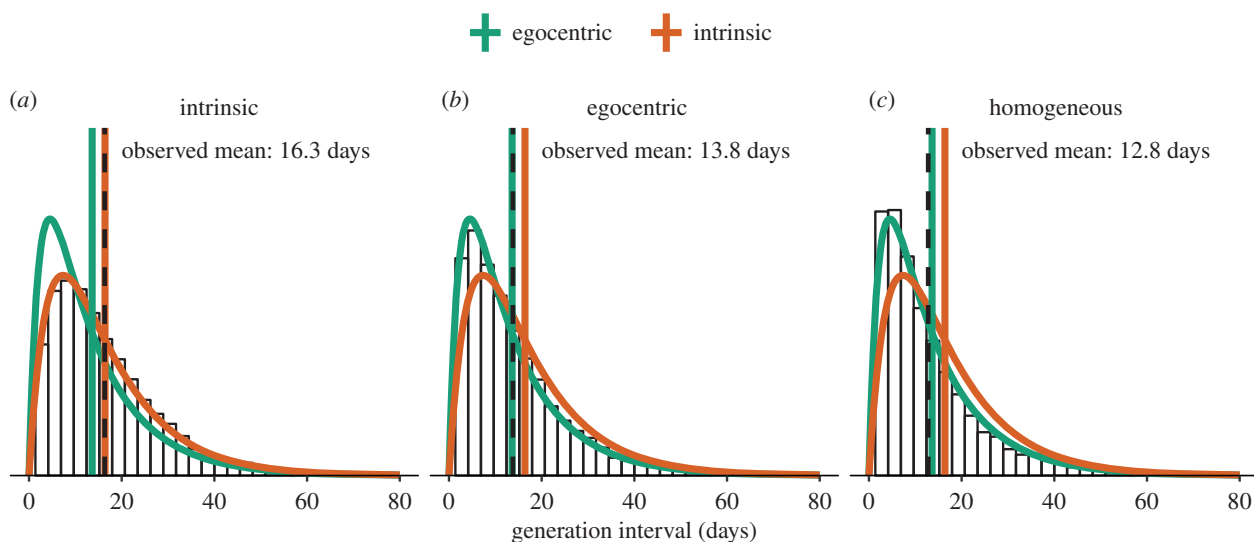


Figure 3. Spatial effects on realized generation intervals. Theoretical distributions and means are shown in colour (and are the same in each panel, for reference). Simulated distributions and means are shown in black. (a) The intrinsic generation-interval distribution corresponds to all infectious contacts by a focal individual, regardless of whether the contact results in infection. (b) The egocentric generation-interval distribution corresponds to the distribution of all infectious contacts by the focal individual with susceptible individuals, in the case where the focal individual is the only possible infector (simulated on a star network). (c) Realized generation-interval distributions have a shorter mean than egocentric distributions in general, because contacts can be wasted when susceptibles become infected through other routes (simulated on a homogeneous network). All figures were generated using 5000 stochastic simulations on a network with five nodes (one infector and four susceptibles) with Ebola-like parameters [25]: mean latent period $1/\sigma = 11.4$ days and mean infectious period $1/\gamma = 5$ days. Per-pair contact rate $\lambda = 0.25 \text{ days}^{-1}$ is chosen to be sufficiently high so that the differences between generation-interval distributions are clear. Each simulation is run until all individuals are either susceptible or have recovered.

interval distribution is

$$\hat{g}(\tau) = \frac{\hat{K}(\tau)}{\int \hat{K}(x) dx}. \quad (4.3)$$

The population-level egocentric generation-interval distribution describes the distribution of times at which secondary infections are realized from an *average* infected-susceptible pair; for convenience, we will often omit ‘population level’. Finally, the initial exponential growth rate and the egocentric reproductive number are linked by the egocentric generation-interval distribution (and the Euler–Lotka equation) [22],

$$\frac{1}{\mathcal{R}_{\text{egocentric}}} = \int \hat{g}(\tau) \exp(-r\tau) d\tau. \quad (4.4)$$

As the egocentric distribution always has a shorter mean than the intrinsic distribution, $\mathcal{R}_{\text{egocentric}}$ will be smaller than $\mathcal{R}_{\text{initial}}$ estimated from the intrinsic distribution; this generation-interval-based argument provides an alternative biological interpretation for the result presented by [22].

For example, consider a susceptible–exposed–infected–recovered (SEIR) model, which assumes that latent and infectious periods are exponentially distributed. The intrinsic generation-interval distribution that corresponds to this model can be written as [30,35]

$$g(\tau) = \frac{\sigma\gamma}{\sigma - \gamma} (e^{-\gamma\tau} - e^{-\sigma\tau}), \quad (4.5)$$

where $1/\sigma$ and $1/\gamma$ are the mean latent and infectious periods, respectively. Assuming a constant per-pair contact rate of λ for any pair, we obtain the following egocentric

generation-interval distribution:

$$\hat{g}(\tau) = \frac{\sigma(\gamma + \lambda)}{\sigma - (\gamma + \lambda)} (e^{-(\gamma + \lambda)\tau} - e^{-\sigma\tau}). \quad (4.6)$$

In this case, with constant transmission rate during the infectious period, the effect of accounting for pairwise contacts is the same as an increase in the recovery rate (by the amount of the per-pair contact rate λ). Infecting a susceptible contact is analogous to recovery because the contactee cannot be infected again—the infector can no longer transmit infection even if they are infectious (effectively losing infectiousness). Therefore, the resulting egocentric generation-interval distribution is equivalent to the intrinsic generation-interval distribution with mean latent period of $1/\sigma$ and mean infectious period of $1/(\gamma + \lambda)$. In practice, directly using the egocentric distribution to link r and $\mathcal{R}_{\text{initial}}$ using the Euler–Lotka equation is unrealistic because it requires that we know the per-pair contact rate. Instead, the per-pair contact rate can be inferred from the growth rate r , assuming that mean and variance of the degree distribution of a network is known (see [22] supplementary material, §1.4.2); we briefly describe this relationship in §7.3.

