

The effect of host population heterogeneity on epidemic outbreaks

M. C. J. Bootsma^{1,2,*}, K. M. D. Chan^{3,4}, O. Diekmann¹, and H. Inaba⁵

¹Department of Mathematics, Faculty of Science, Utrecht University, the Netherlands

²Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, the Netherlands

³Korteweg-de Vries Institute, University of Amsterdam, the Netherlands

⁴Transtrend BV, Rotterdam, The Netherlands

⁵Faculty of Education, Tokyo Gakugei University, Koganei-shi, Tokyo, Japan

*Corresponding author: M.C.J.Bootsma@uu.nl

January 17, 2024

Abstract

In the first part of this paper, we review old and new results about the influence of host population heterogeneity on (various characteristics of) epidemic outbreaks. In the second part we highlight a modelling issue that so far has received little attention: how do contact patterns, and hence transmission opportunities, depend on the size and the composition of the host population? Without any claim on completeness, we offer a range of potential (quasi-mechanistic) submodels. The overall aim of the paper is to describe the state-of-the-art and to catalyse new work.

Keywords: Kermack-McKendrick, epidemic outbreak model, static heterogeneity, final size, contact structures, herd immunity threshold

1 Introduction

As the title indicates, our aim is to investigate how heterogeneity influences various aspects of an epidemic outbreak in a demographically closed host population. In particular, we focus on the Basic Reproduction Number R_0 , the Malthusian parameter r , the final size, the timing and the size of the peak in incidence.

The paper consists of two parts. In the first part we adopt a top down approach. We first introduce a rather general model. Since its formulation employs measures over the trait space, it covers discrete and continuous traits in one go. We present various results that, essentially, have been known for a long time, but that might not exactly be ‘well known’. Next we consider various simplifications and their underlying interpretation and motivation. These lead us to recent results, for instance on the HIT (herd immunity threshold; see Section 7), that were triggered by the outbreak of Covid-19.

An important, yet sometimes rather implicit, ingredient of epidemic models is a specification of the rate at which an individual, with a certain trait, makes contact with other individuals having a specified

trait. With this in mind, Mossong et al. [52] studied age-structured contact patterns empirically, based on a population survey in various European countries. In the context of theoretical ‘what if’ studies, it is crucial to extrapolate such quantitative information. A concrete example is the final size of an outbreak of a new influenza type in Hong Kong [18] and the question of what to expect in the future when, due to ageing, the age-distribution has changed considerably. Another example is motivated by Corona ticket measures: How does the contact process at a venue change when unvaccinated individuals are banned and the vaccination coverage is age-dependent? The aim of the second part of this paper (Sections 8 and 9) is to initiate theoretical work on contact patterns by presenting various motivated options of how contact intensities might depend on the size and composition of the population and how this affects the final size of the outbreak.

For a general introduction to structured epidemic models, see Chapters 7, 8 and 9 of [23] and see [37, 43].

2 Model formulation

The host population we consider is assumed to be static with respect to demography, so neither birth nor death of host individuals is taken into account. Host individuals are characterized by a trait, denoted by symbols like x and ξ and sometimes called ‘type’, rather than trait, in particular in situations when there exist only finitely many types. The trait is static, i.e., does not change during the life of the host individual, but see Appendix B. The trait x takes values in a set Ω , which is a measurable space, i.e., Ω is equipped with a σ -algebra. For concreteness, we let Ω be a subset of \mathbb{R}^n with the Borel σ -algebra. The distribution of x within the host population is described by the (given/known) measure Φ , in order to unify the treatment of a ‘continuum’ setting, with a trait distribution described by a density, and the ‘discrete’ setting with finitely many types. The host population size is denoted by N . So $\Phi(\Omega) = 1$ and the number of individuals with trait belonging to the measurable subset ω of Ω equals $N\Phi(\omega)$. Apart from N and Φ we need just one modelling ingredient:

$$A(\tau, x, \xi) = \begin{array}{l} \text{the expected contribution to the force of infection on an individual with trait } x \\ \text{of an individual with trait } \xi \text{ that became infected } \tau \text{ units of time ago} \end{array} \quad (1)$$

Here A is a measurable non-negative function mapping $\mathbb{R}_+ \times \Omega \times \Omega$ into \mathbb{R}_+ and A is integrable with respect to (τ, ξ) over $\mathbb{R}_+ \times \Omega$. In due time we will make the ‘separable mixing’ assumption that A factorizes as the product of a function of x and a function of (τ, ξ) , but first we shall derive some general results.

Let $S(t, \omega)$ denote the number of susceptible individuals at time t with trait in ω . The model assumes that the measure $S(t, \cdot)$ is absolutely continuous with respect to Φ with bounded ‘derivative’. So for each t a bounded measurable function $s(t, \cdot)$, with $0 \leq s(t, x) \leq 1$, exists such that

$$S(t, \omega) = N \int_{\omega} s(t, x) \Phi(dx). \quad (2)$$

Thus $s(t, x)$ can be interpreted as the probability that an individual with trait x is susceptible at time t . Let $\Lambda(t, x)$ denote the force of infection at time t on individuals with trait x , i.e., assume that

$$\partial_t s(t, x) = -\Lambda(t, x) s(t, x). \quad (3)$$

In doubtful notation we say that the incidence at time t among individuals of trait ξ equals

$$-N \partial_t s(t, \xi) \Phi(d\xi).$$

The point is that, directly from the interpretation (2.1) , we have

$$\Lambda(t, x) = \int_0^\infty \int_\Omega A(\tau, x, \xi) \{-N \partial_t s(t - \tau, \xi) \Phi(d\xi)\} d\tau \quad (4)$$

As we show next, we can combine (2.3) and (2.4) to derive a nonlinear RE (Renewal Equation; see [20]) for s . To do so, we first integrate (2.3) while assuming that $s(-\infty, x) = 1$ for all x (to reflect that a very long time ago, so before the infectious agent made its appearance, the entire population was susceptible):

$$s(t, x) = \exp \left(- \int_{-\infty}^t \Lambda(\sigma, x) d\sigma \right). \quad (5)$$

Next we integrate (2.4) with respect to time from $-\infty$ to t . Interchanging the order of the integrals, and using that s is a primitive of $\partial_t s$, we arrive at

$$\int_{-\infty}^t \Lambda(\sigma, x) d\sigma = N \int_0^\infty \int_\Omega A(\tau, x, \xi) [1 - s(t - \tau, \xi)] \Phi(d\xi) d\tau. \quad (6)$$

Upon substitution of (2.6) into (2.5) we obtain the equation

$$s(t, x) = \exp \left(-N \int_0^\infty \int_\Omega A(\tau, x, \xi) [1 - s(t - \tau, \xi)] \Phi(d\xi) d\tau \right) \quad (7)$$

which provides a complete mathematical representation of the model and serves as the starting point of our analysis in the next section. Here we do not discuss the initial value problem, corresponding to prescribing the history of s on a time interval extending back to $-\infty$, nor the dynamical systems point of view, corresponding to shifting in time along the function s obtained by extending the given history. Instead, we refer to [20], [21] and [22] for an exposition of the relevant ideas. Also see [71].

In conclusion of this section, we mention that in the recent paper [25], it is argued that the discrete time variant

$$s(t, x) = \exp \left(-N \sum_{j=1}^{\infty} \int_\Omega A_j(x, \xi) [1 - s(t - j, \xi)] \Phi(d\xi) \right) \quad (8)$$

offers great computational advantages, in particular when Ω is a finite set. The key point is that there is no user-friendly tool to solve (7) and that (8) is straightforward to implement. For the numerical analysis point of view, we refer to Messina et al. [46–50].

3 The linearized problem

For given t , we think of $s(t, \cdot)$ as an element of

$$\mathcal{Y} = \{\psi : \psi \text{ is a bounded measurable function } \Omega \rightarrow \mathbb{R}\} \quad (9)$$

provided with the norm

$$\|\psi\| = \sup_{x \in \Omega} |\psi(x)|. \quad (10)$$

Equation (7) admits the disease free steady state solution $s = 1$ identically. Inserting

$$s(t, x) := 1 - y(t, x) \quad (11)$$

into (7) and assuming that y is small, we obtain, upon neglecting the higher order terms in the Taylor expansion, the linearized equation

$$y(t, x) = N \int_0^\infty \int_\Omega A(\tau, x, \xi) y(t - \tau, \xi) \Phi(d\xi) d\tau. \quad (12)$$

We refer to Sections 5 and 6 of [70] for an early profound analysis of such linear equations within the setting of positive operator theory.

When, as an Ansatz, we put

$$y(t, x) = \exp(\lambda t) \psi(x) \quad (13)$$

we obtain the nonlinear eigenvalue problem

$$\psi = \mathcal{K}_\lambda \psi \quad (14)$$

with, for λ in a right half plane of \mathbb{C} ,

$$(\mathcal{K}_\lambda \psi)(x) := N \int_\Omega k_\lambda(x, \xi) \psi(\xi) \Phi(d\xi) \quad (15)$$

$$k_\lambda(x, \xi) := \int_0^\infty A(\tau, x, \xi) e^{-\lambda \tau} d\tau. \quad (16)$$

We assume that

$$\sup_{(x, \xi) \in \Omega \times \Omega} k_0(x, \xi) < \infty \quad (17)$$

and interpret $\mathcal{K}_0 : \mathcal{Y} \rightarrow \mathcal{Y}$ as the Next Generation Operator (NGO), but this needs a bit of explanation.

The standard approach, as presented in [23] and [37], is to define the NGO as a bounded positive operator on $\mathcal{L}_1(\Omega)$. The generality achieved by describing the population composition by the measure Φ , precludes this in the present situation. We have to work with measures on Ω that are not necessarily absolutely continuous with respect to the Lebesgue measure. So here, guided by the interpretation just as in the standard approach, we introduce $\mathcal{K} : M(\Omega) \rightarrow M(\Omega)$ by defining

$$(\mathcal{K}m)(\omega) = N \int_\omega \left(\int_\Omega k_0(x, \xi) m(d\xi) \right) \Phi(dx). \quad (18)$$

Now note that \mathcal{K} maps a measure m to a measure of the special form

$$\omega \mapsto N \int_\omega \psi(x) \Phi(dx)$$

for some $\psi \in \mathcal{Y}$. The interpretation is that ψ specifies the distribution over Ω in the form of a fraction. This observation motivates us to define a bounded linear operator $T : \mathcal{Y} \rightarrow M(\Omega)$ by

$$(T\psi)(\omega) = N \int_\omega \psi(x) \Phi(dx). \quad (19)$$

We assume that T is injective or, in other words, that

$$\int_\omega \psi(x) \Phi(dx) = 0, \forall \omega \Rightarrow \psi(x) = 0, \forall x$$

(so, for instance, if Ω is a finite set, then Φ should be positive for all points). In this case we have

$$\mathcal{K}_0 = T^{-1} \mathcal{K} T \quad (20)$$

and the conclusion is that \mathcal{K}_0 is indeed the NGO, but represented in terms of fractions. We define the Basic Reproduction Number R_0 as the spectral radius of \mathcal{K}_0 .

In [2], V. Andreasen presents more or less the same observation for the special case that Ω is a finite set.

We define the Malthusian parameter r as the unique real root of

$$\text{spectral radius } \mathcal{K}_\lambda = 1 \tag{21}$$

if a real root exists. It is of interest to determine precise conditions on A and Φ such that

- R_0 is an eigenvalue with a corresponding positive eigenvector
- r exists
- \mathcal{K}_r has eigenvalue 1 with corresponding positive eigenvector
- $\text{sign}(R_0 - 1) = \text{sign}(r)$

We refer to [37] for results in the $\mathcal{L}_1(\Omega)$ setting and to [29] for results in the $M(\Omega)$ setting. When working with the supremum norm (so when investigating $\mathcal{K}_0 : \mathcal{Y} \rightarrow \mathcal{Y}$) it is helpful to strengthen the conditions on k_0 such that the range of \mathcal{K}_0 consists of continuous functions (in particular, in order to use the Arzela-Ascoli compactness criterion). We shall in fact do this in the next section, when discussing the final size as a function of x . But we do not elaborate these spectral considerations here, since we want to emphasize that an enormous simplification is achieved by assuming that A is the product of functions of less than three variables. This is elaborated in Section 6. In conclusion of this section we refer to [13–15] for recently developed numerical methods for the computation of R_0 . And to [72] for persistence results for an analogous model that does take demographic turnover into account.

4 Herd Immunity

When the outbreak progresses, the size of the susceptible subpopulation declines. At any time t , we can perform a thought experiment: suppose we remove instantaneously and with very high probability every individual that contributes to the future force of infection, does the outbreak make a restart and flare up or does it die out? If it dies out, we say that ‘Herd Immunity’ has been reached. The implication is that the incidence will dwindle, as, on average, new cases generate less than one case (often expressed by saying that the effective reproduction number, denoted by $R_{\text{eff}}(t)$, is less than one). The implication is NOT that in total there will be only a few future cases. Indeed, in actual fact there is a reservoir of already infected individuals that together generate a considerable force of infection and thus a prolonging of the outbreak and an increase of the outbreak size. When herd immunity is reached by vaccination, the population is safe. When herd immunity is reached during an outbreak, better times loom on the horizon, but the danger has not passed, the overshoot may be substantial (see [55]).

For a homogeneous host population, the situation is simple: herd immunity is reached when the susceptible fraction passes the value $1/R_0$. The complementary fraction $1 - 1/R_0$ is called the Herd Immunity Threshold, which is usually abbreviated to HIT. In the SIR compartmental model, this coincides with the prevalence I and the force of infection βI reaching a maximum. As a consequence, there is a tendency to identify “reaching the peak” and “passing the HIT”. But this is unwarranted, if only since different indicators of “severity” may reach a peak at different moments in time. Indeed, even for the SEIR compartmental model it happens that the peak in the force of infection βI does not

coincide with the peak in the prevalence, if one defines prevalence by $E + I$. See Section 10.1 below for other examples.

For a heterogeneous host population, the situation is, in general, rather complicated. For any measurable function $s : \Omega \rightarrow [0, 1]$, one can define R_{eff} by the methodology described in Section 3 and in [38]. The condition $R_{\text{eff}} = 1$ defines a codimension one manifold in the infinite-dimensional space of bounded measurable functions mapping Ω into $[0, 1]$, and the point where this manifold is ‘passed’ may very well depend on the initial condition, i.e., on the precise way in which the outbreak is triggered, see Section 10.2 below and see [57]. As a consequence, the HIT, as the complement of the fraction of the population still susceptible upon reaching herd immunity, is NOT a well defined CONCEPT. (As a side remark we note that for uniform vaccination there is no ambiguity, provided the contact process is in no way influenced by vaccination status; for such a vaccination scenario the HIT is equal to $1 - 1/R_0$, also for a heterogeneous population.)

