Inferring generation-interval distributions from contact tracing data

Sang Woo Park, David Champredon and Jonathan Dushoff

September 19, 2018

1 Introduction

An epidemic can be characterized by its speed (exponential growth rate, r) and its strength (reproductive number, $\mathcal{R}$). Reproductive number, defined as the average number of secondary cases arising from a primary case, is of particular interest as it provides information about the final size of an epidemic [CITE]. However, directly measuring the reproductive number often requires detailed knowledge of disease natural history and may not be feasible early in an epidemic [CITE]. Instead, the reproductive number can be *indirectly* estimated from exponential growth rate, which can easily be estimated from incidence data [CITE]. These two quantities are linked by generation-interval distributions (Wallinga and Lipsitch, 2007).

Generation interval is defined as the time between when a person becomes infected and when that person infects another person. As each individual experiences different course of infection, *individual* generation-interval distribution varies among infectors [CITE sven]. The *intrinsic* generation-interval distribution, which provides link between r and $\mathcal{R}$, is a pooled distribution across all potential infections. [**JD:** *I'd like a more mechanistic definition – it reflects intrinsic infectiousness and is often used by modelers. What do we say in the other MS?*] [**JD:** *Maybe we don't need this s. What is it trying to do? See below.*]

Due to individual variation in infection time, the observed generation-interval distribution can change depending on when and how it is measured (Champredon and Dushoff, 2015). [**JD:** *Other cites first. Maybe cite us later when talking about qualitative effects. We need to talk clearly also about censored intervals*] There are important distinctions to be made when estimating generation intervals: *intrinsic* generation intervals are measured by evaluating the infectiousness of infected people, while *observed* generation intervals refer to the time between actual infection events. Observed generation intervals in turn can be *aggregated* across time, measured *forward* in time by looking at a cohort of individuals infected at the same time and asking when they infected others, or measured *backward in time* by look at a cohort and asking when their infectors were infected. When aggregated generation intervals are observed until a given

time in an ongoing epidemic, we refer to these as *censored* (or right-censored) intervals.

Early in the epidemic when depletion of susceptible is negligible, we expect the forward generation-interval distribution to be similar to the intrinsic generation-interval distribution. As epidemic progresses, an infector is less likely to infect individuals later in time due to decrease in susceptibles and the distribution of forward generation intervals will be shorter than the intrinsic distribution. [**JD:** *See how this relates to where you talk about similar shortening below.*] [**SWP:** *While this statement is true, it relies on the assumption that forward generation will be measured after all infected individuals that are infected early have recovered. We should note (and maybe explore the idea) that forward GI may be affected by censoring.*]

Conversely, when an epidemic is growing 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, and the distribution of backward generation intervals will be shorter than the intrinsic distribution. When epidemic is subsiding, most infections are caused by the remaining infectors, rather than new infectors, and the backward generation interval will be long.

In practice, generation intervals are measured by contact tracing throughout the course of an epidemic, usually aggregated to increase the amount of information available. [**JD:** *I kind of feel like, if you go on forever, aggregated GIs approach intrinsic GIs. This obviously isn't true for a single epidemic (observed GIs are shorter overall). On the other hand, it obviously is true if you add births and get to a stable equilibrium. So I guess the remaining question is, do we get convergence if we have stable cycles? What we think about this will affect how we write the rest of this P.*] If observation goes on throughout the disease cycle, then (something about something). On the other hand, if calculations are being done based on early-outbreak data, the available censored data is best thought of as a weighted average of backward generation-interval distributions; like them, the observed censored intervals will be shorter than intrinsic distributions. [**JD:** *See how this relates to where you talk about similar shortening above and below.*]

[**SWP:** *Need a paragraph about spatial effect?*] [**JD:** *Yes, we do!*]

In this study, we explore the temporal and spatial variation in the observed generation interval obtained from contact tracing. We show that using the observed generation-interval distribution directly will always underestimate the reproductive number [**SWP:** *need to confirm this statement when we're done with the ms*]. We provide a statistical framework of recovering the intrinsic generation-interval distribution from the observed generation-interval distribution.

[**SWP:** *Going to rename the sections... but I think I like the current order...*]

2 Measuring generation intervals - Theory

Following [CITE], let $K(t)$ be the infection kernel. The reproduction number is defined as

$$\mathcal{R} = \int_0^\infty K(t).$$

Then, the intrinsic generation-interval distributions is defined as

$$g(t) = \frac{K(t)}{\mathcal{R}}.$$

Intrinsic generation-interval distribution can be considered as an intrinsic characteristic of a single average infector in a fully susceptible population.