This calculation can be validated by simulating stochastic infection processes on a ‘star’ network (i.e. a single infected individual at the centre connected to multiple susceptible individuals who are not connected with each other). Simulations (figure 3) confirm that in this case the distribution of *contact* times matches the intrinsic generation-interval distribution (a), while the distribution of realized generation intervals (i.e. *infection* times) matches the egocentric generation-interval distribution (b).

The egocentric generation interval (equation (4.3)) only explains some of the reduction in realized generation intervals that occurs on most networks, however. Generation

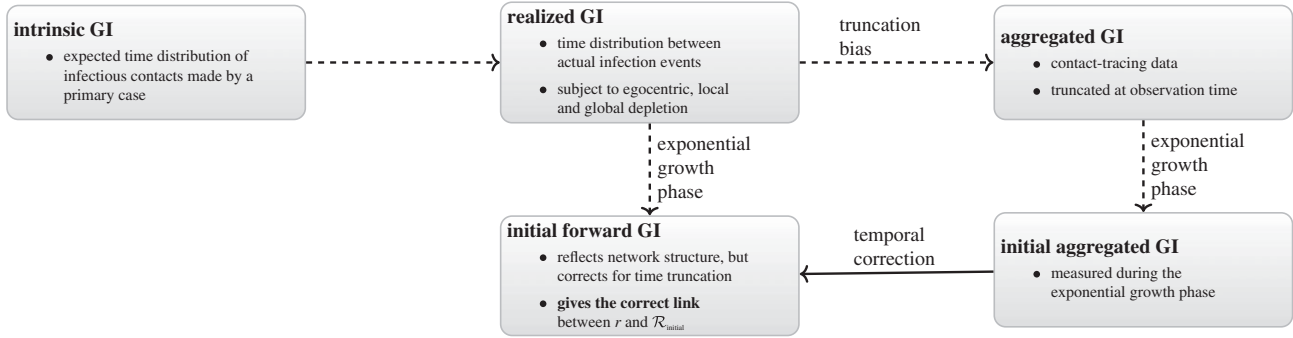


Figure 4. A summary of spatiotemporal effects on generation intervals. The intrinsic generation-interval distribution describes the expected time distribution of infectious contacts made by a primary case. The realized (forward) generation-interval distribution describes the time between actual infection events—egocentric, local and global depletion of the susceptible pool reduces the mean realized generation intervals. The aggregated generation-interval distribution reflects the contact-tracing process during an ongoing epidemic and is subject to right truncation bias. Our goal is to estimate the initial forward generation-interval distribution from the initial aggregated generation-interval distribution, measured during the exponential growth period.

intervals are also shortened by indirect connections: a susceptible individual can be infected through another route before the focal individual makes infectious contacts. Simulations on a small homogeneous network (i.e. complete network) confirm this additional effect (figure 3c). We can think of simulations on this network as an approximation of (local) infection process in a small household, consisting of five individuals; we expect realistic local network structures (and their effects on the realized generation intervals) to lie between a star network and a complete network.

In general, spatial reduction in the mean realized generation interval can be viewed as an effect of susceptible depletion and can be further classified into three levels: egocentric, local and global. Egocentric depletion, as discussed previously, is caused by an infected individual making multiple contacts to the same individual. Local depletion refers to a depletion of susceptible individuals in a household or neighbourhood; we can think of these structures as small homogeneous networks embedded in a larger population structure (and therefore we can expect similar effects to those seen in figure 3c). Both the egocentric and local depletion effects can be observed early in an epidemic, especially in a highly structured population, even if most of the population remains susceptible. Finally, global depletion refers to overall depletion of susceptibility at the population level, and explains the reduction in realized compared with intrinsic generation intervals that occurs even in a well-mixed population (figure 2).

5. Inferring the initial forward generation-interval distribution

In a large homogeneously mixing population, the initial forward generation-interval distribution is equivalent to the intrinsic distribution and provides the correct link between the exponential growth rate r and the initial reproductive number $\mathcal{R}_{\text{initial}}$ (see equation (3.3)). In a non-homogeneous population, the initial forward generation-interval distributions are subject to spatial effects and, therefore, are different from the intrinsic distribution. Since spatial effects have the same effect on how the epidemic spreads as they do on realized generation intervals, we expect the initial forward generation-interval distributions, which implicitly account for the spatial structure, to correctly link r and $\mathcal{R}_{\text{initial}}$ through the Euler–Lotka equation (equation (2.4)). Spatial effects on realized generation intervals are generally expected to be

analytically intractable, even in simple networks (e.g. see [20] for discussion regarding the realized generation intervals in a household with one infector and two susceptibles); therefore, we rely on simulations to validate this prediction.

When realized generation intervals are aggregated over the course of an epidemic, there will be four effects present in the data (figure 4): (i) right-truncation effect, (ii) egocentric depletion effect, (iii) local depletion effect, and (iv) global depletion effect. We can correct explicitly for the egocentric effect and, in the case of exponential growth, the right-truncation effect; these effects shorten the mean realized generation intervals, which in turn will reduce the estimate of the reproductive number [15,16]. While the other two effects are difficult to measure, we can make qualitative predictions about their effects on the realized generation intervals and reproductive numbers: both local and global depletion effects also reduce the number of infections that occur and shorten generation intervals. If we can correct for the truncation bias early in an outbreak, during the exponential growth phase, we should be able to infer the initial forward generation-interval distribution, which incorporates egocentric and local spatial effects but not the global effects, from the aggregated distribution.

Here, we investigate two methods for correcting for temporal bias in aggregated generation-interval data (see Methods for details). We refer to the first method as the population-level method as it relies on realized generation intervals aggregated across the entire population. When an epidemic is growing exponentially, right truncation causes the aggregated generation interval to be discounted by the exponential growth rate (equation (3.8)); hence, we can ‘undo’ the truncation by exponentially weighting the aggregated generation-interval distribution [19–21],

$$f_0(\tau) \propto a_0(\tau) \exp(r\tau), \quad (5.1)$$

where r is the exponential growth rate.