In Section 7, we shall show that the heterogeneous situation becomes as simple as the homogeneous situation if the dynamics can be fully described in terms of a scalar function of time. Following the footsteps [31, 51] of Gabriela Gomes this will allow us to show that heterogeneity can lead to a major “reduction” of the HIT (terminology is dangerous here; what we mean is, that the fraction of the population still susceptible when herd immunity is reached, is substantially higher than $1/R_0$). Note that in the very recent paper [4] an example is presented where heterogeneity, in this case household structure, has the opposite effect!

We close with a word of warning: here we stay within the idealized world of models; using early phase data, to predict when herd immunity will be reached, is afflicted with serious and subtle difficulties, see [16] and the references given there.

5 The Final Size Equation

The interpretation requires that, for fixed x , $s(t, x)$ is a monotone non-increasing function of t . (Standard arguments can be used to show that, if one prescribes for each x the history of s on an interval of the form $(-\infty, t_0]$ by a non-increasing function with values in $[0, 1]$, then the equation defines a unique non-increasing extension to $(-\infty, +\infty)$ with values in $[0, 1]$.) As a bounded monotone function must have a limit, we know that $s(\infty, x)$ exists. By passing to the limit in (7) we obtain the so-called final size equation

$$s(\infty, x) = \exp \left(-\mathcal{K}_0 (1 - s(\infty, \cdot))(x) \right). \quad (22)$$

Please note the general form of this equation: it only depends on the particular model by way of the representation of the NGO in terms of fractions. Equation (22) has $s(\infty, \cdot) \equiv 1$ as trivial solution. This describes the situation in which the pathogen is *not* introduced in the host population. From now on we adopt the hypotheses

$$\forall \epsilon > 0, \exists \delta > 0 \text{ such that } \int_{\Omega} |k_0(x_1, \xi) - k_0(x_2, \xi)| \Phi(d\xi) < \epsilon \text{ when } |x_1 - x_2| < \delta \quad \text{H}_{A_1}$$

and

$$\exists m \text{ such that } \inf_{(x, \xi) \in \Omega \times \Omega} k^n(x, \xi) > 0 \text{ for } n \geq m \quad \text{H}_{A_2}$$

where k^n is defined inductively by

$$\begin{aligned} k^1(x, \xi) &:= k_0(x, \xi) \\ k^{n+1}(x, \xi) &:= \int_{\Omega} k(x, \eta) k^n(\eta, \xi) \Phi(d\eta), \end{aligned}$$

Below we demonstrate that these reasonable conditions on A guarantee that

- for $R_0 > 1$, equation (22) has precisely one nontrivial solution taking values in $[0, 1]$; in fact the values are bounded away from both 0 and 1
- for $R_0 \leq 1$ no such nontrivial solution exists.

So what happens when $R_0 \leq 1$ and we do infect a very small fraction of the host individuals with the pathogen? As explained in the elaboration of Exercise 1.22.iv in [23], the final size will be a Lipschitz continuous function of the size of the introduction, hence will be of the same order of magnitude as the introduction. In contrast, when $R_0 > 1$ such an introduction, no matter how small, causes a large outbreak as described by the nontrivial solution of (22). Here we add a warning: our deterministic description ignores the demographic stochasticity inherent to small *numbers*. We refer to section 1.3.4 in [23] for a description of the effects of demographic stochasticity.

The condition H_{A_1} on A has a double effect: it first guarantees that functions in the range of \mathcal{K}_0 are continuous (a weaker condition would suffice for that) and, next, that the restriction of \mathcal{K}_0 to the continuous functions on Ω is compact (by the Arzela-Ascoli Theorem) if Ω is compact. The condition H_{A_2} guarantees that \mathcal{K}_0 is irreducible in the strong sense that a power of \mathcal{K}_0 maps the positive cone, with the zero element excluded, into the interior of that cone (just like a primitive matrix). These properties of \mathcal{K}_0 will be used in the proofs below. These proofs are inspired by [65, 68, 69] and Appendix B in [60]. For other interesting aspects of the final size equation we refer to [2, 5, 10, 36, 41].

Defining

$$y(x) = 1 - s(\infty, x) \tag{23}$$

we rewrite (22) as

$$y = F(y) \tag{24}$$

where

$$F(y)(x) := 1 - e^{-(\mathcal{K}_0 y)(x)} \tag{25}$$

In the following we use “ \geq ” to denote the order relation induced by the cone of nonnegative functions on Ω . So

$$y_1 \geq y_2 \iff y_1(x) \geq y_2(x), \forall x \in \Omega. \tag{26}$$

\mathcal{K}_0 , being a positive linear operator, is *order preserving*:

$$y_1 \geq y_2 \Rightarrow \mathcal{K}_0 y_1 \geq \mathcal{K}_0 y_2 \tag{27}$$

Lemma 5.1. *F is order preserving, i.e.,*

$$y_1 \geq y_2 \Rightarrow F(y_1) \geq F(y_2). \tag{28}$$

Proof. Combine (27) with the fact that $z \mapsto 1 - e^{-z}$ is monotone increasing. \square

We want to find solutions $y \geq 0$ of (24). The form (25) of F implies that any such solution satisfies $y < 1$, in the sense that $y(x) < 1$ for all x .

Theorem 5.2. *If $R_0 < 1$, equation (24) has only the trivial solution $y = 0$.*

Proof. Since $1 - e^{-z} \leq z$ for $z \geq 0$ (see Lemma A.i) the inequality $F(y) \leq \mathcal{K}_0 y$ holds. So (24) yields $y \leq \mathcal{K}_0 y$ and, by induction, $y \leq \mathcal{K}_0^n y$ for $n \geq 1$. Taking the supremum with respect to x at both sides, it follows that

$$\sup_{x \in \Omega} y \leq \|\mathcal{K}_0^n\| \sup_{x \in \Omega} y$$

Now recall Gelfand's formula for the spectral radius

$$R_0 = \rho(\mathcal{K}_0) = \lim_{n \rightarrow \infty} \|\mathcal{K}_0^n\|^{\frac{1}{n}}.$$

Let n be so large that $\|\mathcal{K}_0^n\|^{\frac{1}{n}} < 1$, then also $\|\mathcal{K}_0^n\| < 1$ and the inequality above can only hold if $\sup y = 0$. □

Lemma 5.3. *Let y be a nontrivial solution of (24). If H_{A_2} holds, then $y(x) > 0$ for all $x \in \Omega$.*

Proof. Define $z_0 = \sup_{x \in \Omega} (\mathcal{K}_0 y)(x)$ and use the inequality (ii) of Lemma A to deduce that

$$y = F(y) \geq \left(\frac{1 - e^{-z_0}}{z_0} \right) \mathcal{K}_0 y.$$

By induction

$$y \geq \left(\frac{1 - e^{-z_0}}{z_0} \right)^n \mathcal{K}_0^n y$$

and hence H_{A_2} implies that $y(x) > 0$. □

Theorem 5.4. *Let Ω be compact and assume that H_{A_1} and H_{A_2} hold. Then equation (24) has at most one nontrivial solution.*

Proof. Let y and z be nontrivial solutions. Both are continuous and strictly positive on the compact domain Ω . Define

$$\theta = \min_{x \in \Omega} \frac{y(x)}{z(x)}$$

and assume that $\theta < 1$. Let $\bar{x} \in \Omega$ be such that $y(\bar{x}) = \theta z(\bar{x})$. Now use Lemma A.iv to obtain the contradiction

$$\begin{aligned} y(\bar{x}) &= F(y)(\bar{x}) = 1 - e^{-(\mathcal{K}_0 y)(\bar{x})} \geq 1 - e^{-\theta(\mathcal{K}_0 z)(\bar{x})} \\ &> \theta(1 - e^{-(\mathcal{K}_0 z)(\bar{x})}) = \theta z(\bar{x}). \end{aligned}$$

So $\theta \geq 1$ and $y(x) \geq z(x)$. But if we consider

$$\tilde{\theta} = \min_{x \in \Omega} \frac{z(x)}{y(x)}$$

exactly the same argument yields $\tilde{\theta} \geq 1$ and $z(x) \geq y(x)$. We conclude that $y(x) = z(x)$ for all $x \in \Omega$. □

Theorem 5.5. *Let Ω be compact and assume that H_{A_1} and H_{A_2} hold. If $R_0 > 1$, equation (24) has precisely one nontrivial solution.*

Proof. Assumption H_{A_1} guarantees that \mathcal{K}_0 is compact and assumption H_{A_2} that \mathcal{K}_0 is irreducible. On account of the (sharpening of the) Krein-Rutman Theorem (by de Pagter) we conclude that \mathcal{K}_0 has an eigenvector $y_0 \geq 0$ corresponding to R_0 , normalized by $\max_{x \in \Omega} y_0(x) = 1$. Choose $\epsilon > 0$ small enough to have $(1 - \epsilon)R_0 > 1$. Choose $\delta = \delta(\epsilon)$ as in Lemma A.iii. Let $\tilde{\delta} = \delta/R_0$. Then

$$\begin{aligned} F(\tilde{\delta} y_0) &= 1 - e^{-\tilde{\delta} \mathcal{K}_0 y_0} = 1 - e^{-\delta y_0} \geq (1 - \epsilon) \delta y_0 \\ &= (1 - \epsilon) R_0 \tilde{\delta} y_0 \geq \tilde{\delta} y_0. \end{aligned}$$

Iterating F , starting with $\tilde{\delta} y_0$, we obtain an increasing and bounded sequence that converges uniformly on Ω to a solution. □

Boundedness of the trait space Ω is essential for the uniform convergence towards the final situation. Indeed, when the trait specifies the spatial position on the line or in the plane, expansion of a locally introduced infection is characterized by an asymptotic speed of propagation, which is equal to the smallest possible speed c_0 of plane wave solutions, see [60] and the references in there.

Intuitively, it is crystal clear that the Basic Reproduction Number associated with the situation AFTER the outbreak should be less than one. Remarkably, a simple straightforward proof does not exist, as far as we know. The proof presented below is inspired by the proof given in [65] for the finite dimensional case.

Theorem 5.6. *Assume that Ω is compact, that H_{A_1} and H_{A_2} hold and that $R_0 > 1$. The NGO corresponding to the situation after the outbreak has spectral radius less than one.*

Proof. To describe the situation after the outbreak, we replace Φ by the measure

$$\omega \mapsto \int_{\omega} s(\infty, x) \Phi(dx) \quad (29)$$

where $s(\infty, \cdot)$ is the unique nontrivial solution of (22). Introducing

$$y(x) := 1 - s(\infty, x) \quad (30)$$

we have

$$\mathcal{K}_0^{\text{after}} = \mathcal{K}_0 L \quad (31)$$

where $L : \mathcal{Y} \rightarrow \mathcal{Y}$ is defined by

$$L\psi = (1 - y)\psi \quad (32)$$

So if we define

$$\bar{\mathcal{K}}_0 = L\mathcal{K}, \quad (33)$$

then

$$\bar{\mathcal{K}}_0 = L\mathcal{K}_0^{\text{after}}L^{-1} \quad (34)$$

We conclude that $\bar{\mathcal{K}}_0$ and $\mathcal{K}_0^{\text{after}}$ are similar, consequently have the same spectrum and therefore have the same spectral radius.

From H_{A_1} it follows that $\bar{\mathcal{K}}_0$, as a linear operator on $C(\Omega)$, is compact. Consequently, its adjoint, acting on $M(\Omega)$, is compact too. By the Krein-Rutman Theorem, the spectral radius is an eigenvalue of this adjoint with a nontrivial positive measure μ as the corresponding eigenvector. We denote the spectral radius by ρ . Now rewrite (22) as

$$\mathcal{K}_0 y + \ln(1 - y) = 0 \quad (35)$$

multiply both sides by $1 - y$ and integrate against μ over Ω . This leads to the identity

$$N \int_{\Omega} \mu(dx) [\rho y(x) + (1 - y(x)) \ln(1 - y(x))] = 0 \quad (36)$$

For $\rho \geq 1$ and $0 < z < 1$ the inequality

$$\rho z + (1 - z) \ln(1 - z) > 0 \quad (37)$$

holds. As a consequence, the left-hand side of (36) is strictly positive for $\rho \geq 1$ and it follows that necessarily $\rho < 1$. □

6 Separable Mixing: reduction to a scalar renewal equation

A preliminary conclusion: type-structure complicates epidemic models, but all the well-known results from the single-type situation do have a multi-type analogon. What we want to know, however, is what impact heterogeneity has on the dynamics. When it comes to doing calculations, the difference between the structured and the unstructured situation can be enormous.

Following Section 8.4 of [23] we now show how the assumption of separable mixing, (6.1) below, allows us to work with scalar quantities and thus facilitates the computational aspect tremendously. The key point is that various operators have a one-dimensional range. The interpretation is as follows: whenever the type at the moment of becoming infected is following an a priori given distribution (in particular independently of the type of the infecting individual), newly infected individuals are identical in a stochastic sense and therefore we know how to take averages. The aim of this section is to demonstrate that this principle is not restricted to R_0 , but extends to other aspects of the spread of infectious agents.

In recent Covid-driven work ([31], [51], [76], [53], [80], also see [45]), this feature was very effectively exploited. See [56] for a pre-Covid description of the main idea.

The key mathematical assumption is that the kernel A decomposes into two factors, one describing the influence of the type x of the one that may become infected and one describing the influence of the type ξ and the age-since-infection τ of the potential infector, i.e.,

$$A(\tau, x, \xi) = a(x)g(\tau, \xi) \quad (38)$$

When (38) holds, the operators \mathcal{K}_λ defined in (15) have one-dimensional range spanned by a . Consequently

$$R_0 = N \int_{\Omega} \left(\int_0^{\infty} g(\tau, \eta) d\tau \right) a(\eta) \Phi(d\eta) \quad (39)$$

while r is the unique REAL root of the Euler-Lotka equation

$$1 = N \int_{\Omega} \left(\int_0^{\infty} g(\tau, \eta) \exp(-\lambda\tau) d\tau \right) a(\eta) \Phi(d\eta) \quad (40)$$

Moreover, when we introduce the function w of t by putting

$$s(t, x) = \exp(-a(x)w(t)) \quad (41)$$

then insertion of (41) into (7) yields for w the RE

$$w(t) = N \int_0^{\infty} \int_{\Omega} g(\tau, \eta) \{1 - \exp(-a(\eta)w(t - \tau))\} \Phi(d\eta) d\tau \quad (42)$$

with corresponding final size equation

$$w(\infty) = N \int_{\Omega} \left(\int_{[0, \infty)} g(\tau, \eta) d\tau \right) \{1 - \exp(-a(\eta)w(\infty))\} \Phi(d\eta). \quad (43)$$

The Renewal Equation (42) is a delay equation, i.e., a rule for extending a function of time towards the future on the basis of the, assumed to be, known past. Solving such equations numerically is not really difficult, but user-friendly software does not exist (but see Messina et al. [46–50] for promising developments). A recently developed methodology for numerical bifurcation analysis via systematic approximation by ODE is described in [61]. For models that incorporate demographic turnover, see

e.g. [3,13], one could use this approach to study the stability of the endemic equilibrium, but for studying an outbreak in a demographically closed population it seems a bit of overkill (although one could, of course, use the ODE to compute an approximation of the solution of the RE). In [25], it is explained how one can formulate a discrete time variant of (42) with parameters that allow a clear interpretation (which is very helpful when it comes to identification on the basis of data). If Ω is a discrete set, with points numbered by an index i , and time is represented by the integers, corresponding to, say, days, this variant reads

$$w(t+1) = \sum_{k=1}^{k_{max}} \sum_i g_{ki} [1 - \exp(-a_i w(t-k))] N_i. \quad (44)$$

Here $N_i := N\Phi(i)$, a_i is a measure for the relative susceptibility of type i and g_{ki} describes the expected infectiousness of an individual of type i at day k after becoming infected. Often one is inclined to assume that g_{ki} factors into the product of a function of k and a function of i . Numerical implementation of (44) is straightforward. Also note that (44) is the separable mixing simplification of (8).