[**SWP:** *Insert renewal equation approach*]

2.1 Right-censored interval

Assume that contact tracing is performed from the beginning of an epidemic to time t. The number of infection occuring at time s caused by infectors who were themselves infected at time $s - \tau$ is given by

$$i_{s-\tau}(s) = \mathcal{R} i(s-\tau) g(\tau) S(s) \tag{1}$$

Then, total number of secondary infections that are τ time steps apart and occur before time t:

$$\mathcal{R} \int_\tau^t i(s-\tau) g(\tau) S(s) ds. \tag{2}$$

Then, the censored interval at time t is given by

$$c_t(\tau) = \frac{\mathcal{R} \int_\tau^t i(s-\tau) g(\tau) S(s) ds}{\mathcal{R} \int_0^t \int_x^t i(s-x) g(x) S(s) ds dx}. \tag{3}$$

We note that the expression in the denominater is equivalent to cumulative incidence at time t. The intuition behind this is that we are normalizing acrosss all incidence before time t. Then, we have

$$c_t(\tau) = \frac{\mathcal{R} \int_\tau^t i(s-\tau) g(\tau) S(s) ds}{\int_0^t i(s) ds}. \tag{4}$$

For convenience, we ignore normalizing constants and write

$$c_t(\tau) \propto g(\tau) \int_0^t i(s-\tau) S(s) ds. \tag{5}$$

For a single epidemic, the observed mean generation interval through contact tracing will always be shorter than intrinsic mean generation interval (Figure 1). [**JD:** *See how this relates to where you talk about similar shortening*

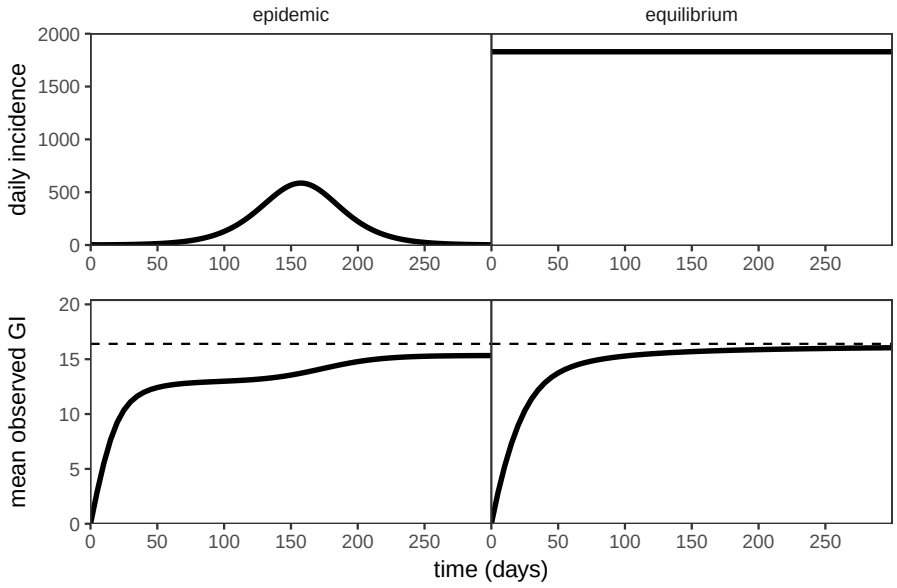

Figure 1: Fill this out.

above.] There are two reasons for this phenomenon. First, as contact tracing is performed up to time t, any infection events that occur after time t are not observed. Any individuals that are infected before time t can only complete infection events that are shorter than time t and the distribution of all infection events that occur before time t will be concentrated on shorter intervals. In particular, when an epidemic is growing exponentially ($i(t) \propto \exp(rt)$), the censored generation-interval distribution is equivalent to the inverse exponentially weighted intrisnc generation-interval distribution:

$$c_{\exp}(\tau) \propto g(\tau) \exp(-r\tau), \tag{6}$$

Second, number of susceptibles decrease over the course of an epidemic and any infector is less likely to infect susceptible individuals through long generation intervals than it would have in a fully susceptible population (Champredon and Dushoff, 2015). As a result, even if contact tracing is performed through an entire epidemic, mean generation interval will be underestimated.