We refer to the second method as the individual-level method because it relies on individual contact information. We model each infection as a non-homogeneous Poisson process arising from the infector (equation (7.11)); incorporating information about time of infection of an infector, time of infection of an infectee and time since the beginning of an epidemic allows us to explicitly model the truncation process in the realized generation intervals. For both methods, the mean and coefficient of variation (CV) of the initial forward generation-interval distributions are estimated by maximum likelihood; the inferred generation-interval distributions are then used

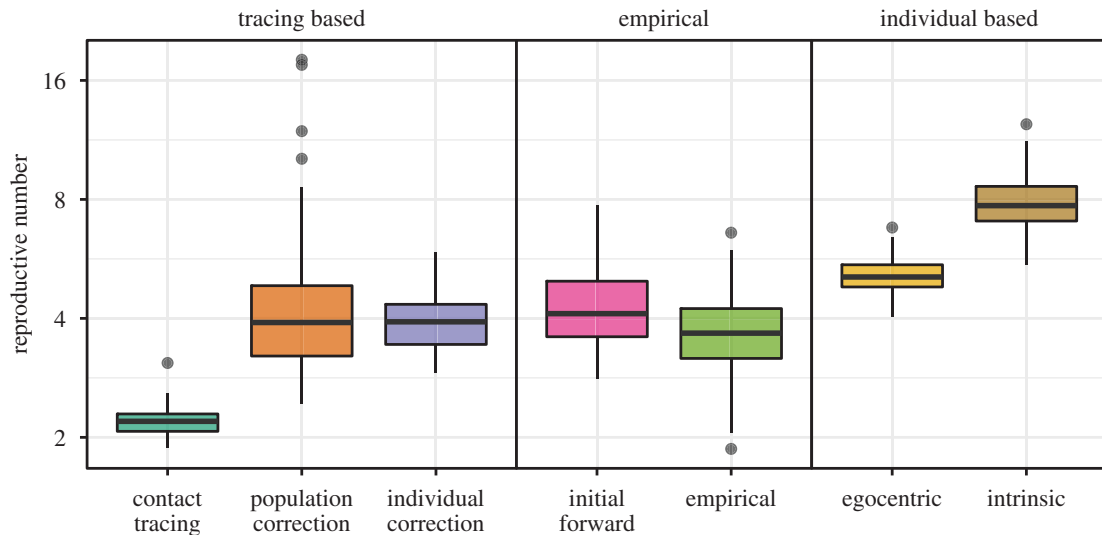


Figure 5. Comparison of estimates of reproductive number based on various methods. Using the aggregated generation-interval distributions (based on the first 1000 realized generation intervals) without correcting for right-truncation (labelled as ‘contact tracing’) severely underestimates the reproductive number. Similarly, using the intrinsic generation-interval distribution overestimates the reproductive number because it fails to account for local spatial effects; the egocentric distribution corrects for this only partially. Both population-level and individual-level methods provide estimates of reproductive number that are consistent with the empirical estimates, which we define as the average number of secondary cases generated by the first 75 infected individuals, as well as the estimates based on the initial forward generation intervals, which are calculated by applying the Euler-Lotka equation to the realized generation intervals of all infections caused by the first 75 infected individuals. Boxplots are generated using 100 stochastic simulations of the SEIR model on an empirical network using Ebola-like parameters [25]: mean latent period $1/\sigma = 11.4$ days and mean infectious period $1/\gamma = 5$ days. Per-pair contact rate $\lambda = 0.08 \text{ days}^{-1}$ is chosen to be sufficiently high such that differences are clear.

to estimate the initial reproductive number $\mathcal{R}_{\text{initial}}$ from the observed growth rate r using the Euler–Lotka equation.

To test these methods, we simulate 100 epidemics with Ebola-like parameters on an empirical network [36] and compare the estimates of the initial reproductive number with empirical reproductive numbers, which we define as the average number of secondary cases generated by the first 75 infected individuals, as well as the initial reproductive number calculated from the empirical initial forward generation intervals, which we define as the generation intervals for all infections caused by the first 75 infected individuals (figure 5). For simplicity, we assume that realized generation intervals are observed without error and assume that there is no under-reporting of generation intervals. We do not expect under-reporting to affect the inference of generation-interval distributions (see electronic supplementary material, appendix A.3) unless there are systematic biases in the observation process. On the other hand, it is difficult to measure generation intervals precisely because (i) infection events are often unobserved and (ii) there may be multiple potential infectors; these factors can introduce biases to the estimates of the initial reproductive number [21]. We do not pursue these directions in this study.

As expected, estimating the reproductive number based on the intrinsic generation-interval distribution overestimates the empirical reproductive number; estimates based on the egocentric generation-interval distribution (equation (4.3)) address this problem only partially, as they do not account for indirect (local) spatial effects. Direct estimates based on the aggregated generation intervals from contact tracing (via Euler–Lotka) severely underestimate the empirical estimates. While both population- and individual-level corrections provide similar estimates to the empirical estimates (as well as to estimates based on the untruncated empirical initial forward generation-interval distribution) on average, population-level estimates are more variable as they are more sensitive to outliers in generation intervals and our estimates of the initial exponential growth rate. For smaller values of $\mathcal{R}_{\text{initial}}$, we expect the differences to become

smaller. In the electronic supplementary material, we present the same figure using smaller $\mathcal{R}_{\text{initial}}$ (see electronic supplementary material, appendix A.1) and using Erlang-distributed latent periods (see electronic supplementary material, appendix A.2), which better corresponds to Ebola. Overall, our qualitative conclusions do not change.

6. Discussion

The intrinsic generation-interval distribution, which describes the expected time distribution of infectious contacts, provides a direct link between speed (initial exponential growth rate, r) and strength (initial reproductive number, $\mathcal{R}_{\text{initial}}$) of an epidemic in a homogeneously mixing population [13,15,16,24]. However, realized generation-interval distributions can vary depending on how and when they are measured [17,19–21]; determining which distribution correctly links r and $\mathcal{R}_{\text{initial}}$ can be challenging. Here, we analyse how realized generation intervals aggregated over the course of an epidemic, possibly through contact tracing, differ from intrinsic generation intervals. Changes due to right truncation reflect observation bias, whereas changes due to spatial or network structure reflect the dynamics of the outbreak. Thus, correcting the aggregated distribution for temporal, but not spatial, effects provides the correct link between r and $\mathcal{R}_{\text{initial}}$.