7 Separable mixing: detailed analysis, with special attention for Gamma distributed traits and the Herd Immunity Threshold

When we assume that

$$g(\tau, \eta) = b(\tau)c(\eta) \quad (45)$$

i.e.,

$$A(\tau, x, \xi) = a(x)b(\tau)c(\xi) \quad (46)$$

we can rewrite (42) in the form

$$w(t) = \int_0^\infty b(\tau)\Psi(w(t-\tau))d\tau \quad (47)$$

with $\Psi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ defined by

$$\Psi(w) = N \int_\Omega c(\eta) \left(1 - e^{-a(\eta)w}\right) \Phi(d\eta). \quad (48)$$

So note that the influence of the model ingredients N , c , a , Φ is captured by Ψ , a scalar function of one variable. In the present section we analyse (47)-(48). The aim is to deduce epidemiological relevant conclusions that reveal the impact of heterogeneity. As part of our analyses we shall make specific choices for a , c , Φ . For instance, note that the choice $a \equiv 1$, $c \equiv 1$ captures the homogeneous situation with

$$\Psi(w) = N \left(1 - e^{-w}\right) \quad (49)$$

In the companion paper [26] we show how a specific choice of b (involving a matrix exponential, cf. [24]) leads to a large class of ODE systems that correspond to heterogeneous versions of familiar compartmental models.

By linearization we find that

$$R_0 = \Psi'(0) \int_0^\infty b(\tau)d\tau \quad (50)$$

with

$$\Psi'(0) = N \int_{\Omega} c(\eta)a(\eta)\Phi(d\eta) \quad (51)$$

and that the Euler-Lotka equation reads

$$1 = \Psi'(0) \int_0^{\infty} b(\tau)e^{-\lambda\tau} d\tau. \quad (52)$$

Consider an ongoing outbreak. We want to quantify the effect of the reduction in the susceptible subpopulation brought about so far by the infection process itself. To do so, we perform a thought experiment: pretend that the reservoir of already infected individuals does not exist and that the current susceptible subpopulation is the entire population in which the pathogen is introduced. We denote by R_{eff} the corresponding reproduction number. As long as $R_{\text{eff}} > 1$ the outbreak is still in its accelerating phase, but as soon as $R_{\text{eff}} < 1$ deceleration starts (yet many more victims are to be expected, simply since in reality this reservoir of already infected individuals may well be huge). So $R_{\text{eff}} = 1$ characterizes the turning point at which a gradual improvement slowly sets in. The fraction of the population that is still susceptible when R_{eff} reaches the value 1, is called the Herd Immunity Threshold (HIT).

In order to determine this HIT for the present model, we require that linearization at the value \bar{w} should yield the value 1 for the corresponding reproduction number (so the root 0 for the corresponding Euler-Lotka equation). This amounts to

$$1 = \Psi'(\bar{w}) \int_0^{\infty} b(\tau)d\tau$$

which we rewrite in the form

$$\Psi'(\bar{w}) = \frac{\Psi'(0)}{R_0}. \quad (53)$$

Since Ψ'' is negative, there exists a unique solution $\bar{w} > 0$ if $R_0 > 1$. The HIT is the fraction of the population that is susceptible when w reaches the value \bar{w} , and is given by

$$\bar{s} = \int_{\Omega} e^{-a(x)\bar{w}}\Phi(dx). \quad (54)$$

For $t \rightarrow \infty$, w tends to $w(\infty)$ characterized by

$$w(\infty) = \Psi(w(\infty)) \int_0^{\infty} b(\tau)d\tau = \frac{\Psi(w(\infty))}{\Psi'(0)} R_0 \quad (55)$$

and the fraction of the population that escapes is accordingly given by

$$\bar{s} = \int_{\Omega} e^{-a(x)w(\infty)}\Phi(dx). \quad (56)$$

Note that (55) implies that $\Psi'(w(\infty)) < \frac{\Psi'(0)}{R_0}$ (since $\Psi(y) > y\Psi'(y)$ for $y > 0$) and hence that $w(\infty) > \bar{w}$.

When Ω is a subset of \mathbb{R} , it makes sense to organize the parameterization of susceptibility in such a way that

$$a(x) = x. \quad (57)$$

(Note that in the decomposition (46) we may accommodate multiplicative constants in any of the factors. Here we choose to keep the specification of a and c simple.). Concerning the impact of the

trait on the infectiousness, we may now contrast (following work of G. Gomes and co-workers [31, 51]) the case

$$c(\xi) = 1 \tag{58}$$

where there is no impact at all, with the case

$$c(\xi) = \xi \tag{59}$$

where susceptibility and infectiousness are perfectly correlated. A straightforward computation based on (50)-(51), shows that in the second case R_0 is a factor

$$\text{mean} + \frac{\text{variance}}{\text{mean}}$$

bigger than in the first case, where “mean” and “variance” refer to the distribution of the trait as described by Φ . This is a well-known result, cf. Section 7.4.1 of [23], leading to the key insight that for sexual activity structured populations the variance contributes to R_0 . In principle, we can in a similar manner compare the value \bar{w} for the two cases, but since it is hard to say something in general about the solution of (53), we first choose a particular Ω and $\bar{\Phi}$ that will enable us to do explicit calculations.

Let

$$\Omega = (0, \infty)$$

and let Φ be the Gamma Distribution with mean 1 and variance $\frac{1}{p}$, meaning that Φ has density

$$x \mapsto \frac{p^p}{\Gamma(p)} x^{p-1} e^{-px}. \tag{60}$$

(That the mean equals one is not a loss of generality, it can be achieved by scaling of the variable x , and the scaling constant can, as noted before, be incorporated in the factor b of A .) As noted in [53, 56, 76], a key feature of the Gamma Distribution is that, when it is reduced according to (41), with a given by (57), we obtain again a Gamma Distribution with exactly the same variance, but a reduced mean. Another important feature is that the Laplace Transform is given explicitly by

$$\bar{\Phi}(\lambda) = \left(\frac{\lambda}{p} + 1 \right)^{-p} \tag{61}$$

which facilitates the computation of moments via derivatives of the Laplace Transform evaluated in $\lambda = 0$. In this context, also note that when (57) holds and $\Omega = (0, \infty)$ we have

$$\int_0^\infty e^{-xw(t)} \Phi(dx) = \bar{\Phi}(w(t)). \tag{62}$$

Moreover, denoting (58) as “Case I” and (59) as “Case II”, we have

$$\Psi(w) = N \begin{cases} (1 - \bar{\Phi}(w)) & \text{Case I} \\ (-\bar{\Phi}'(0) + \bar{\Phi}'(w)) & \text{Case II.} \end{cases} \tag{63}$$

So if Φ is the Gamma Distribution we obtain

$$\Psi(w) = \begin{cases} 1 - \left(\frac{w}{p} + 1 \right)^{-p} & \text{Case I} \\ 1 - \left(\frac{w}{p} + 1 \right)^{-p-1} & \text{Case II} \end{cases} \tag{64}$$

and hence

$$\bar{w} = \begin{cases} p \left(R_0^{\frac{1}{p+1}} - 1 \right) & \text{Case I} \\ p \left(R_0^{\frac{1}{p+2}} - 1 \right) & \text{Case II} \end{cases} \quad (65)$$

and, according to (54),

$$\bar{s} = \begin{cases} R_0^{-1 + \frac{1}{p+1}} & \text{Case I} \\ R_0^{-1 + \frac{1}{\frac{1}{2}p+1}} & \text{Case II} \end{cases} \quad (66)$$

These expressions should be contrasted with the HIT

$$\bar{s} = R_0^{-1}$$

for the homogeneous situation. We see that in both cases heterogeneity with large variance (i.e., small p) brings about a substantial reduction of the HIT. The reason is, of course, that among the individuals infected so far the highly susceptible individuals are over-represented. So here we see that letting the outbreak run its course is a far more efficient way of immunizing a population than at random vaccinating individuals, when there is substantial variation in susceptibility.

This effect plays already a role in the early stage of the outbreak. For small w we have

$$\Psi'(w) = \Psi'(0) - N \int_{\Omega} c(\xi) a^2(\xi) \Phi(d\xi) w + \dots$$

and

$$\begin{aligned} s = \text{fraction susceptible} &= 1 - \int_{\Omega} a(\xi) \Phi(d\xi) w + \dots \\ &= 1 - w + \dots \end{aligned}$$

if we normalize a by requiring

$$\int_{\Omega} a(\xi) \Phi(d\xi) = 1.$$

It follows that the reduction in reproduction number relates to the reduction in the fraction susceptible according to

$$\frac{\Psi'(w)}{\Psi'(0)} = 1 - \theta(1 - s) + o(1 - s)$$

with

$$\begin{aligned} \theta &:= \frac{\int_{\Omega} c(\xi) a^2(\xi) \Phi(d\xi)}{\int_{\Omega} c(\xi) a(\xi) \Phi(d\xi)} \\ &= \begin{cases} 1 + \frac{1}{p} & \text{Case I} \\ 1 + \frac{2}{p} & \text{Case II.} \end{cases} \end{aligned} \quad (67)$$

A. Tkachenko e.a. write in [76]: “We named the coefficient θ the immunity factor because it quantifies the effect that a gradual build up of population immunity has on the spread of an epidemic”.

In conclusion of this section we refer to our recent paper [9] for an analysis of the effect of mask wearing on HIT and final size. Our study was inspired by [58]. See [28, 54, 75] for the wider context.

8 The influence of population size on the contact process

Transmission is superimposed on contact. In the present section we recall some observations made in [23], in particular in Section 1.3.3 and Chapter 12, concerning the influence of the population size N on the contact intensity. In the next section we shall focus on the influence of population composition as described by Φ . For the sake of exposition, consider the homogeneous SIR model. Whether we write

$$\frac{dS}{dt} = -\beta SI \quad (68)$$

or

$$\frac{dS}{dt} = -\frac{\tilde{\beta}}{N}SI \quad (69)$$

is irrelevant as long as N is a given constant, since by

$$\tilde{\beta} = \beta N \quad (70)$$

we can identify the two equations. But what if we want to compare the spread of the same disease in different geographical areas, for instance countries? First of all we should ascertain whether the variables are numbers or spatial densities, cf. Section 1.3.5 of [23]. In stochastic models we deal with finite numbers. In deterministic models we let numbers go to infinity and yet want to work with something finite, such as a spatial density (number/area) or a fraction (of the total). If we introduce

$$s = \frac{S}{N}, \quad i = \frac{I}{N} \quad (71)$$

then (68) and (69) transform into, respectively

$$\frac{ds}{dt} = -\beta N s i \quad (72)$$

$$\frac{ds}{dt} = -\tilde{\beta} s i. \quad (73)$$

If we want to allow for variable N , is it more appropriate to consider β as constant (i.e., as independent of N) or should we consider $\tilde{\beta}$ as constant? If we think in terms of spatial densities and aerosol transmission, ‘contacts’ are reminiscent of colliding molecules in a gas and (72) with β constant seems most appropriate. For STD’s (Sexually Transmitted Diseases), for mosquito transmission in a host-vector context and for sun-bathing seals, [23], Section 1.3.3 and references given there, (73) with $\tilde{\beta}$ constant seems most appropriate. In the first case, the average number of contacts that an individual has per unit of time scales with population size N , so is homogeneous of degree 1. In the second case it is homogeneous of degree 0, so independent of N . A somewhat intermediate situation is described in [33, 74] and Section 12.2 of [23]. It is based on the idea that contacts have a certain duration, so take time. As a consequence there is an upper bound for the number of contacts per unit of time. Yet at small values of N the number of contacts per unit of time is proportional to N .

The underlying contact sub-model assumes that individuals can be ‘single’ or ‘paired with another individual’. Let N measure the total number of individuals, X the number of singles and $2P$ the number of individuals that form a pair with another individual (in other words, there are P pairs). Then

$$X + 2P = N. \quad (74)$$

Concerning the dynamics, assume that the rate at which a single becomes part of a pair depends on the availability of potential partners in the sense that it is proportional to X , with constant r . Assume

that a pair spontaneously dissolves at rate s (so has an exponentially distributed life time with mean duration $1/s$). Then

$$\begin{aligned}\frac{dX}{dt} &= -rX^2 + 2sP \\ \frac{dP}{dt} &= \frac{1}{2}rX^2 - sP.\end{aligned}\tag{75}$$

By making use of (74) we reduce to a scalar differential equation and next it is easy to show that the solution converges to the steady state $\bar{P} = \frac{1}{2}\theta\bar{X}^2$ with

$$\theta := \frac{r}{s}\tag{76}$$

and

$$\bar{X} := \frac{1}{2\theta} \left(\sqrt{1 + 4\theta N} - 1 \right).\tag{77}$$

The probability that a randomly chosen individual participates in a pair equals

$$C(N) := \frac{2\bar{P}}{N} = \frac{1 + 2\theta N - \sqrt{1 + 4\theta N}}{2\theta N}\tag{78}$$

in steady state. The idea is now to assume that

- the processes of pair formation and dissolution occur on a much faster time scale than disease transmission
- disease status has no influence at all on pair formation and separation
- transmission only occurs within pairs.

So at any moment in time a susceptible belongs to a pair with probability $C(N)$ and, if so, its partner is infectious with probability I/N and, if so, there is a certain probability per unit of time, say β , that transmission occurs. This leads to

$$\begin{aligned}\frac{dS}{dt} &= -\beta SC(N) \frac{I}{N} \\ \frac{dI}{dt} &= \beta SC(N) \frac{I}{N} - \alpha I\end{aligned}\tag{79}$$

in the familiar SIR setting and to

$$\frac{dS}{dt} = SC(N) \frac{1}{N} \int_0^\infty \beta(\tau) S(\cdot - \tau) d\tau\tag{80}$$

in the general Kermack-McKendrick framework. Note that $C(N) \sim \theta N$ for small N and that $C(N) \rightarrow 1$ for $N \rightarrow \infty$. For a more detailed justification of (79) we refer to [33] and Section 12.2 of [23]. As far as we know, the ‘derivation’ of (79) is so far purely formal. One of the things we shall consider in the next section, is a multitype version of the pair formation sub-model described above. This too is based on [33]. We refer to [78] for biologically motivated modelling considerations.