When an epidemic is in an equilibrium state (i.e., incidence $i(t)$ and the number of susceptibles $S(t)$ is constant over time), the observed generation-interval distribution is equivalent to the intrinsic generation-interval distribution. On the other hand, when an epidemic is in a cycle, we get ??? **[SWP:** *TODO: Continue this exciting research!*]

2.2 Local depletion of susceptibles

[**SWP:** *cite trapman somewhere*] The intrinsic generation-interval distribution is often defined as the distribution of times at which secondary infections occur [CITE]. This definition implicitly assumes that all infectious contacts made give rise to infection. Given limited contact network, a susceptible individual can be contacted multiple times and the depletion of susceptibles occur locally. Infectious contacts give rise to an infection only when the susceptible individual is contacted for the first time. Therefore, the intrinsic generation-interval distribution is actually the distribution of times at which infectious contacts are made.

The *effective* generation-interval distribution – the distribution of times at which secondary infections occur from a single primary infector – depends on the probability that a susceptible individual has not been contacted yet and the intrinsic generation-interval distribution. Given a per pair contact rate $\beta(t)$ that does not vary across pairs for a single primary infector, the effective generation-interval distribution is given by [**SWP:** *Need to double check this and/or make it more general.*]

$$g_{\text{eff}}(\tau) \propto g(\tau) \exp\left(-\int_0^\tau \beta(t)dt\right). \tag{7}$$

Typically, when homogeneous mixing is assumed, the per pair contact rate is sufficiently small ($\beta \approx \mathcal{R}/N$ where N is the population size) that the effect of multiple contacts is negligible.

The difference between effective and intrinsic generation-interval distribution can be best demonstrated by simulating stochastic infection processes on a tree network with artificially high $\mathcal{R}$ (Figure ??? to be made). We observe that the distribution of generation times matches with the effective generation-interval distribution, whereas the distribution of contact times matches with the intrinsic generation-interval distribution. On the other hand, if a stochastic infection processes are simulated on a small homogeneous network (e.g., a household) with all else being equal, the distribution of generation times is even shorter than the effective generation intervals because a susceptible individual may be contacted by multiple infectors (and only the first contact gives rise to infection). [**JD:** *Let's think about how to explain this. In particular, the first one is where we have our correction that works . . .*]. Hence, we expect there to be varying degrees of local depletion of susceptibles in a heterogeneous contact structure and the accurate measurement of generation interval will be more difficult.

2.3 Spatial spread of an epidemic

3 Statistical approach to estimating generation-interval distribution

The observed generation-interval distribution is a weighted intrinsic generation-interval distribution (equation 5), Then, the intrinsic generation interval can be

recovered by taking the inverse weights:

$$g(\tau) \propto g_t(\tau) \frac{1}{\int_0^t i(s-\tau)S(s)ds} \tag{8}$$

However, this method requires a knowledge of susceptible dynamics and may not be feasible in practice.

During exponential growth period, we can write $i(\tau) \propto \exp(rt)$, where r is the exponential growth rate. Assuming that $S(t) \approx 1$, the observed generation-interval distribution during growth period can be written as follows:

$$g_{\exp}(\tau) \propto g(\tau) \exp(-r\tau), \tag{9}$$

Taking the inverse weight, we obtain the following expression for the intrinsic generation-interval distribution:

$$g(\tau) \propto g_{\exp}(\tau) \exp(r\tau) \tag{10}$$

and the reproductive number:

$$\mathcal{R} = \int_0^\infty g_{\exp}(\tau) \exp(r\tau) d\tau. \tag{11}$$

Therefore, applying the Lotka-Euler equation using the observed generation interval via contact tracing results in underestimation of reproductive number:

$$\int_0^\infty g_{\exp}(\tau) \exp(r\tau) d\tau > \left(\int_0^\infty g_{\exp}(\tau) \exp(-r\tau) d\tau \right)^{-1} \tag{12}$$

This method provides a non-parametric approach for inferring the intrinsic generation-interval distribution and the reproductive number from contact tracing data.

[SWP: *Need to rewrite this section. It's merely a place holder for now:*] While the non-parametric method is simple, it does not use all available information from contact tracing data. In particular, it does not take into account who infected whom. Assuming a poisson process, we can obtain a likelihood for observing infections:

$$\mathcal{R}^{n_e} \cdot \prod g(\tau_e) \cdot \exp\left(-\mathcal{R} \int_0^{c-t_{\inf}} g(s) ds \right) \tag{13}$$

This method requires us to make an assumption about the generation-interval distributions... See example...

4 Methods

References

Champredon, D. and J. Dushoff (2015). Intrinsic and realized generation intervals in infectious-disease transmission. *282*(1821).

Wallinga, J. and M. Lipsitch (2007). How generation intervals shape the relationship between growth rates and reproductive numbers. *Proceedings of the Royal Society of London B: Biological Sciences 274*(1609), 599–604.