Realized generation intervals that have been aggregated over the course of an epidemic are subject to right truncation—it is not possible to trace individuals who have not been infected yet. The aggregated distributions can be thought of as averages of ‘backward’ generation intervals (measured by looking at infectors of a cohort of individuals infected at the same time) [17–21]. During an ongoing outbreak, the aggregated generation-interval distribution will always have a shorter mean than the intrinsic-interval distribution because of right truncation. Early in the outbreak, the initial aggregated intervals are

expected to match the initial backward intervals. Near the end of an outbreak, the effect of right truncation becomes negligible but the aggregated generation intervals are still shorter on average than intrinsic generation intervals, because of depletion of the susceptible population.

We think of susceptible depletion as operating on three levels: egocentric, local and global. Egocentric susceptible depletion refers to the effect of an infected individual making multiple contacts to the same susceptible individual. Accounting for the egocentric effect allows us to link the results by [22] to established results based on generation intervals. Local susceptible depletion refers to the effect of multiple 'linked' individuals (e.g. in the same household or neighbourhood) making infectious contacts to the same susceptible individual. Global susceptible depletion refers to the decrease in the susceptible proportion of the whole population.

Susceptible depletion happening at all three levels shortens realized generation intervals but acts on different time scales. Egocentric and local depletion effects are present from the beginning of an epidemic, even when depletion in the global susceptible population is negligible and can strongly affect the initial spread of an epidemic. Therefore, we predict the realized generation intervals during an exponential growth phase to contain information about the contact structure, allowing us to estimate the initial forward generation-interval distribution by simply accounting for the right truncation. Simulation studies confirm our prediction: using the initial forward generation-interval distribution provides the correct link between r and $\mathcal{R}_{\text{initial}}$.

We compare two methods for estimating the initial forward generation-interval distribution and assume that the initial forward generation-interval distribution follows a gamma distribution. The gamma approximation of the generation-interval distribution has been widely used because of its simplicity [9,37–40]; we previously showed that a gamma approximation (requiring estimation of only two parameters) can be sufficient to understand the role of generation-interval distributions in linking r and $\mathcal{R}_{\text{initial}}$ for Ebola, rabies and measles [16]. However, further investigation of our methods suggests that making a wrong distributional assumption can lead to biased estimates of the mean and CV of a generation-interval distribution (see electronic supplementary material, appendix A.4), even though the estimated gamma distribution may 'look' indistinguishable from the true shape of the intrinsic generation-interval distribution (derived from the SEIR model). These results are particularly alarming because it is impossible to know the true shape of the generation-interval distribution for real diseases. Nonetheless, biases in the parameter estimates of a generation-interval distribution may have opposite effects on the estimate of $\mathcal{R}_{\text{initial}}$ (e.g. shorter mean generation interval leads to lower $\mathcal{R}_{\text{initial}}$ whereas narrower generation-interval distribution leads to higher $\mathcal{R}_{\text{initial}}$) and, therefore, may have small effects on the overall estimate of $\mathcal{R}_{\text{initial}}$ (see electronic supplementary material, appendix A.4).

Generation-interval-based approaches to estimating the reproductive number often assume that an epidemic grows exponentially [12,14–16]. In practice, heterogeneity in population structure can lead to subexponential growth [41–46]; we therefore expect our simulations on an empirical network to be better characterized by subexponential growth models [46]. However, our simulations suggest that the initial exponential growth assumption still provides a viable approach for estimating the reproductive number.

Contact tracing provides an effective way of collecting epidemiological data and controlling an outbreak [47–49]. In particular, using tracing information allows us to infer real-time estimates of the time-varying reproductive number [50–53]. Generation-interval distributions, which can be either assumed or estimated, often play a central role in analysing tracing data. Our study illustrates that realized generation intervals over the course of an epidemic contain information about the underlying contact structure, which can be implicitly reflected in the estimates of the reproductive number; this perspective can be particularly useful for characterizing an epidemic because detailed information about the contact structure is often unavailable.

The generation-interval distribution is a key, and often under-appreciated, component of disease modelling and forecasting. Different definitions, and different measurement approaches, produce different estimates of these distributions. We have shown that estimates based on aggregated generation intervals (e.g. measured through contact tracing) differ in predictable ways from intrinsic estimates based on underlying measures of infectiousness (e.g. from shedding studies). These predictable differences can arise from temporal effects, egocentric spatial effects, local spatial (or network) effects and population-level effects. Correcting aggregated intervals for temporal effects allows us to estimate a spatially informed initial forward distribution, which accurately describes how disease spreads in a population. Future studies should carefully consider how measurement influences estimated generation-interval distributions, and how these distributions influence the spread of disease.

7. Methods

7.1. Deterministic SEIR model

To study the effects of right truncation on the realized generation intervals, we use the deterministic SEIR model. The SEIR model describes how disease spreads in a homogeneously mixing population; it assumes that infected individuals become infectious after a latent period. We use a SE^mIⁿR model, which extends the SEIR model to have multiple equivalent stages in the latent and infectious periods. This gives latent and infectious periods with Erlang distributions (gamma distributions with integer shape parameters, including the exponential distribution), which are often more realistic than the exponentially distributed periods in the standard SEIR model [54,55],

$$\left. \begin{aligned} \frac{dS}{dt} &= -\beta S \sum_{k=1}^{n_I} I_k, \\ \frac{dE_1}{dt} &= \beta S \sum_{k=1}^{n_I} I_k - n_E \sigma E_1, \\ \frac{dE_m}{dt} &= n_E \sigma (E_{m-1} - E_m) \quad \text{for } m = 2, 3, \dots, n_E, \\ \frac{dI_1}{dt} &= n_E \sigma E_m - n_I \gamma I_1 \\ \text{and } \frac{dI_n}{dt} &= n_I \gamma (I_{n-1} - I_n) \quad \text{for } n = 2, 3, \dots, n_I, \end{aligned} \right\} \quad (7.1)$$

where S is the proportion of susceptible individuals, E_m is the proportion of exposed individuals in the m -th compartment and I_n is the proportion of infectious individuals in the n -th compartment. Parameters of the model are specified as follows: β is the transmission rate, $1/\sigma$ is the mean latent period, n_E is the

number of latent compartments, $1/\gamma$ is the mean infectious period and n_i is the number of infectious compartments. We scale the proportions of individuals in each compartment by the total population size N . In the main text, we present results based on exponentially distributed latent and infectious periods; we show results based on Erlang distributed latent periods ($n_E = 2$), which better match the incubation period distribution of Ebola virus disease (see electronic supplementary material).