9 The influence of population composition on the contact process

The general model ingredient $A(\tau, x, \xi)$ incorporates information about how expected intrinsic infectiousness depends on ξ and τ , how intrinsic susceptibility depends on x , but also on how the probability per unit of time for an individual with trait x to have contact with an individual of type ξ depends on the combination of x and ξ . The aim of this section is to describe a kind of catalogue of possibilities for this last aspect, as first presented in the unpublished manuscript [18], which is based on [17]. The work reported in this manuscript originated from a very concrete question: how do we extrapolate information about the impact of the H1N1-2009 Influenza outbreak in Hong Kong to a future outbreak of a similar new influenza strain, taking into account the predictable demographic changes, in particular the ageing of the population, i.e., the relative increase of the older part of the population? The manuscript is available upon request to K.M.D. Chan.

Recently, the question on how the contact structure depends on the size of different age groups popped up in another context, viz., the effectiveness of measures to prevent the spread of SARS-CoV-2: When only vaccinated individuals are allowed access to certain premises like theaters and restaurants, and the vaccination coverage is inhomogeneous among age groups, the age distribution of individuals at the premises will change. Consequently, the contact intensities between age groups will change as well and the question is how to model these changes in the absence of direct observations of the changes in the contact process at those premises [8].

So here we want to extend the considerations of the foregoing section to the multi-type situation. (For a recent overview focusing on compartmental models see [34].)

Recall that we use the words ‘type’ and ‘trait’ interchangeably, but have a tendency to use the former when there are finitely many types and the latter when the trait-space might be, or contain, a continuum. Here we do indeed restrict to finitely many types, partly for technical reasons, partly to avoid modelling difficulties related to how accurately individuals can distinguish one type of individual from another (what is the difference between an individual on its 70th birthday and an individual that had its 70th birthday a fortnight ago?). In particular we specialize (7) to

$$s_i(t) = \exp \left(- \int_0^\infty \sum_{j=1}^m A_{ij}(\tau) [1 - s_j(t - \tau)] N_j d\tau \right) \quad (81)$$

where x and ξ are replaced by integers numbering the m points in the support of Φ and $N_i := N\Phi(i^{\text{th}}$ point of support). The corresponding final size equation reads

$$s_i(\infty) = \exp \left(- \sum_{j=1}^m \int_0^\infty A_{ij}(\tau) d\tau [1 - s_j(\infty)] N_j \right). \quad (82)$$

Next, assume that

$$A_{ij}(\tau) = \frac{1}{N_i} k_{ij} b_{ij}(\tau) \quad (83)$$

where

$$k_{ij} := \text{expected number of contacts per unit of time that a type-}j \text{ individual has with type-}i \text{ individuals} \quad (84)$$

(so the factor $1/N_i$ serves to translate to a ‘per i -type individual’ probability per unit of time) and $b_{ij}(\tau)$ specifies the product of the infectiousness of an (j, τ) individual and the susceptibility of an

i -type individual (it is tempting to put $b_{ij}(\tau) = a_i \tilde{b}_j(\tau)$; but at this point we would like to include situations in which, for instance, an individual that is aware of its Covid-19 infection status might choose to care for its children, while avoiding to meet its parents; also note that the factors in a product are never unique, so we are free to put such a reduction of contact intensity into b , even though the description in words might suggest to put it into k). Clearly the consistency relation

$$k_{ij}N_j = k_{ji}N_i \quad (85)$$

should hold. Let K denote the matrix with elements k_{ij} . We allow K to depend on the vector

$$\mathbf{N} = (N_1, \dots, N_m). \quad (86)$$

We call a specification of how K depends on \mathbf{N} a CONTACT PATTERN.

In general, a contact pattern K can be represented by a function $K : \mathbb{R}^m \rightarrow \mathbb{R}^{\frac{m(m+1)}{2}}$, as relation (85) implies that the upper-triangular part of the matrix K specifies the full matrix K . Moreover, one expects a contact pattern to be continuous, such that small changes in \mathbf{N} lead to small changes in K .

When K is homogeneous of degree 1 and, more precisely, when

$$k_{ij} = q_{ij}N_i \quad (87)$$

with $q_{ij} = q_{ji}$ and q_{ij} constant, i.e., independent of \mathbf{N} for all $1 \leq i, j \leq m$, we call the contact pattern DENSITY DEPENDENT. When, on the other hand, K is homogeneous of degree 0, i.e., when for all $c > 0$ we have

$$K(c\mathbf{N}) = K(\mathbf{N}) \quad (88)$$

we call the contact pattern FREQUENCY DEPENDENT. Note that in the multi-type frequency dependent case, knowledge of K for a certain \mathbf{N} , does not fully specify the contact pattern. Only when all group sizes change with the same factor, the contact matrix K will remain identical.

In the special case

$$k_{ij} = \frac{c_i c_j N_i}{\sum_{\ell=1}^m c_\ell N_\ell} \quad (89)$$

(with c_i constant) we speak about PROPORTIONATE MIXING, while if for all pairs (i, j) with $i \neq j$ k_{ij} depends on N_i and N_j but NOT on N_ℓ for $\ell \neq i, j$ we speak about a BILATERAL pattern. In a bilateral pattern each pair of types of individuals ‘decides’ on how changes in the group size of the two types affect their contact intensities. This ‘decision process’ may differ for each pair of types. Hence, there exist many bilateral patterns. There are two mathematically convenient ways to ‘decide’ on changes in the contact intensities between two different groups $i \neq j$.

One is the POWER LAW

$$k_{ij} = q_{ij} \left(\frac{N_i}{N_j} \right)^d, \quad (90)$$

which is homogeneous of degree 0, but satisfies the consistency condition (85) only when $d = 1/2$. The other is when there is a DOMINATING type which keeps its contact rate constant, while the other type has to adapt its contact rate to satisfy the consistency condition (85), i.e., when type j dominates type i we have

$$\begin{aligned} k_{ij} &= q_{ij} \\ k_{ji} &= q_{ij} \frac{N_j}{N_i}. \end{aligned} \quad (91)$$

Note that neither the power-law nor the dominating pattern does nail down the within-group contact intensities k_{jj} , even in the case of only two types. One may, of course, assume that the within-group

contact intensities k_{jj} do not depend on \mathbf{N} , such that equations (90) and (91) are valid as well for $i = j$. But other assumptions make sense too, e.g., that the total contact intensity of a type- j -individual, $\sum_{i=1}^m k_{ij}$, does not depend on \mathbf{N} .

In general, contact patterns are not bilateral. In practical applications, the contact matrix K is often estimated for a given \mathbf{N} , and instead of attempting to determine the full contact pattern, one would like to know the contact matrix \tilde{K} for a specific $\tilde{\mathbf{N}} \neq \mathbf{N}$. As a contact matrix has $\frac{m(m+1)}{2}$ degrees of freedom, it is a challenge to avoid arbitrariness.

One way to deal with the contact pattern problem is to define an explicit model for pair formation, see Haderer [32].

In the Heesterbeek-Metz approach [33], it is assumed that pair formation occurs according to the law of mass action and that pairs disband at a certain rate (or, in other words, pairs exist for an exponentially distributed amount of time). The short time scale dynamic equations for pair formation and dissolution are

$$\begin{aligned} \frac{dX_i}{dt} &= - \left(\sum_{j=1}^m r_{ij} X_j \right) X_i + 2 \sum_{j=1}^m s_{ij} P_{ij} \\ \frac{dP_{ij}}{dt} &= \frac{1}{2} r_{ij} X_i X_j - s_{ij} P_{ij} \end{aligned} \quad (92)$$

where we use a similar notation as in (75), but now with indices indicating the type or the two types forming a pair.

Here we assume that pairs P_{ij} are symmetric entities and that accordingly

$$r_{ij} = r_{ji} \text{ and } s_{ij} = s_{ji} \text{ and } P_{ij} = P_{ji} \quad (93)$$

in the sense that the first two identities are requirements for the ingredients r and s while the third is, subsequently, a consequence of (92). The factor $1/2$ in front of r_{ij} serves to be able to treat $i = j$ and $i \neq j$ in an identical way. The consequence is that the number of pairs consisting of an i -type individual and a j -type individual equals $2P_{ij}$ if $i \neq j$ (or, if you prefer, equals $P_{ij} + P_{ji}$). Accordingly we have

$$X_i + 2 \sum_{j=1}^m P_{ij} = N_i \quad (94)$$

It can be shown that convergence to a (unique) steady state is guaranteed, see [33] and references in there.

To facilitate the notation, we now omit bars when we consider variables in steady state. In steady state we have to have

$$P_{ij} = \frac{1}{2} \theta_{ij} X_i X_j \quad (95)$$

with

$$\theta_{ij} := \frac{r_{ij}}{s_{ij}} \quad (96)$$

So we can rewrite (94) as

$$X_i + \sum_{j=1}^m \theta_{ij} X_i X_j = N_i. \quad (97)$$

(Incidentally, note that if we divide this identity by N_i , the first term at the left hand side is the probability that a type- i individual is single, while the term with index j in the sum gives the probability that a type- i individual is paired to a type- j individual. This observation shall be used below.)

We would like to relate the steady state to a known contact matrix K . As the ij^{th} element of the contact matrix K represents the number of contacts a type j individual has, per unit of time, with type i individuals, we want to find values of r_{ij} and s_{ij} such that

$$k_{ij} = 2P_{ij}s_{ij}, \quad (98)$$

since $2P_{ij}s_{ij}$ is the number of ij -pairs that dissolve per time unit, which, in the equilibrium, equals the number of new contacts per time unit. By the symmetry postulated in equation (93), both $(r_{ij})_{1 \leq i, j \leq m}$ and $(s_{ij})_{1 \leq i, j \leq m}$ have $m(m+1)/2$ degrees of freedom. So in total there are $m(m+1)$ degrees of freedom for $(r_{ij})_{1 \leq i, j \leq m}$ and $(s_{ij})_{1 \leq i, j \leq m}$ combined. As a contact matrix K is determined by its upper triangular part, (98) puts $m(m+1)/2$ restrictions on s_{ij} and r_{ij} . This means that for a general contact matrix K , there is a $m(m+1)/2$ -dimensional set of $(r_{ij})_{1 \leq i, j \leq m}$ and $(s_{ij})_{1 \leq i, j \leq m}$ such that the corresponding equilibrium contact process leads to the contact matrix K . Hence, additional restrictions on $(r_{ij})_{1 \leq i, j \leq m}$ and $(s_{ij})_{1 \leq i, j \leq m}$ are needed to uniquely determine how the r, s coefficients, and thus how the elements of the contact matrix, change if \mathbf{N} changes. Here we consider two options.

In the first option, we simplify the situation. We assume that the two individuals involved have an independent influence on both pair formation and separation, in the sense that both r_{ij} and s_{ij} are the product of an i -dependent factor and a j -dependent factor, leading to

$$\theta_{ij} = \rho_i \rho_j. \quad (99)$$

If we insert

$$X_i = \zeta_i N_i \quad (100)$$

into (97), use (99) and divide the identity by N_i we obtain

$$\zeta_i + \sum_{j=1}^m \rho_i \rho_j \zeta_i \zeta_j N_j = 1 \quad (101)$$

which we rearrange into

$$\sum_{j=1}^m \rho_j \zeta_j N_j = \frac{1 - \zeta_i}{\rho_i \zeta_i}. \quad (102)$$

As the left hand side does not depend on i , the same must hold for the right hand side. Let us call the common value Q . Then

$$\zeta_i = \frac{1}{1 + \rho_i Q} \quad (103)$$

while Q itself should satisfy the equation

$$\sum_{j=1}^m \frac{\rho_j N_j}{1 + \rho_j Q} = Q \quad (104)$$

which has, as a simple graphical argument shows, a unique positive solution. Thus we have constructively defined the solution of the steady state problem for any given (N_1, \dots, N_m) . Now how do we use these steady state expressions in the context of an epidemic model? The probability that an i -type individual is paired to j -type individual is given by

$$C_{ij} = \frac{2P_{ij}}{N_i} = \frac{\rho_i \rho_j X_i X_j}{N_i}. \quad (105)$$

Under the assumptions formulated below (78) and focussing on an SIR or SEIR setting we should put

$$\frac{dS_i}{dt} = -S_i \sum_{j=1}^m \beta_{ij} C_{ij} \frac{I_j}{N_j} \quad (106)$$

while the analogue of (80) reads

$$\frac{dS_i}{dt} = -S_i \sum_{j=1}^m C_{ij} \frac{1}{N_j} \int_0^\infty \beta_{ij}(\tau) \frac{dS_j}{dt}(\cdot - \tau) d\tau. \quad (107)$$

In [33] and the references given there, one finds far more information about how to prove the existence, uniqueness and global stability of the steady state of (92) even when the simplifying assumption (99) does not hold.

In the second option, we do require that the total number of contacts per time unit of an individual does not depend on \mathbf{N} . This idea is inspired by the observation that during the SARS-CoV-2 outbreak, people visiting certain premises, did not change their behaviour substantially [64]. In this context, the type specifies the age group to which an individual belongs. We denote the original, known, contact matrix by K and the original population size by \mathbf{N} . By putting tildes on top of these parameters we denote the new situation for which we want to determine the contact matrix \tilde{K} . Let $\tilde{N}_j := \rho_j N_j$, i.e., the lower ρ_j , the lower the attendance of type- j -individuals in the new situation.

We define an ordering σ such that $\rho_{\sigma(1)} \leq \rho_{\sigma(2)} \leq \dots \leq \rho_{\sigma(m)}$, so type- $\sigma(1)$ has the highest relative reduction in participation.

Originally, the average total number of contacts per unit of time of an individual of age group $\sigma(1)$ equalled $\sum_{i=1}^m k_{i\sigma(1)}$. We assume that $\tilde{k}_{j\sigma(1)}$, the number of contacts of a type- $\sigma(1)$ -individual with type- j individuals per unit of time after the intervention, equals:

$$\tilde{k}_{j\sigma(1)} = \frac{\rho_j k_{j\sigma(1)}}{\sum_{i=1}^m \rho_i k_{i\sigma(1)}} \sum_{i=1}^m k_{i\sigma(1)} \quad (108)$$

In this way, the total number of contacts of a type- $\sigma(1)$ -individual remains $\sum_{i=1}^m k_{i\sigma(1)}$ and the intensity of contacts with age group j are proportional to ρ_j and $k_{j\sigma(1)}$.