7.2. Stochastic SEIR model

We simulate an individual-based SEIR model on a contact network, using an algorithm based on the Gillespie algorithm [56,57]. We begin by randomly selecting individuals assumed to be infected at $t = 0$. For each infected individual i , we randomly draw the latent period E_i from an Erlang distribution with mean $1/\sigma$ and shape n_E . We then construct the random infectious period and infectious contact times simultaneously as follows. For each of the n_i stages of the infectious period, we draw the number of infectious contacts (before transitioning to the next compartment) from a geometric distribution with probability $n_i\gamma/(S_i\lambda + n_i\gamma)$, where S_i is the number of susceptible acquaintances and λ is the per-pair contact rate. We then choose the time between consecutive events (the chosen number of contacts, followed by exit from the given stage of infection) from an exponential distribution with rate $S_i\lambda + n_i\gamma$. For each contact, a contactee is uniformly sampled from the set of susceptible acquaintances of the individual i . The infectious period I_i is the sum of all of these waiting times.

After repeating the contact process for all initially infected individuals, all contacts are put into a sorted queue. The first person in the queue becomes infected (thus decreasing S_i by 1 for all individuals i that are acquaintances of the newly infected individual), and the current time is updated to infection time of this individual. Any subsequent contacts made to this individual are removed from the queue because they will no longer be effective. We repeat the contact process for this newly infected individual. Then, new contacts are added to the sorted queue. The simulation continues until there are no more contacts left in the queue.

7.3. Egocentric relationship between r and \mathcal{R} (SEIR model)

Here, we show that the egocentric relationship between r and \mathcal{R} derived by Trapman *et al.* [22] (see the original source for detailed derivations) matches what would be calculated by applying the Euler–Lotka equation to the egocentric (rather than the intrinsic) generation-interval distribution. Assume that latent and infectious periods are exponentially distributed with mean $1/\sigma$ and $1/\gamma$, respectively. Assuming a constant per-pair contact rate of λ for any pair, the egocentric generation-interval distribution can be written

$$\hat{g}(\tau) = \frac{\sigma(\gamma + \lambda)}{\sigma - (\gamma + \lambda)} (e^{-(\gamma + \lambda)\tau} - e^{-\sigma\tau}). \quad (7.2)$$

Substituting into equation (4.4), we get

$$\mathcal{R}_{\text{egocentric}} = \left(1 + \frac{r}{\sigma}\right) \left(1 + \frac{r}{\gamma + \lambda}\right), \quad (7.3)$$

where r is the exponential growth rate. Alternatively, the egocentric reproductive number can be expressed based on the degree distribution (mean μ and variance v) of a network,

$$\mathcal{R}_{\text{egocentric}} = \underbrace{\kappa\lambda}_{\text{average contact rate}} \times \underbrace{\frac{1}{\gamma + \lambda}}_{\text{mean effective infectious period}}, \quad (7.4)$$

where $\kappa = v/\mu + \mu - 1$, referred to as the mean degree excess [58],

describes the expected number of susceptible individuals that an average infected individual will encounter early in an outbreak. Combining the two equations, we get

$$\lambda = \frac{(\gamma + r)(\sigma + r)}{(\kappa - 1)\sigma - r}, \quad (7.5)$$

which completes the relationship between the growth rate and the egocentric reproductive number [22],

$$\mathcal{R}_{\text{egocentric}} = \frac{\gamma + r}{\gamma\sigma/(\sigma + r) + r/\kappa}. \quad (7.6)$$

7.4. Estimating the initial forward generation-interval distribution

The *population-level method* estimates the initial forward generation-interval distribution by reversing the inverse exponential weighting in the aggregated generation-interval distribution without explicitly accounting for the infection process (i.e. who infected whom) [19–21],

$$f_0(\tau) \propto a_0(\tau) \exp(r\tau), \quad (7.7)$$

where r is the initial exponential growth rate. In order to do so, we first approximate the aggregated distribution a_0 with a gamma distribution by assuming that realized generation intervals (subject to right truncation) during the exponential growth phase come from the same gamma distribution; specifically, we estimate the mean \bar{G} and shape α of a gamma distribution by maximum likelihood. Then, the initial forward generation-interval distribution follows a gamma distribution with mean $\alpha/(\alpha/\bar{G} - r)$ and shape α . We then use the estimated initial forward generation-interval distribution to infer the initial reproductive number $\mathcal{R}_{\text{initial}}$ from the estimated growth rate (using the Euler–Lotka equation).

The *individual-level method* models each infection i from an infected individual j as a non-homogeneous Poisson process between the time at which infector j was infected (t_j) and the truncation time (t_{truncate}), with time-varying Poisson rate at time t equal to $\Lambda f_0(t - t_j)$, where $f_0(t)$ is the initial forward generation-interval distribution [59]. We use a gamma distribution (parameterized by its mean and shape) to model the initial forward generation-interval distribution. Then, the probability that an individual j infects n_j individuals between t_j and t_{truncate} is equal to

$$\frac{\Lambda^{n_j} F_0(t_{\text{truncate}} - t_j; \theta)^{n_j} \exp(-\Lambda F_0(t_{\text{truncate}} - t_j; \theta))}{n_j!}, \quad (7.8)$$

where θ is a (vector) parameter of the initial forward generation-interval distribution f_0 (and the corresponding cumulative distribution function F_0). On the other hand, the probability density that the realized generation interval between infector j and infectee i is equal to $0 \leq s_{ij} \leq t_{\text{truncate}} - t_j$ can be expressed using a truncated distribution [60,61],

$$\frac{f_0(s_{ij}; \theta)}{F_0(t_{\text{truncate}} - t_j; \theta)}. \quad (7.9)$$

Therefore, the probability density that individual j infects n_j individuals between t_j and t_{truncate} with realized generation intervals $s_{i,j}$ for $i = 1, \dots, n_j$ is a product of equation (7.8) and equation (7.9),

$$\frac{\Lambda^{n_j} \exp(-\Lambda F_0(t_{\text{truncate}} - t_j; \theta)) \prod_{i=1}^{n_j} f_0(s_{i,j}; \theta)}{n_j!}. \quad (7.10)$$

The full likelihood of contact-tracing data, which include aggregated generation intervals as well as information about who

infected whom, until time t_{truncate} can be written as

$$\mathcal{L}(\Lambda, \theta | \mathbf{s}, \mathbf{t}, \mathbf{n}, t_{\text{truncate}}) = \prod_{j=1}^{N_I} \left(\frac{\Lambda^{n_j} \exp(-\Lambda F_0(t_{\text{truncate}} - t_j; \theta)) \prod_{i=1}^{n_j} f_0(s_{i,j}; \theta)}{n_j!} \right), \quad (7.11)$$

where N_I is the total number of infected individuals (in the data). This likelihood is a special case of the likelihood suggested in [62], which assumes that the second event (observation of infection) occurs simultaneously with the corresponding initiating event (infection of an individual). Here, we estimate parameters Λ and θ by maximum likelihood. In theory, the forward generation intervals arising from the same infector may be correlated because of non-independence in the contact process [35]; although we do not account for this potential correlation in our likelihood, our simulations (figure 5) suggest that approximating the initial forward generation-interval distribution with a single distribution provides a viable approach for estimating the reproductive number.

We use the estimated distribution f_0 to infer the initial reproductive number $\mathcal{R}_{\text{initial}}$ from the estimated growth rate (using the Euler–Lotka equation). When the entire transmission process is known, we expect Λ to match $\mathcal{R}_{\text{initial}}$; otherwise, Λ will be sensitive to under-reporting of the number of infections caused by each infected individual. As we show in electronic supplementary material, appendix A.3, the estimates of $\mathcal{R}_{\text{initial}}$ using the Euler–Lotka equation from f_0 remain unbiased even in the presence of random under-reporting.

7.5. Measuring the exponential growth rate

We estimate the initial exponential growth rate r of an epidemic from daily incidence by modelling the cumulative incidence $c(t)$ with a logistic function [11],

$$c(t) = \frac{K}{1 + [(K/c_0) - 1]e^{-rt}}. \quad (7.12)$$

While the exponential growth rate (i.e. the rate of change in log

incidence) of the logistic function changes throughout an epidemic, we focus strictly on estimating the *initial* exponential growth rate (when $t \rightarrow -\infty$). The method of estimating the initial exponential growth rate by fitting a logistic curve has been previously validated against simulations of stochastic compartmental models [11].

Fitting directly to cumulative incidence can lead to overly confident results [63]; instead, we fit interval incidence $x(t) = c(t + \Delta t) - c(t)$, where Δt is 1 day, to daily incidence, assuming that daily incidence follows a negative binomial distribution with overdispersion parameter θ . We estimate parameters r , K , c_0 and θ by maximum likelihood. The fitting time window is defined from the last trough before the peak of an epidemic to the first day after the peak of an epidemic.

7.6. Empirical network

To simulate epidemics on a realistic network, we use the ‘condensed matter physics’ network from the Stanford Large Network Dataset Collection [36]. This graph describes co-authorship among anyone who submitted a paper to the Condensed Matter category in arXiv between January 1993 and April 2003 [64]. It consists of 23 133 nodes and 93 497 edges. The same network was used by [22] to study how network structure affects the estimate of the basic reproductive number.

Data accessibility. All code is available at https://github.com/parksw3/contact_trace.

Authors’ contributions. SW.P. led the literature review, performed analytic calculations and simulations, and wrote the first draft of the manuscript; J.D. conceived the study, performed analytic calculations and wrote the first draft of the manuscript; all authors contributed to refining the study design, literature review and final manuscript writing. All authors gave final approval for publication.

Competing interests. The authors declare that they have no competing interests.

Funding. This work was supported by the Canadian Institutes of Health Research (funding reference no. 143486).