To keep contacts symmetric, we need that $\tilde{k}_{\sigma(1)j}$, the number of contacts per unit of time of a type- j -individual with type- $\sigma(1)$ individuals, equals:

$$\tilde{k}_{\sigma(1)j} = \frac{k_{\sigma(1)j} \rho_{\sigma(1)}}{k_{j\sigma(1)} \rho_j} \tilde{k}_{j\sigma(1)}. \quad (109)$$

We have constructed the contact of and with group $\sigma(1)$ which has the highest reduction in attendance. Next, we will define the contact rate of and with group $\sigma(2)$ individuals. However, $\tilde{k}_{\sigma(1)\sigma(2)}$ and $\tilde{k}_{\sigma(2)\sigma(1)}$ are already defined. As the total contact rate of each type of individual remains constant, the total contact rate of type- $\sigma(2)$ -individual with types other than $\sigma(1)$ needs to be:

$$R_{\sigma(2)} := \left(\sum_{i=1}^m k_{i\sigma(2)} \right) - \tilde{k}_{\sigma(1)\sigma(2)} \quad (110)$$

We distribute the remaining contact rate $R_{\sigma(2)}$ of type $\sigma(2)$ -individuals over all types $j \neq \sigma(1)$ in a similar way as we did in (108), i.e.,

$$\tilde{k}_{j\sigma(2)} := \frac{\rho_j k_{j\sigma(2)}}{\sum_{i=2}^m \rho_{\sigma(i)} k_{\sigma(i)\sigma(2)}} R_{\sigma(2)} \quad (111)$$

To keep contacts symmetric, we need that $\tilde{k}_{\sigma(2)j}$, the number of contacts per unit of time of a type- $j \neq \sigma(1)$ -individual with type- $\sigma(2)$ individuals, equals:

$$\tilde{k}_{\sigma(2)j} = \frac{k_{\sigma(2)j} \rho_{\sigma(2)}}{k_{j\sigma(2)} \rho_j} \tilde{k}_{j\sigma(2)}. \quad (112)$$

We now recursively define all contact rates this way. More precisely, suppose we know the new contact rate of the $n-1$ types with the highest reduction in attendance, i.e., we know $\tilde{k}_{\sigma(i)j}$ and $\tilde{k}_{j\sigma(i)}$ for $1 \leq i \leq n-1 < m$ and $1 \leq j \leq m$. The total contact rate of type- $\sigma(n)$ -individuals with all type $\sigma(j)$ -individuals with $n-1 < j \leq m$ equals:

$$R_{\sigma(n)} := \left(\sum_{i=1}^m k_{i\sigma(n)} \right) - \sum_{i=1}^{n-1} \tilde{k}_{\sigma(i)\sigma(n)} \quad (113)$$

We define the contact rate of a type n -individual with type $\sigma(j)$ individual, with $n \leq j \leq m$, as;

$$\tilde{k}_{\sigma(j)\sigma(n)} := \frac{k_{\sigma(j)\sigma(n)} \rho_{\sigma(j)}}{\sum_{i=n}^m \rho_{\sigma(i)} k_{\sigma(i)\sigma(n)}} R_{\sigma(n)} \quad (114)$$

i.e., contacts with types $\sigma(1), \dots, \sigma(n-1)$ are already defined, and the contacts with the remaining types are such that they are proportional to both the original contact rate with that type and the reduction factor of that type. The contacts are scaled such that the total number of contacts of individuals of age-group $\sigma(n)$ is the same as in the original contact matrix. By symmetry of the contacts we have that:

$$\tilde{k}_{\sigma(n)\sigma(j)} = \frac{k_{\sigma(n)\sigma(j)} \rho_{\sigma(n)}}{k_{\sigma(j)\sigma(n)} \rho_{\sigma(j)}} \tilde{k}_{\sigma(j)\sigma(n)}. \quad (115)$$

Thus we recursively construct the contact rates between all types. This provides us with a new contact matrix $\tilde{\mathbf{K}}$ which still satisfies the symmetry-condition (85) and keeps the total contact rate of individuals fixed. This iterative procedure was used in a report for the Dutch government to assess the effectiveness of Corona ticket measures [8].

Note that these two options do not at all exhaust all possibilities!

10 Numerical illustration of some subtle issues

10.1 Peaks

Even without heterogeneity, one needs numerical methods to determine the size and timing of a peak in the incidence and to investigate how these quantities depend on the model ingredients, see e.g. [25]. With heterogeneity, the need for numerical methods intensifies. And, more importantly, new phenomena arise. First, before we can even look for peaks, we need to agree upon the quantity that we graph as a function of time. Here we choose two measures of the incidence, i.e., the number of cases per time unit that become infected and the number of cases per time unit that become infectious. Depending on whether individuals in the latent period (before they become infectious) are symptomatic or not, the first or the latter may be closest to surveillance data in a situation where one may be unaware of relevant heterogeneity. Now imagine, as a thought experiment, two well-mixed subpopulations characterized by a major difference of the latent period (reflecting, for instance, a genetically determined difference in immune physiology). Figure 1a shows TWO peaks in the graph of total incidence of new infections as a function of time. If we graph the incidence of new infections in the two subpopulations separately, the

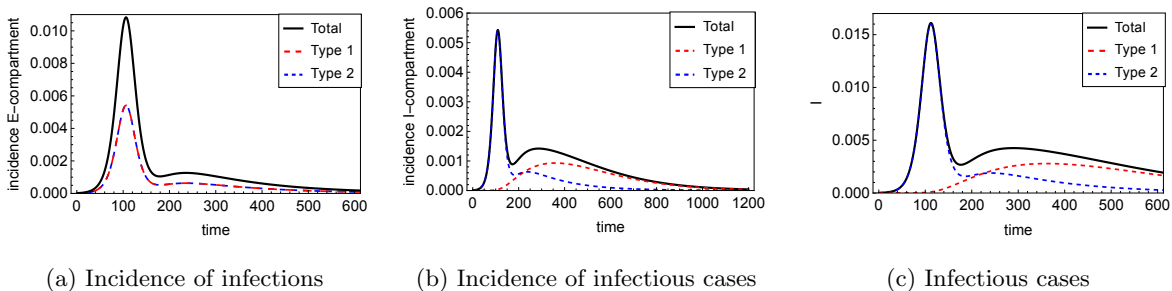


Figure 1: **Incidence with heterogeneity in latent period.** We model two groups of equal size and random mixing of both groups. The groups only differ in the duration of the latent period. Both are Gamma-distributed with shape-parameter 3 while the scale parameter is 1 for group 1 and 0.01 for group 2. The infectious period is exponentially distributed with mean 3 for both groups. R_0 equals 3. (a) Incidence of new infections. (b) Incidence of new infectious individuals. (c) I-compartment as fraction of the total population

two graph are identical and the peaks occur at the same points in time for the two subpopulations. But if we graph the incidence of infectious cases Figure 1b or the contribution of the two subpopulations to the force of infection, Figure 1c, i.e., the number of individuals that are infectious, a different picture emerges: these contributions are out of phase. It is this difference in phase that causes the double peak.

In the example above, contact is uniform but physiology is not. Let's now reverse this and consider neighbouring countries inhabited by identical individuals, but with weak coupling (in the sense of contacts). More precisely, assume that the two subpopulations are of equal size and have equal within-subpopulation contact rate. We introduce the between-subpopulation contact rate as a small parameter ϵ . For $\epsilon = 0$ the two subpopulations are uncoupled, each shows a single peak, but the timing of the peak depends, of course, on the timing of the introduction of the pathogen. Now make ϵ a tiny bit positive. Technically we obtain irreducibility: when we introduce the pathogen in one of the two subpopulations, ultimately both will be hit and, in terms of final size, in equal measure. Yet the outbreaks are bound to be out of phase.

And indeed, in Figure 2 we can observe TWO peaks in the graphs of the total incidence (both for new infections and for new infectious cases) as a function of time. This time (and in contrast with the first example) each peak relates to the incidence in a subpopulation. Thus, this example illustrates that the question “when do we speak about one population and when about two ?” has a subtle quantitative aspect, in addition to the more obvious public health administration aspect. At the end of Section 7.3 (page 175) in [23] this is described as ‘quantitative aspects of irreducibility’: it may happen that nonlinear dynamics sets in BEFORE the distribution takes the shape predicted by the stable distribution of the linearized model (this happened for instance with HIV; the gay community suffered considerably before the disease was observable in the heterosexual community, simply since the connection, though non-zero, is so very weak). Or, in other words, the basic reproduction number R_0 of the coupled population definitely has critical value one, and yet the linearization might give a prediction/suggestion that isn't necessarily right. In conclusion: whether a peak occurs also depends on what you plot and, even without control measures or arrival of new variants, multiple peaks may be found

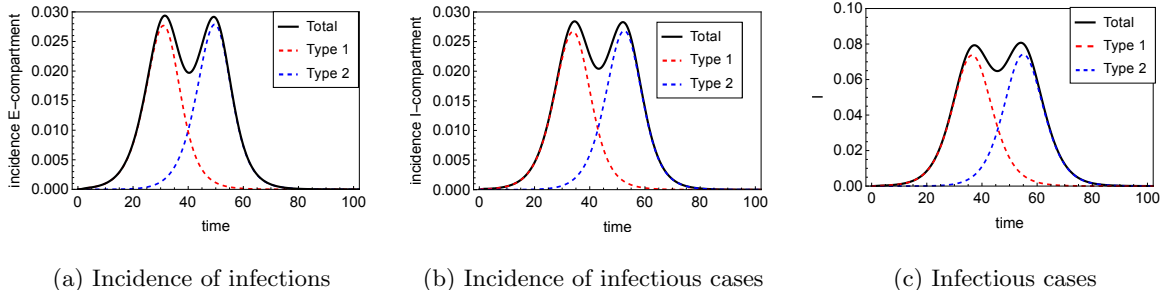


Figure 2: **Incidence with heterogeneity due to weak coupling.** We model two groups of equal size and with weak coupling between the two groups: The within group transmission parameter is 100 times higher than the between-group transmission parameter. The groups have the same parameters (latent period Gamma-distributed with shape-parameter 3 and scale parameter 1; infectious period exponentially distributed with mean 3 for both groups), R_0 equals 3. We introduce the disease one of the groups. (a) Incidence of new infections. (b) Incidence of new infectious individuals. (c) I-compartment as fraction of the total population

10.2 How well defined is the HIT?

As shown in [37] and [29], the ∞ -dimensional (i.e., Krein-Rutman) version of Perron-Frobenius theory yields that within the positive cone there is, for the linearized problem, modulo translation only one solution growing away from the disease free steady state. Unstable manifold theory, see [20, 22], next extends this uniqueness modulo translation to the nonlinear setting. Implicitly we have used this idea when writing (7) and paying little to no attention to the initial condition. Even though we do not know a published proof, we believe that one can define the HIT unambiguously in terms of the positive unstable manifold of the disease free steady state. But, armed with the insights obtained in the preceding subsection, we might wonder how relevant the HIT thus defined is when coupling is only weak? To find out, imagine a small subpopulation with a high within-contact rate, very weakly coupled to a large subpopulation that has a within-contact rate large enough to have the within- R_0 bigger than 1. If we introduce the pathogen in the small subpopulation, the HIT will be reached more or less when the susceptible fraction of the large subpopulation reaches $1/\text{within-}R_0$. But if we introduce the pathogen in the large subpopulation, then upon reaching $1/\text{within-}R_0$ the small subpopulation will still have its susceptible fraction ABOVE its own $1/\text{within-}R_0$. While we wait for the susceptible fraction of the small subpopulation to reach this critical level, an overshoot happens in the large subpopulation. So we expect that in this second scenario the overall susceptible fraction upon reaching the overall HIT is smaller than in the first scenario. Figure 3 illustrates that this can indeed happen. In conclusion: for the same model, the HIT may differ substantially depending on the precise details of the introduction, in particular the subpopulation in which the small introduction occurs (we reiterate: the distinction between one population and several populations is not as clear cut as the mathematical definition of irreducibility seems to suggest at first).

10.3 A simple yet illuminating example (showing, among other things, that there is no clear relationship between R_0 and final size)

Let Ω consist of just two points, labeled 1 and 2. Following notational custom, we now represent the last two arguments of A as indices. So $A_{ij}(\tau)$ is the expected force of infection exerted on an individual

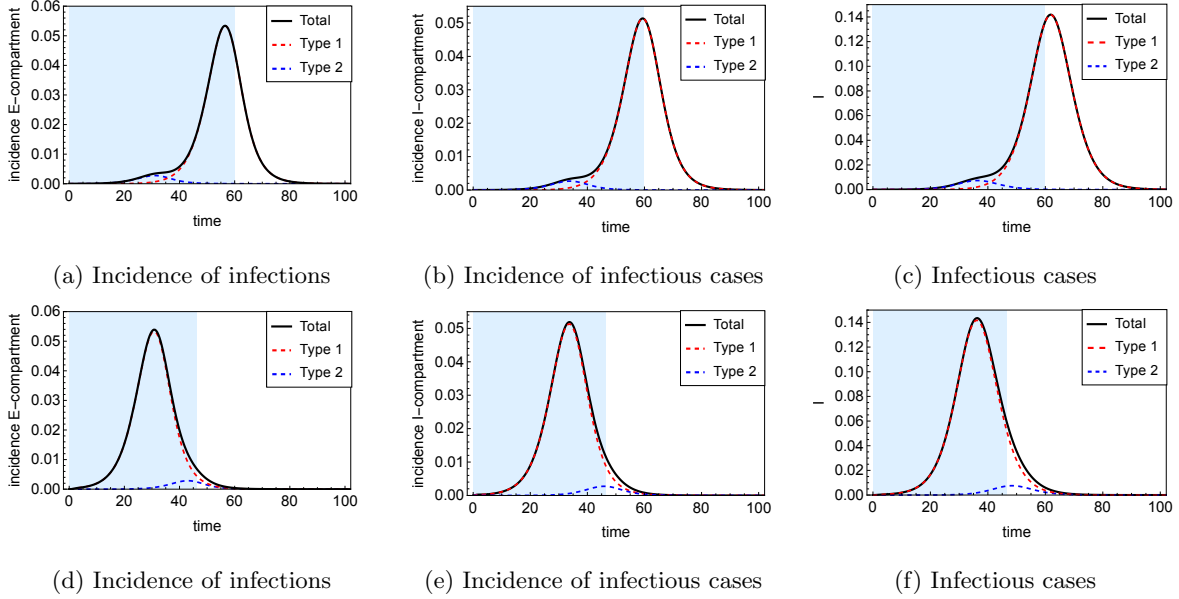


Figure 3: **HIT in case of a small ‘isolated’ community.** We model two groups, comprising 5% and 95% of the population. 99% of contacts of individuals in the small group are with other individuals in the small group. The groups have the same parameters (the latent period is Gamma-distributed with shape-parameter 3 and scale parameter 1, the infection period is exponentially distributed with mean 3). In the initial phase, the expected number of secondary cases per primary case is 3, irrespective of the group. Initially there is no immunity. In (a), (b), and (c) a fraction 0.001 of the small population is infectious while there are initially no infectious individuals in the large population. In (d), (e), and (f) a fraction 0.001 of the large population is infectious while there are initially no infectious individuals in the small population. The shading changes at the HIT. (a) and (d): Incidence of new infections. (b) and (e): Incidence of new infectious individuals. (c) and (f): I-compartment as fraction of the total population

of type i by an individual of type j that was infected time τ ago. Similarly we denote N times the fraction of the population that is of type i by N_i .

The 2×2 matrix \mathcal{K}_0 is accordingly defined by

$$(\mathcal{K}_0)_{ij} = N_j \int_0^\infty A_{ij}(\tau) d\tau \quad (116)$$

for $i, j \in \{1, 2\}$.

As explained in Section 3, \mathcal{K}_0 is the NGM in terms of fractions. The more traditional NGM in terms of numbers is K with

$$K_{ij} = N_i \int_0^\infty A_{ij}(\tau) d\tau \quad (117)$$

and the two are related by (20) with T the scaling of the axes given by

$$Tx = \begin{pmatrix} N_1 x_1 \\ N_2 x_2 \end{pmatrix} \quad (118)$$

When (i) the type j only affects the contact intensity c_j and (ii) a fraction $c_j N_j / (c_1 N_1 + c_2 N_2)$ of

the contacts of any individual is with individuals of type j , we are in the separable mixing situation described in the Sections 5 and 6 and

$$A_{ij}(\tau) = \frac{c_i c_j}{c_1 N_1 + c_2 N_2} b(\tau) \quad (119)$$

In this case the range of \mathcal{K}_0 is spanned by $\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ and this vector therefore is the eigenvector corresponding to the unique non-zero eigenvalue

$$R_0 = \frac{c_1^2 N_1 + c_2^2 N_2}{c_1 N_1 + c_2 N_2} \int_0^\infty b(\tau) d\tau \quad (120)$$

In terms of the distribution of c in the population, the first factor can be written as

$$\text{mean} + \text{variance}/\text{mean}$$

(this classical result was important in the context of HIV, as it showed that ignoring the variance of the (homo)sexual activity distribution leads to a wrong estimate of R_0 ; the underlying reason is that very active individuals are BOTH more susceptible AND more infectious). It may happen that $N_1 \ll N_2$ but $c_1 \gg c_2$. In such a situation the type 1 subpopulation forms a core group of superspreaders, meaning that it is a small group contributing heavily to transmission. Note that the distribution of the two types is:

$$\begin{array}{ll} N_1 : N_2 & \text{among all individuals,} \\ c_1 N_1 : c_2 N_2 & \text{in the incidence during the initial phase of the outbreak,} \\ c_1^2 N_1 : c_2^2 N_2 & \text{among the infectors during the initial phase.} \end{array}$$

A simple computation shows that

$$\frac{\partial R_0}{\partial c_2} < 0 \text{ if } \frac{N_2}{N_1} < \frac{c_1}{c_2} \left(\frac{c_1}{c_2} - 2 \right) \quad (121)$$

At first it might seem counterintuitive that R_0 can DECREASE when individuals of a subgroup INCREASE their contact rate. But once one realises that a side effect might be that the members of the core group have less WITHIN-core-group contacts, it should be clear how the intuition should be up-dated. The relation (41) now takes the form

$$s_i(t) = e^{-c_i w(t)} \quad (122)$$

while (42) boils down to

$$w(t) = \int_0^\infty b(\tau) \left(\frac{c_1 N_1}{c_1 N_1 + c_2 N_2} (1 - s_1(t - \tau)) + \frac{c_2 N_2}{c_1 N_1 + c_2 N_2} (1 - s_2(t - \tau)) \right) d\tau \quad (123)$$

combined with (122). In both of these relations we can take the limit $t \rightarrow \infty$. This yields an equation for $w(\infty)$ and the identities

$$s_i(\infty) = e^{-c_i w(\infty)} \quad (124)$$

and from these we can compute the overall final escape fraction

$$s(\infty) := \frac{s_1(\infty) N_1 + s_2(\infty) N_2}{N_1 + N_2} \quad (125)$$

In the special case $c_1 = c = c_2$ there is, after all, no heterogeneity. In the special case $c_1 = c$, $c_2 = 0$ the infection only circulates in subpopulation 1. In both special cases we have

$$R_0 = c \int_0^\infty b(\tau) d\tau \quad (126)$$

and the equation

$$w(\infty) = \int_0^\infty b(\tau) d\tau \left(1 - e^{-cw(\infty)}\right) \quad (127)$$

showing that $w(\infty)$ is the same in these two cases.

On the other hand, we have

$$s(\infty) = 1 - \frac{w(\infty)}{\int_0^\infty b(\tau) d\tau} \quad (128)$$

in the homogeneous case (which allows to rewrite (127) in the more familiar form $s(\infty) = e^{-R_0(1-s(\infty))}$), while in the ‘disconnected’ case we find

$$s(\infty) = \frac{N_1}{N_1 + N_2} \left(1 - \frac{w(\infty)}{\int_0^\infty b(\tau) d\tau}\right) + \frac{N_2}{N_1 + N_2} \quad (129)$$

clearly showing that for small N_1 almost the entire population ‘escapes’ infection, simply since almost all individuals belong to the isolated type 2 subpopulation. By continuity we obtain a similar situation for small values of c_2 : a small core group of superspreaders has a large impact on R_0 but, due to its smallness, not necessarily on the size of the outbreak. We conclude that heterogeneity can ‘destroy’ the monotone relationship between R_0 and final size.

11 Summary and Discussion

In 1927, inspired by work of Sir Ronald Ross and Hilda Hudson [62, 63], William Ogilvy Kermack and Anderson Gray McKendrick [42] introduced a simple yet very powerful idea into the Mathematical Epidemiology of Infectious Diseases: they took as the key model ingredient a description of how a newly infected individual is expected to contribute to the future force of infection on other individuals. More precisely, they employed a function A of a time variable τ , with A the expected contribution to the force of infection and τ the time elapsed since exposure (often τ is called ‘infection-age’). In exactly this spirit we here incorporate host heterogeneity by taking as the starting point a function A of three variables, x , τ and ξ , where A and τ keep their interpretation, while x specifies the trait of the individual subjected to the force of infection and ξ specifies the trait of the infected individual. Except for a brief excursion in Appendix B, we assume that the trait of an individual is a static characteristic. We use a measure Φ to describe the trait distribution in the host population, thus unifying models with a subdivision into finitely many discrete traits (then usually called ‘types’) and models incorporating a continuum of traits.

A first aim of our work is to show that many familiar qualitative features of epidemic outbreak models can easily be characterized in this general setting. A highlight, in our opinion, is the formulation of the final size equation in terms of the Next Generation Operator acting on fractions (a precursor of this result can be found in V. Andreasen’s paper [2], but here we make it more explicit in a much more general setting). Concerning quantitative aspects, our recommendation is to resort to a discrete time formulation as in [25], but we do not elaborate on this.

A second aim of our work is to reveal the simplification that occurs when the function A of three variables is actually the product of three functions of one variable, i.e., in case of separable mixing, as it is called in modelling terms. The main result is that in this case we are essentially back to the scalar ‘homogeneous’ case, but with a modified nonlinearity that incorporates the effects of the heterogeneity. Triggered by the Covid-19 outbreak, there has recently been a lot of attention for the influence of heterogeneity on the Herd Immunity Threshold (HIT). The scalar character that results from separable mixing greatly facilitates the characterization of the HIT. When the trait space is $(0, \infty)$ and Φ is the Gamma Distribution one even obtains explicit expressions showing that the HIT is quickly reached when the variance of the trait distribution is high, a point stressed by G. Gomes and others in [11, 31, 51, 53, 76]. Here we showed that this holds for general τ dependence of the infectiousness (and not just for SIR or SEIR compartmental models; please note that, as shown in the companion paper [26], our general framework allows to derive in a rather easy way separable mixing heterogeneous variants of these and other compartmental models). Not surprisingly, the effect of heterogeneity on the HIT is reduced when the trait is dynamic, see [77] and Appendix B.

In order to specify the model ingredient A in more detail, one has to reflect on the influence of host population size and composition on contact intensities. This gives rise to a highly nontrivial modelling difficulty: if we have information about the contact structure for a certain host population size and composition, how can we extrapolate this information to situations in which the size and composition are different? A third aim of our work is to call attention to this challenge, partly by reviving the Heesterbeek-Metz pair formation model introduced in [33]. Our main contribution to this topic, is to present it as an open problem!

Acknowledgements

We thank Horst Thieme and an anonymous referee for helpful suggestions to improve the exposition.

References

- [1] L. Almeida, P. Bliman, G. Nadin, B. Perthame and N. Vauchelet, *Final size and convergence rate for an epidemic in heterogeneous populations*, Math. Models Methods Appl. Sci. **31**(2021), 1021-1051.
- [2] V. Andreasen, *The Final Size of an Epidemic and Its Relation to the Basic Reproduction Number*, Bull. Math. Biol. **73**(2011), 2305-2321.
- [3] F. Avram, R. Adenane, L. Basnarkov, G. Bianchin, D. Goreac and A. Halanay, *An Age of Infection Kernel, an R Formula, and Further Results for Arino-Brauer A, B Matrix Epidemic Models with Varying Populations, Waning Immunity, and Disease and Vaccination Fatalities.*, Mathematics **11**(2023):1307.
- [4] F. Ball, L. Critcher, P. Neal and D. Sirl, *The impact of household structure on disease-induced herd immunity*. J. Math. Biol. **87**, 83 (2023), 83.
- [5] C. Barril, P. Bliman and S. Cuadrado, *Final Size for Epidemic Models with Asymptomatic Transmission*, Bull. Math. Biol. **85**(2023), 52.
- [6] A. Bátkai, M. Kramar Fijavž and A. Rhandi, *Positive Operator Semigroups : from Finite to Infinite Dimensions*, Birkhäuser, Basel, 2017.

- [7] H. Berestycki, B. Desjardins, J. Weitz and J. Oury, *Epidemic modeling with heterogeneity and social diffusion*, J. Math. Biol. **86**(2023), 60.
- [8] M. Bootsma, B. Kolen, M. de Vries, M. Kretzschmar, and N. Mouter. *Effectiviteit van verschillende toepassingen van het Coronatoegangsbewijs* (in Dutch)(2022) <https://open.overheid.nl/repository/ronl-aa2988ce-324f-4c1c-ad39-459358e32bfe/1/pdf/effectiviteit-coronatoegangsbewijs-eindversie-tu-delft4.pdf>
- [9] M. Bootsma, K. Chan, O. Diekmann and H. Inaba. *Separable mixing: The general formulation and a particular example focusing on mask efficiency*, Math. Biosci. Eng. **20**(2023), 10.
- [10] F. Brauer and J. Watmough, *Age of infection epidemic models with heterogeneous mixing*, J. Biol. Dyn., **3**(2009), 324-330.
- [11] T. Britton, F. Ball and P. Trapman, *A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV2*, Science **369**(2020), 846-849.
- [12] D Breda, O Diekmann, W. de Graaf, A Pugliese and R Vermiglio, *On the formulation of epidemic models (an appraisal of Kermack and McKendrick)*, J. Biol. Dyn. **6**(2012), 103-117.
- [13] D. Breda, F. Florian, J. Ripoll and R. Vermiglio, *Efficient numerical computation of the basic reproduction number for structured populations*, J. Comp. Appl. Math. **384**(2021), 113165.
- [14] D. Breda, T. Kuniya, J. Ripoll and R. Vermiglio, *Collocation of next-generation operators for computing the basic reproduction number of structured populations*, J. Sci. Comput. **85**(2020), 40.
- [15] D. Breda, S. De Reggi, F. Scarabel, R. Vermiglio and J. Wu, *Bivariate collocation for computing R_0 in epidemic models with two structures*, submitted.
- [16] M. Castro, S. Ares, J. A. Cuesta and S. Manrubia, *The turning point and end of an expanding epidemic cannot be precisely forecast*, Proc. Natl. Acad. Sci. **117**(2020), 26190-26196.
- [17] K. Chan, *Impact of demographical change on the severity of an epidemic*, Master thesis, Utrecht Univeristy, 2013, <https://studenttheses.uu.nl/handle/20.500.12932/13069>.
- [18] K. Chan, H. Nishiura, O Diekmann, M. Bootsma, *The impact of population ageing on the severity of an epidemic outbreak*, unpublished manuscript, (2016).
- [19] J. Cui, Y. Sun and Huaiping Zhu, *The impact of media on the control of infectious diseases*, J. Dyn. Diff. Equa. **20**(2008), 31-53.
- [20] O. Diekmann, Ph. Getto and M. Gyllenberg, *Stability and bifurcation analysis of Volterra functional equations in the light of suns and stars*, SIAM J. Math. Anal. **39**(2008), 1023-1069.
- [21] O. Diekmann and M. Gyllenberg, *Abstract delay equations inspired by population dynamics*, In : *Functional Analysis and Evolution Equations. The Günter Lumer Volume*. Birkhauser, Basel, 2007, 187-200.
- [22] O. Diekmann and M. Gyllenberg, *Equations with infinite delay: blending the abstract and the concrete*, J. Differ. Equ. **252**(2011), 819-851.
- [23] O. Diekmann, J. Heesterbeek and T. Britton, *Mathematical Tools for Understanding Infectious Disease Dynamics*, Princeton University Press, Princeton, 2013.