References

- Anderson RM, May RM. 1991 *Infectious diseases of humans: dynamics and control*. Oxford, UK: Oxford University Press.
- Diekmann O, Heesterbeek JAP, Metz JA. 1990 On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28**, 365–382. (doi:10.1007/BF00178324)
- Kribs-Zaleta CM, Velasco-Hernández JX. 2000 A simple vaccination model with multiple endemic states. *Math. Biosci.* **164**, 183–201. (doi:10.1016/S0025-5564(00)00003-1)
- Van den Driessche P, Watmough J. 2002 Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* **180**, 29–48. (doi:10.1016/S0025-5564(02)00108-6)
- Smith DL, Battle KE, Hay SI, Barker CM, Scott TW, McKenzie FE. 2012 Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. *PLoS Pathog.* **8**, e1002588. (doi:10.1371/journal.ppat.1002588)
- Dietz K. 1993 The estimation of the basic reproduction number for infectious diseases. *Stat. Methods Med. Res.* **2**, 23–41. (doi:10.1177/096228029300200103)
- Chowell G, Fenimore PW, Castillo-Garsow MA, Castillo-Chavez C. 2003 SARS outbreaks in Ontario, Hong Kong and Singapore: the role of diagnosis and isolation as a control mechanism. *J. Theor. Biol.* **224**, 1–8. (doi:10.1016/S0022-5193(03)00228-5)
- Mills CE, Robins JM, Lipsitch M. 2004 Transmissibility of 1918 pandemic influenza. *Nature* **432**, 904–906. (doi:10.1038/nature03063)
- Nishiura H, Castillo-Chavez C, Safan M, Chowell G. 2009 Transmission potential of the new influenza A(H1N1) virus and its age-specificity in Japan. *Euro Surveill.* **14**, 19227. (doi:10.2807/ese.14.22.19227-en)
- Nishiura H, Chowell G, Safan M, Castillo-Chavez C. 2010 Pros and cons of estimating the reproduction number from early epidemic growth rate of influenza A(H1N1) 2009. *Theor. Biol. Med. Model.* **7**, 1. (doi:10.1186/1742-4682-7-1)
- Ma J, Dushoff J, Bolker BM, Earn DJ. 2014 Estimating initial epidemic growth rates. *Bull. Math. Biol.* **76**, 245–260. (doi:10.1007/s11538-013-9918-2)
- Wearing HJ, Rohani P, Keeling MJ. 2005 Appropriate models for the management of infectious diseases. *PLoS Med.* **2**, e174. (doi:10.1371/journal.pmed.0020174)
- Svensson Å. 2007 A note on generation times in epidemic models. *Math. Biosci.* **208**, 300–311. (doi:10.1016/j.mbs.2006.10.010)
- Roberts M, Heesterbeek J. 2007 Model-consistent estimation of the basic reproduction number from the incidence of an emerging infection. *J. Math. Biol.* **55**, 803–816. (doi:10.1007/s00285-007-0112-8)
- Wallinga J, Lipsitch M. 2007 How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B* **274**, 599–604. (doi:10.1098/rspb.2006.3754)
- Park SW, Champredon D, Weitz JS, Dushoff J. 2019 A practical generation-interval-based approach to inferring the strength of epidemics from their