- [24] O. Diekmann, M. Gyllenberg and J. Metz, *Finite Dimensional State Representation of Linear and Nonlinear Delay Systems* J. Dyn. Diff. Equat. **30**(2018), 1439-1467.
- [25] O. Diekmann, H. Othmer, R. Planqué and M. Bootsma, *The discrete-time Kermack-McKendrick model: A versatile and computationally attractive framework for modeling epidemics*, Proc. Natl. Acad. Sci. **118**(2021), e2106332118.
- [26] O. Diekmann and H. Inaba, *A systematic procedure for incorporating separable static heterogeneity into compartmental epidemic models*, J. Math. Biol. **86**(2023), 29.
- [27] A. d’Onofrio and P. Manfredi, *Information-related changes in contact patterns may trigger oscillations in the endemic prevalence of infectious diseases*, J. Theor. Biol. **256**(2009), 473-478.
- [28] S. Eikenberry, M. Mancuso, E. Iboi, T. Phan, K. Eikenberry, Y. Kuang, E. Kostelich and A. Gumelet, *To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic*, Infect. Dis. Model. **5**2020, 293-308.
- [29] E. Franco, O. Diekmann and M. Gyllenberg, *Modelling physiologically structured populations: renewal equations and partial differential equations*, J. Evol. Equ. **23**(2023), 46.
- [30] S. Funk, M. Salathé and V. Jansen, *Modelling the influence of human behaviour on the spread of infectious diseases: a review*, J. R. Soc. Interface **7**(2010), 1247-1256.
- [31] M. Gomes, M. Ferreira, R. Corder, J. King, C. Souto-Maior, C. Penha-Gonçalves, G. Gonçalves, M. Chikina, W. Pegden and R. Aguas, *Individual variation in susceptibility or exposure to SARS-CoV-2 lowers the herd immunity threshold*, J. Theor. Biol. **540**(2022), 111063.
- [32] K. Hadeler, *Pair formation*, J. Math. Biol. **64**(2012), 613-645.
- [33] J. Heesterbeek and J. Metz, *The saturating contact rate in marriage and epidemic models*, J. Math. Biol. **31**(1993), 529-539.
- [34] A. Hill, J. Glasser and Z. Feng, *Implications for infectious disease models of heterogeneous mixing on control thresholds*, J. Math. Biol. **86**(2023), 53.
- [35] H. Inaba, *On a new perspective of the basic reproduction number in heterogeneous environments*, J. Math. Biol. **65**(2012), 309-348.
- [36] H. Inaba, *On a pandemic threshold theorem of the early Kermack-McKendrick model with individual heterogeneity*, Math. Popul. Stud. **21**(2014), 95-111.
- [37] H. Inaba, *Age-Structured Population Dynamics in Demography and Epidemiology*, Springer, Singapore, 2017.
- [38] H. Inaba, *Basic concepts for the Kermack and McKendrick model with static heterogeneity*, arXiv, preprint, 2023, <https://arxiv.org/abs/2311.11247>.
- [39] D. Juher, D. Rojas and J. Saldaña, *Saddle-node bifurcation of limit cycles in an epidemic model with two levels of awareness*, Phys. D: Nonlinear Phenom. **448**(2023), 133714.
- [40] W. Just, J. Saldaña and Y. Xin, *Oscillations in epidemic models with spread of awareness*, J. Math. Biol. **76**(2018), 1027-1057.

- [41] G. Katriel, *The size of epidemics in populations with heterogeneous susceptibility*, J. Math. Biol. **65**(2012), 237-262.
- [42] W. Kermack and A. McKendrick, *A contribution to the mathematical theory of epidemics*, Proc. R. Soc. Lond. A **115**(1927), 700-721.
- [43] X. Li, J. Yang and M. Martcheva, *Age Structured Epidemic Modeling*, Springer, Cham, (2020).
- [44] R. Liu, J. Wu and H. Hu, *Media/psychological impact on multiple outbreaks of emerging infectious diseases*, Comput. Math. Methods Med. **8**(2007), 153-164.
- [45] S. Manrubia, *The Uncertain Future in How a Virus Spreads*, Physics **13**(2020),166.
- [46] E. Messina, M. Pezzella and A. Vecchio, *Positive numerical approximation of an integro-differential epidemic model*, Axioms **11**(2022), 69.
- [47] E. Messina, M. Pezzella and A. Vecchio, *A non-standard numerical scheme for an age-of-infection epidemic model*, J. Comput. Dyn. **9**(2022), 239-252.
- [48] E. Messina, M. Pezzella and A. Vecchio, *A long-time behavior preserving numerical scheme for age-of-infection epidemic models with heterogeneous mixing*, Appl. Numer. Math. (2023).
- [49] E. Messina, M. Pezzella and A. Vecchio, *Asymptotic solutions of non-linear implicit Volterra discrete equations*, J. Comput. Appl. Math. **425**(2023), 115068.
- [50] E. Messina, M. Pezzella and A. Vecchio. *Nonlocal finite difference discretization of a class of renewal equation models for epidemics*, Math. Biosci. Eng. **20**(2023), 11656-11675.
- [51] A. Montalbán, R. Corder and M. Gomes, *Herd immunity under individual variation and reinfection*, J. Math. Biol. **85**(2022), 2.
- [52] J. Mossong, N. Hens, M. Jit, Ph. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. Scalia Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska and W. Edmunds *Social contacts and mixing patterns relevant to the spread of infectious diseases*, PLOS Medicine **5**(2008), e74.
- [53] J. Neipel, J. Bauermann, S. Bo, T. Harmon and F. Jülicher, *Power-Law Population Heterogeneity Governs Epidemic Waves*, PLoS ONE **15**(2020), e0239678.
- [54] C. Ngonghala, E. Iboi, S. Eikenberry, M. Scotch, C. MacIntyre, M. Bonds and A. Gumel, *Mathematical assessment of the impact of non-pharmaceutical interventions on curtailing the 2019 novel Coronavirus*, Math. Biosci. **325**(2020).
- [55] M. Nguyen, A. Freedman, S. Ozbay and S. Levin, *Power-Law Population Heterogeneity Governs Epidemic Waves*, submitted.
- [56] A. Novozhilov, *On the spread of epidemics in a closed heterogeneous population* Math. Biosci. **215**(2008), 177-185.
- [57] G. Oliva, S. Bonfigli, P. Cavallo and A. Scala, *Navigating the Herd Immunity Surface: A Novel Framework for Optimising Epidemic Response Strategies*, Qeios (2023).
- [58] R. Pastor-Satorras and C. Castellano, *The advantage of self-protecting interventions in mitigating epidemic circulation at the community level*, Sci. Rep. **12**(2022).

- [59] P. Poletti, B. Caprile, M. Ajelli, A. Pugliese and S. Merler, *Spontaneous behavioural changes in response to epidemics*, J. Theor. Biol. **260**(2009), 31-40.
- [60] L. Rass and J. Radcliffe, *Spatial Deterministic Epidemics*, American Mathematical Society, Providence, RI, 2003.
- [61] F. Scarabel, O. Diekmann and R. Vermiglio, *Numerical bifurcation analysis of renewal equations via pseudospectral approximation*, J. Comput. Appl. Math. **397**(2021), 113611.
- [62] R. Ross, *An application of the theory of probabilities to the study of a priori pathometry*, Proc. R. Soc. A **92**(1916), 204-230 (Part I).
- [63] R. Ross and H. Hudson, *An application of the theory of probabilities to the study of a priori pathometry* Proc. R. Soc. A **93**(1917), 212-225 (Part II), 225-240 (Part III).
- [64] L. Van Schaik, D. Duives, S. Hoogendoorn-Lanser, J. Hoekstra, W. Daamen, A. Gavriilidou, P. Krishnakumari, M. Rinaldi and S.P. Hoogendoorn. *Understanding physical distancing compliance behaviour using proximity and survey data: A case study in the Netherlands during the COVID-19 pandemic* Transportation Research Procedia (in press).
- [65] B. Schmidt, *Die epidemische Dynamik von Infektionskrankheiten mit multiplem Krankheitsverlauf in strukturierten Bevölkerungen Mathematische Modellierung, Sensitivitätsanalyse und numerische Simulation der Ausbreitung von AIDS*, PhD thesis, Universität zu Köln, 1990.
- [66] B. Tang, W. Zhou, X. Wang, H. Wu and Y. Xiao, *Controlling multiple Covid-19 waves: an insight from a multi-scale model linking the behaviour change dynamics to disease transmission dynamics*, Bull. Math. Biol. **84**(2022), 106.
- [67] A. Teslya, H. Nunner, V. Buskens and M. Kretzschmar, *The effect of competition between health opinions on epidemic dynamics*, Proc. Natl. Acad. Sci. Nexus **1**(2022), 1-14.
- [68] H. Thieme, *On a class of Hammerstein integral equations*, Manuscr. math. **29**(1979), 49-84.
- [69] H. Thieme, *On the boundedness and the asymptotic behaviour of the non-negative solutions to Volterra-Hammerstein integral equations.*, Manuscr. math. **31**(1980), 379-412.
- [70] H. Thieme, *Renewal theorems for linear periodic Volterra integral equations*, J. Integral Equations **7**(1984), 253-277.
- [71] H. Thieme, *Renewal theorems for some mathematical models in epidemiology*, J. Integral Equations **8**(1985), 185-216.
- [72] H. Thieme, *Distributed susceptibility: a challenge to persistence theory in infectious disease models*, Disc. Cont. Dyn. Sys. B **12**(2009), 865-864.
- [73] H. Thieme, *Spectral bound and reproduction number for infinite-dimensional population structure and time heterogeneity*, SIAM J. Appl. Math. **70**(2009), 188-211.
- [74] H. Thieme and J. Yang, *On the complex formation approach in modeling predator prey relations, mating, and sexual disease transmission*, Elect. J. Diff. Eqns. **5**(2000), 255-283.
- [75] Y. Tian, A. Sridhar, C. Wu, S. Levin, K. Carley, H. Poor and O. Yağan, *Role of masks in mitigating viral spread on networks*, Phys. Rev. E **108**(2023).

- [76] A. Tkachenko, S. Maslov, A. Elbanna, G. Wong, Z. Weiner and N. Goldenfeld, *Time-dependent heterogeneity leads to transient suppression of the COVID-19 epidemic, not herd immunity*, Proc. Natl. Acad. Sci. **118**(2021), e2015972118.
- [77] A. Tkachenko, S. Maslov, T. Wang, A. Elbanna, G. Wong and N. Goldenfeld, *Stochastic social behavior coupled to COVID-19 dynamics leads to waves, plateaus, and an endemic state*, eLife (2021) <https://doi.org/10.7554/eLife.68341>].
- [78] M. Toorians, A. MacPherson and T. Davies, *Revisiting pathogen transmission in epidemiological models and its role in the disease-diversity relation*, Preprints (2021), 2021100295. <https://doi.org/10.20944/preprints202110.0295.v4>.
- [79] J. Wambua, N. Loedy, C. Jarvis, K. Wong, C. Faes, R. Grah, B. Prasse, F. Sandmann, R. Niehus, H. Johnson, W. Edmunds, P. Beutels, N. Hens and P. Coletti, *The influence of COVID-19 risk perception and vaccination status on the number of social contacts across Europe: insights from the CoMix study*, BMC Public Health **23**(2023).
- [80] G. Wong, Z. Weiner, A. Tkachenko, A. Elbanna, S. Maslov and N. Goldenfeld, *Modeling COVID-19 Dynamics in Illinois under Nonpharmaceutical Interventions*, Phys. Rev. X **10**(2020), 041033.
- [81] X. Zhang, F. Scarabel, K. Murty and J. Wu, *Renewal equations for delayed population behaviour adaptation coupled with disease transmission dynamics: A mechanism for multiple waves of emergent infections*, Math. Biosci. **365**(2023).

A Mathematical statements

In this appendix we prove some elementary auxiliary results that are used in the main text. Throughout this section z denotes a scalar.

Lemma A.

- i) $1 - e^{-z} \leq z$ for $z \geq 0$.
- ii) $1 - e^{-z} \geq \frac{1 - e^{-z_0}}{z_0} z$ for $z_0 > 0$ and $0 \leq z \leq z_0$.
- iii) $\forall \epsilon > 0 \exists \delta = \delta(\epsilon)$ such that $1 - e^{-z} \geq (1 - \epsilon)z$ for $0 \leq z \leq \delta(\epsilon)$.
- iv) $1 - e^{-\theta z} > \theta(1 - e^{-z})$ for $z > 0$ and $0 < \theta < 1$.
- v) $(1 - z) \log(1 - z) + \rho z > 0$ for $0 < z < 1$ and $\rho > 1$.

Proof.

- i) Let $h(z) := 1 - e^{-z} - z$. Then $h(0) = 0$ and $h'(z) = e^{-z} - 1 < 0$ for $z > 0$.
- ii) Define $h(z) := 1 - e^{-z} - \frac{1 - e^{-z_0}}{z_0} z$ then $h(0) = 0 = h(z_0)$ and $h'(z) = e^{-z} - \frac{1 - e^{-z_0}}{z_0}$, $h''(z) = -e^{-z} < 0$. We note that $h'(0) = 1 - \frac{1 - e^{-z_0}}{z_0} > 0$ (see the proof of Lemma A.i), while $h'(z_0) = \frac{(z_0 + 1)e^{-z_0} - 1}{z_0} < 0$. It follows that $h(z) > 0$ for $0 < z < z_0$.
- iii) Let now $h(z) := 1 - e^{-z} - (1 - \epsilon)z$. Then $h(0) = 0$ and $h'(z) = e^{-z} - (1 - \epsilon)$. So $h'(0) = \epsilon > 0$ and consequently h is positive for small positive z .

- iv) Define, for given $z > 0$, $H(\theta) := 1 - e^{-\theta z} - \theta(1 - e^{-z})$. Then H is continuous, $H(0) = 0$, $H(1) = 0$, $H'(\theta) = ze^{-\theta z} - 1 + e^{-z}$ and $H''(\theta) = -z^2 e^{-\theta z} < 0$. So H cannot have two or more zero's in between 0 and 1. Since $H'(0) = z - 1 + e^{-z} > 0$ (see Lemma A.i)), H is positive for $0 < \theta < 1$.
- v) Let $h(z) := (1 - z) \log(1 - z) + \rho z$, then $h(0) = 0$ and $h'(z) = -1 - \log(1 - z) + \rho > 0$ for $0 \leq z < 1$, so h is positive on $[0, 1]$.

□

B Dynamic heterogeneity

So far we considered a host population consisting of individuals with different STATIC traits, where ‘static’ refers to the assumption that individuals do not change trait as the pathogen spreads through the population. The aim of this appendix is to warn readers that results may change significantly when the trait itself is a dynamic variable.

We distinguish:

- 1 models with trait-dynamics *not* influenced by disease status,
- 2 models with trait-dynamics influenced by disease status,

with the second class of models typically requiring more model assumptions and exhibiting more complex dynamics.

B.1 dynamic heterogeneity: *not* influenced by disease status

We introduce a model with the goal to study the impact of dynamic heterogeneity in a framework where dynamics are not influenced by disease status. Individuals are characterized by the number of contacts they make per unit of time. Type 1 has contact rate c_1 , type 2 has contact rate c_2 . Transitions occur at rates

$$\nu \begin{pmatrix} -\theta & \frac{\theta}{\theta-1} \\ \theta & -\frac{\theta}{\theta-1} \end{pmatrix}, \quad (130)$$

with $\nu > 0$ the frequency (note that a round trip takes on average $\frac{1}{\nu\theta} + \frac{\theta-1}{\nu\theta} = \frac{1}{\nu}$) and $\theta > 1$ a measure for the asymmetry in the sojourn times, and hence for the asymmetry of the two steady state subpopulations. We assume the subpopulations to be in their steady states. Hence, if N denotes the total population size and N_1 and N_2 the subpopulation sizes then

$$\begin{pmatrix} N_1 \\ N_2 \end{pmatrix} = \frac{N}{\theta} \begin{pmatrix} 1 \\ \theta - 1 \end{pmatrix}. \quad (131)$$

The average contact rate c is hence given by

$$c = \frac{c_1}{\theta} + \frac{\theta - 1}{\theta} c_2 = c_2 + \frac{c_1 - c_2}{\theta}. \quad (132)$$

As a normalization, we fix c and require that $c_2 \leq c$.