- speed. *Epidemics* **27**, 12–18. (doi:10.1016/j.epidem.2018.12.002)
17. Champredon D, Dushoff J. 2015 Intrinsic and realized generation intervals in infectious-disease transmission. *Proc. R. Soc. B* **282**, 20152026. (doi:10.1098/rspb.2015.2026)
 18. Kenah E, Lipsitch M, Robins JM. 2008 Generation interval contraction and epidemic data analysis. *Math. Biosci.* **213**, 71–79. (doi:10.1016/j.mbs.2008.02.007)
 19. Nishiura H. 2010 Time variations in the generation time of an infectious disease: implications for sampling to appropriately quantify transmission potential. *Math. Biosci. Eng.* **7**, 851–869. (doi:10.3934/mbe.2010.7.851)
 20. Scalia Tomba G, Svensson Å, Asikainen T, Giesecke J. 2010 Some model based considerations on observing generation times for communicable diseases. *Math. Biosci.* **223**, 24–31. (doi:10.1016/j.mbs.2009.10.004)
 21. Britton T, Scalia Tomba G. 2019 Estimation in emerging epidemics: biases and remedies. *J. R. Soc. Interface* **16**, 20180670. (doi:10.1098/rsif.2018.0670)
 22. Trapman P, Ball F, Dhersin JS, Tran VC, Wallinga J, Britton T. 2016 Inferring R_0 in emerging epidemics—the effect of common population structure is small. *J. R. Soc. Interface* **13**, 20160288. (doi:10.1098/rsif.2016.0288)
 23. Hampson K, Dushoff J, Cleaveland S, Haydon DT, Kaare M, Packer C, Dobson A. 2009 Transmission dynamics and prospects for the elimination of canine rabies. *PLoS Biol.* **7**, e1000053. (doi:10.1371/journal.pbio.1000053)
 24. Svensson Å. 2015 The influence of assumptions on generation time distributions in epidemic models. *Math. Biosci.* **270**, 81–89. (doi:10.1016/j.mbs.2015.10.006)
 25. WHO Ebola Response. 2014 Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *N. Engl. J. Med.* **371**, 1481–1495. (doi:10.1056/NEJMoa1411100)
 26. Heesterbeek J, Dietz K. 1996 The concept of R_0 in epidemic theory. *Stat. Neerl.* **50**, 89–110. (doi:10.1111/j.1467-9574.1996.tb01482.x)
 27. Diekmann O, Heesterbeek JAP. 2000 *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, vol. 5. New York, NY: John Wiley & Sons.
 28. Roberts M. 2004 Modelling strategies for minimizing the impact of an imported exotic infection. *Proc. R. Soc. B* **271**, 2411–2415. (doi:10.1098/rspb.2004.2865)
 29. Aldis G, Roberts M. 2005 An integral equation model for the control of a smallpox outbreak. *Math. Biosci.* **195**, 1–22. (doi:10.1016/j.mbs.2005.01.006)
 30. Champredon D, Dushoff J, Earn DJD. 2018 Equivalence of the Erlang-distributed SEIR epidemic model and the renewal equation. *SIAM J. Appl. Math.* **78**, 3258–3278. (doi:10.1137/18M1186411)
 31. Lotka AJ. 1907 Relation between birth rates and death rates. *Science* **26**, 21–22. (doi:10.1126/science.26.653.21-a)
 32. Fraser C. 2007 Estimating individual and household reproduction numbers in an emerging epidemic. *PLoS ONE* **2**, e758. (doi:10.1371/journal.pone.0000758)
 33. Ball F, Mollison D, Scalia-Tomba G. 1997 Epidemics with two levels of mixing. *Ann. Appl. Probab.* **7**, 46–89. (doi:10.1214/aop/1034625252)
 34. Ball F, Neal P. 2002 A general model for stochastic SIR epidemics with two levels of mixing. *Math. Biosci.* **180**, 73–102. (doi:10.1016/S0025-5564(02)00125-6)
 35. Yan P. 2008 Separate roles of the latent and infectious periods in shaping the relation between the basic reproduction number and the intrinsic growth rate of infectious disease outbreaks. *J. Theor. Biol.* **251**, 238–252. (doi:10.1016/j.jtbi.2007.11.027)
 36. Leskovec J, Krevl A. 2014 SNAP datasets: Stanford large network dataset collection. See <http://snap.stanford.edu/data>.
 37. McBryde E, Bergeri I, van Gemert C, Rotty J, Headley E, Simpson K, Lester R, Hellard M, Fielding JE. 2009 Early transmission characteristics of influenza A(H1N1 v in Australia: Victorian state, 16 May–3 June 2009. *Euro Surveill.* **14**, 19363. (doi:10.2807/ese.14.42.19363-en)
 38. Roberts MG, Nishiura H. 2011 Early estimation of the reproduction number in the presence of imported cases: pandemic influenza H1N1-2009 in New Zealand. *PLoS ONE* **6**, e17835. (doi:10.1371/journal.pone.0017835)
 39. Trichereau J, Verret C, Mayet A, Manet G, Decam C, Meynard JB, Deparis X, Migliani R. 2012 Estimation of the reproductive number for A(H1N1) pdm09 influenza among the French armed forces, September 2009–March 2010. *J. Infect.* **64**, 628–630. (doi:10.1016/j.jinf.2012.02.005)
 40. Nishiura H, Chowell G. 2015 Theoretical perspectives on the infectiousness of Ebola virus disease. *Theor. Biol. Med. Model.* **12**, 1. (doi:10.1186/1742-4682-12-1)
 41. Szendroi B, Csányi G. 2004 Polynomial epidemics and clustering in contact networks. *Proc. R. Soc. B* **271**, S364–S366. (doi:10.1098/rsbl.2004.0188)
 42. Chowell G, Viboud C, Hyman JM, Simonsen L. 2015 The Western Africa Ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. *PLoS Curr.* **7**. (doi:10.1371/currents.outbreaks.8b55f4bad99ac5c5db3663e916803261)
 43. Chowell G, Viboud C. 2016 Is it growing exponentially fast? – Impact of assuming exponential growth for characterizing and forecasting epidemics with initial near-exponential growth dynamics. *Infect. Dis. Model.* **1**, 71–78. (doi:10.1016/j.idm.2016.07.004)
 44. Chowell G, Viboud C, Simonsen L, Moghadas SM. 2016 Characterizing the reproduction number of epidemics with early subexponential growth dynamics. *J. R. Soc. Interface* **13**, 20160659. (doi:10.1098/rsif.2016.0659)
 45. Kiskowski M, Chowell G. 2016 Modeling household and community transmission of Ebola virus disease: epidemic growth, spatial dynamics and insights for epidemic control. *Virulence* **7**, 163–173. (doi:10.1080/21505594.2015.1076613)
 46. Viboud C, Simonsen L, Chowell G. 2016 A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks. *Epidemics* **15**, 27–37. (doi:10.1016/j.epidem.2016.01.002)
 47. Clarke J. 1998 Contact tracing for chlamydia: data on effectiveness. *Int. J. STD. AIDS* **9**, 187–191. (doi:10.1258/0956462981921945)
 48. Eames KT, Keeling MJ. 2003 Contact tracing and disease control. *Proc. R. Soc. B* **270**, 2565–2571. (doi:10.1098/rspb.2003.2554)
 49. Donnelly CA *et al.* 2003 Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet* **361**, 1761–1766. (doi:10.1016/S0140-6736(03)13410-1)
 50. Cauchemez S, Boëlle PY, Thomas G, Valleron AJ. 2006 Estimating in real time the efficacy of measures to control emerging communicable diseases. *Am. J. Epidemiol.* **164**, 591–597. (doi:10.1093/aje/kwj274)
 51. Hens N, Calatayud L, Kurkela S, Tamme T, Wallinga J. 2012 Robust reconstruction and analysis of outbreak data: influenza A(H1N1)v transmission in a school-based population. *Am. J. Epidemiol.* **176**, 196–203. (doi:10.1093/aje/kws006)
 52. Jewell CP, Roberts GO. 2012 Enhancing Bayesian risk prediction for epidemics using contact tracing. *Biostatistics* **13**, 567–579. (doi:10.1093/biostatistics/kxs012)
 53. Soetens L, Klinkenberg D, Swaan C, Hahné S, Wallinga J. 2018 Real-time estimation of epidemiologic parameters from contact tracing data during an emerging infectious disease outbreak. *Epidemiology* **29**, 230–236. (doi:10.1097/EDE.0000000000000776)
 54. Anderson D, Watson R. 1980 On the spread of a disease with gamma distributed latent and infectious periods. *Biometrika* **67**, 191–198. (doi:10.1093/biomet/67.1.191)
 55. Bailey NT. 1964 Some stochastic models for small epidemics in large populations. *J. R. Stat. Soc. C Appl. Stat.* **13**, 9–19. (doi:10.2307/2985218)
 56. Bartlett MS. 1953 Stochastic processes or the statistics of change. *J. R. Stat. Soc. C Appl. Stat.* **2**, 44–64. (doi:10.2307/2985327)
 57. Gillespie DT. 1977 Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361. (doi:10.1021/j100540a008)
 58. Newman ME. 2003 The structure and function of complex networks. *SIAM Rev.* **45**, 167–256. (doi:10.1137/S003614450342480)
 59. Daley DJ, Vere-Jones D. 2007 *An introduction to the theory of point processes: volume II: general theory and structure*. New York, NY: Springer-Verlag.

60. Lagakos SW, Barraj LM, Gruttola Vd. 1988 Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika* **75**, 515–523. (doi:10.1093/biomet/75.3.515)
61. Kalbfleisch J, Lawless J. 1991 Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica* **1**, 19–32.
62. Kalbfleisch J, Lawless JF. 1989 Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. *J. Am. Stat. Assoc.* **84**, 360–372. (doi:10.1080/01621459.1989.10478780)
63. King AA, de Celles MD, Magpantay FM, Rohani P. 2015 Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc. R. Soc. B* **282**, 20150347. (doi:10.1098/rspb.2015.0347)
64. Leskovec J, Kleinberg J, Faloutsos C. 2007 Graph evolution: densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* **1**, 2. (doi:10.1145/1217299.1217301)