It follows that there are three ‘free’ parameters: $\nu > 0$, $\theta > 1$, $0 \leq c_2 \leq c$. We assume proportionate mixing: if an individual makes a contact, it is with probability

$$\frac{c_1 N_1}{c_1 N_1 + c_2 N_2} \quad (133)$$

with an individual of type 1.

B.1.1 The SIR model

We will combine the ‘dynamic heterogeneous’ ingredients in the previous section with a homogeneous version of the epidemic model, the SIR model, as described by

$$\begin{aligned}\frac{dS}{dt} &= -\beta c \frac{SI}{N} \\ \frac{dI}{dt} &= \beta c \frac{SI}{N} - \alpha I,\end{aligned}\tag{134}$$

with S and I the amount of susceptible and infectious individuals respectively. Here β with $0 < \beta \leq 1$, is the probability of transmission in a contact between an infectious and a susceptible individual. The expected duration of the infectious period equals $\frac{1}{\alpha}$. By scaling of the time variable, we achieve that $\alpha = 1$. By scaling of c we achieve that $\beta = 1$. We now list some relevant features:

$$\text{basic reproduction number } R_0 = c,\tag{135}$$

$$\text{herd immunity threshold (HIT) } \bar{s} = \frac{\bar{S}}{N} = \frac{1}{R_0} = \frac{1}{c},\tag{136}$$

$$\text{final size equation } s(\infty) = e^{-c(1-s(\infty))}.\tag{137}$$

We define the overshoot as $\bar{s} - s(\infty)$.

B.1.2 The combined model

Recall that we assume that type transitions are not influenced by disease status and that $\alpha = 1$ and $\beta = 1$. Recall that

$$c = c_1 \frac{N_1}{N} + c_2 \frac{N_2}{N} = \frac{c_1}{\theta} + (1 - \frac{1}{\theta})c_2.\tag{138}$$

Keep in mind that if a contact is with a type i individual, it is with probability $\frac{S_i}{N_i}$ with a susceptible individual. The combined model is defined as follows

$$\begin{aligned}\frac{dS_1}{dt} &= -\frac{c_1 S_1}{c_1 N_1 + c_2 N_2} (c_1 I_1 + c_2 I_2) - \nu \theta S_1 + \nu \frac{\theta}{\theta - 1} S_2 \\ \frac{dS_2}{dt} &= -\frac{c_2 S_2}{c_1 N_1 + c_2 N_2} (c_1 I_1 + c_2 I_2) + \nu \theta S_1 - \nu \frac{\theta}{\theta - 1} S_2 \\ \frac{dI_1}{dt} &= \frac{c_1 S_1}{c_1 N_1 + c_2 N_2} (c_1 I_1 + c_2 I_2) - \nu \theta I_1 + \nu \frac{\theta}{\theta - 1} I_2 - I_1 \\ \frac{dI_2}{dt} &= \frac{c_2 S_2}{c_1 N_1 + c_2 N_2} (c_1 I_1 + c_2 I_2) + \nu \theta I_1 - \nu \frac{\theta}{\theta - 1} I_2 - I_2\end{aligned}\tag{139}$$

Define $s_j = \frac{S_j}{N}$ and $i_j = \frac{I_j}{N}$. Then

$$\begin{aligned}\frac{ds_1}{dt} &= -\frac{c_1}{c} (c_1 i_1 + c_2 i_2) s_1 - \nu \theta s_1 + \nu \frac{\theta}{\theta - 1} s_2 \\ \frac{ds_2}{dt} &= -\frac{c_2}{c} (c_1 i_1 + c_2 i_2) s_2 + \nu \theta s_1 - \nu \frac{\theta}{\theta - 1} s_2 \\ \frac{di_1}{dt} &= \frac{c_1}{c} (c_1 i_1 + c_2 i_2) s_1 - \nu \theta i_1 + \nu \frac{\theta}{\theta - 1} i_2 - i_1 \\ \frac{di_2}{dt} &= \frac{c_2}{c} (c_1 i_1 + c_2 i_2) s_2 + \nu \theta i_1 - \nu \frac{\theta}{\theta - 1} i_2 - i_2\end{aligned}\tag{140}$$

B.1.3 Static heterogeneity

In the combined model of the previous subsection we put $\nu = 0$, but keep (131). The resulting model is one with static heterogeneity. If we freeze the values s_1 and s_2 and introduce $x := c_1 i_1 + c_2 i_2$ (as a metric for the subpopulation of infectious individuals), we obtain

$$\frac{dx}{dt} = \left(\frac{c_1^2}{c} s_1 + \frac{c_2^2}{c} s_2 - 1 \right) x. \quad (141)$$

We conclude that

$$R_0 = \frac{c_1^2}{c} \frac{1}{\theta} + \frac{c_2^2}{c} \left(1 - \frac{1}{\theta}\right) \quad (142)$$

and that the HIT is characterized by

$$\frac{c_1^2}{c} \bar{s}_1 + \frac{c_2^2}{c} \bar{s}_2 = 1. \quad (143)$$

The latter can be reduced to an equation for the scalar variable w , where

$$s_1(t) = \frac{1}{\theta} e^{-c_1 w(t)}, \quad s_2(t) = \left(1 - \frac{1}{\theta}\right) e^{-c_2 w(t)}. \quad (144)$$

The equation for the HIT in w reads

$$\frac{c_1^2}{c} \frac{1}{\theta} e^{-c_1 \bar{w}} + \frac{c_2^2}{c} \left(1 - \frac{1}{\theta}\right) e^{-c_2 \bar{w}} = 1, \quad (145)$$

and for that value of w the HIT, as a fraction of the total population, is given by

$$\text{HIT} : \quad \bar{s}_1 + \bar{s}_2 = \frac{1}{\theta} e^{-c_1 \bar{w}} + \left(1 - \frac{1}{\theta}\right) e^{-c_2 \bar{w}}. \quad (146)$$

The equation

$$\frac{dw}{dt} = -w + \frac{c_1}{c} \frac{1}{\theta} (1 - e^{-c_1 w}) + \frac{c_2}{c} \left(1 - \frac{1}{\theta}\right) (1 - e^{-c_2 w}), \quad (147)$$

can be easily derived by combining the defining relation (see (140))

$$\frac{dw}{dt} = \frac{c_1}{c} i_1 + \frac{c_2}{c} i_2, \quad (148)$$

with

$$\frac{ds_j}{dt} + \frac{di_j}{dt} = -i_j, \quad j = 1, 2 \quad (149)$$

(Use (149) to write the r.h.s. of (148) as a time derivative; next integrate from $-\infty$ to t and use (144) and (148)).

The limit $w(\infty)$ is accordingly characterized by

$$w(\infty) = \frac{c_1}{c} \frac{1}{\theta} (1 - e^{-c_1 w(\infty)}) + \frac{c_2}{c} \left(1 - \frac{1}{\theta}\right) (1 - e^{-c_2 w(\infty)}). \quad (150)$$

The fraction that escapes infection is given by

$$s_1(\infty) + s_2(\infty) = \frac{1}{\theta} e^{-c_1 w(\infty)} + \left(1 - \frac{1}{\theta}\right) e^{-c_2 w(\infty)}, \quad (151)$$

and the overshoot by

$$\bar{s}_1 + \bar{s}_2 - s_1(\infty) - s_2(\infty). \quad (152)$$

We obtain by combining (132) with (142)

$$R_0 = \frac{\theta}{c} \left(\left(1 - \frac{1}{\theta}\right)c_2^2 - 2c\left(1 - \frac{1}{\theta}\right)c_2 + c^2 \right). \quad (153)$$

The first and second derivative of R_0 to c_2 are

$$\frac{dR_0}{dc_2} = \frac{\theta}{c} \left(2\left(1 - \frac{1}{\theta}\right)c_2 - 2c\left(1 - \frac{1}{\theta}\right) \right) \quad (154)$$

and

$$\frac{d^2R_0}{dc_2^2} = \frac{\theta}{c} \left(2\left(1 - \frac{1}{\theta}\right) \right) \quad (155)$$

respectively. Recall $\theta > 1$, hence R_0 is a quadratic function in c_2 with strictly positive second derivative and minimum in $c_2 = c$ with value c . Thus, when $c_2 \neq c$ we have $R_0 > c$ for a model with static heterogeneity ($\nu = 0$).

B.1.4 Comparing R_0 , $s(\infty)$ and HIT in the extremes, i.e., $\nu = 0$ and $\nu = \infty$

We consider the homogeneous SIR model as describing the limit $\nu \rightarrow \infty$ (we refrain from providing a formal justification). We then have $R_0 = c$, see (135). Using the statement in the last paragraph of Section B.1.3 we can conclude that $R_0(\nu = 0) > R_0(\nu = \infty)$ for $c_2 \neq c$. Thus, when the contact rate of different types differ, the basic reproduction number is strictly higher when considered in a ($\nu = 0$) heterogeneous setting than in a ($\nu = \infty$) homogeneous setting.

For the final size $s(\infty)$ and HIT we find through numerical investigation a similar relation between $\nu = 0$ and $\nu = \infty$ as shown in Figure 4.

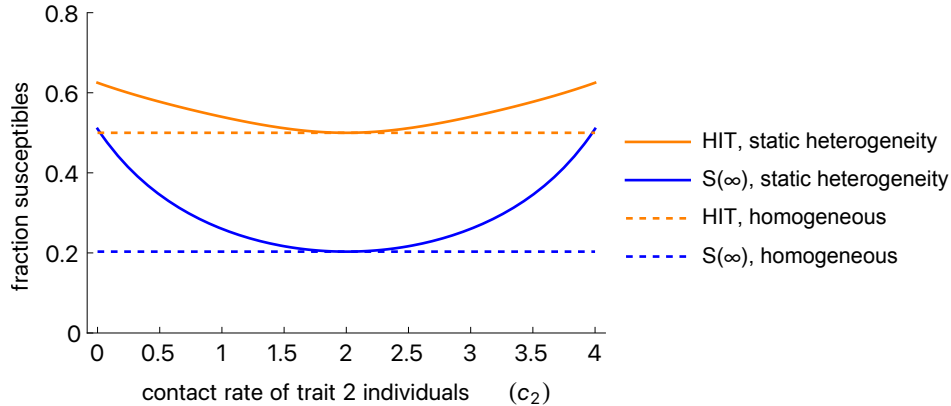


Figure 4: HIT and the final size as a function of contact rate c_2 in a static heterogeneous setting ($\nu = 0$) and in a homogeneous setting ($\nu = \infty$). $\theta = 2$. c_1 is a function of c_2 such that the average contact rate equals $c = 2$. The results for $\nu = 0$ follow from numerically solving equations (145) and (150). The results for $\nu = \infty$ follow directly from (136) and from solving (137).

B.1.5 Numerical investigation of dynamic heterogeneity ($0 < \nu < \infty$)

By interpolating the results in the extremes one may tend to conclude that R_0 , the HIT and final size are a decreasing function of the rate of trait change ν . However, numerical investigation for $0 < \nu < \infty$ shows otherwise, see Figure 5.

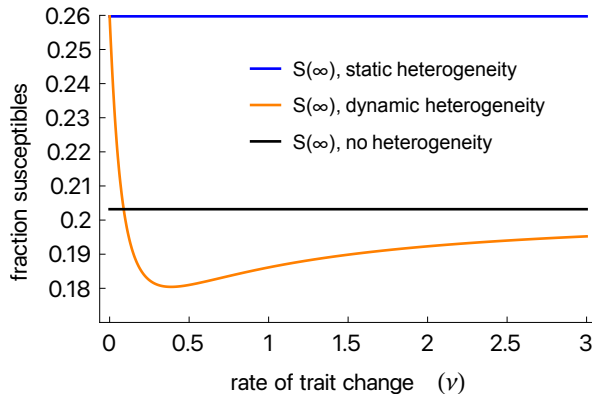


Figure 5: Final size as a function of ν in a homogeneous, a static heterogeneous and a dynamic heterogeneous setting. $\theta = 2$, $c = 2$, $c_1 = 3$, $c_2 = 1$ The results follow from numerically solving (140).

Our interpretation of Figure 5 suggests 3 types of mechanisms at play when considering dynamic heterogeneity (not influenced by disease-status).

- 1 When heterogeneity is (nearly) static, or $\nu = 0$, individuals of the group with higher contact rate infect more individuals, but also are being infected more. One would therefore expect a relatively fast and large outbreak in the group with higher contact rate. After this, the dissemination is mainly driven by individuals with lower contact rate. This causes relatively less infections among the lower contact rate individuals and eventually less infections in total.
- 2 A new mechanism emerges in addition to (1) when heterogeneity is dynamic and ν is of similar order as the rate of susceptible individuals becoming infected. While mechanism (1) still partly holds, the number of higher-contact-rate susceptible individuals decreases slowly due to ‘inflow’ of susceptible lower-contact-rate individuals. This leads to more infections overall.
- 3 When the rate of trait-change is much higher than the rate of susceptibles becoming infected, we find the impact of mechanisms (1) and (2) being diluted. This does not come as a surprise as individuals can hardly be distinguished anymore from individuals with an average contact rate c in a homogeneous model.

Our main message is that new mechanisms arise when a model is considered incorporating a dynamic heterogeneity setting instead of a static heterogeneity one. Even when the total number of infections is lower in a static heterogeneity setting than in a homogeneous one, the result can be opposite when considering from a dynamic heterogeneity setting as shown in Figure 5.

B.2 Models in which trait dynamics incorporates feedback (in particular information about incidence, prevalence and/or own disease status)

The recent outbreak of Covid-19 provides much motivation for formulating and analysing models that incorporate dynamic heterogeneity and allow the dynamics of the trait of an individual to be influenced by (information about) both the population level epidemic dynamics and its own health (and vaccination) status. We refer to [7, 39, 66, 67, 76, 81] for recent examples, to [79] for data aspects and

to [19, 27, 30, 40, 44, 59] for pre-Covid pioneering work. The subject is still in its infancy and we expect much more work in years to come.

To organize our thoughts, we first focus on trait dynamics, while assuming that, both at the individual level and at the population level, the disease related dynamics is known/given. This allows us to think of the trait as a stochastic variable that follows a Markov process in a non-constant environment, so with time-dependent transition rates. The modelling task is to specify these transition rates. Here we limit ourselves to the simpler task of indicating the variables on which the transition rates depend (without specifying the dependence itself).

The following classification attempts to structure the sea of possibilities a little bit, without claiming completeness. These rates may depend on:

1. the health status of the individual itself
2. the perceived current incidence or prevalence (with the perception being based on information; how information is handled, in particular whether it is trusted, may depend in part on the trait itself); in network models one can work with a local version based on information about the health status of acquaintances
3. governmental advice/rules (with compliance possibly depending on the trait itself)

Likewise one can describe the epidemic dynamics while pretending that the dynamic trait distribution is known. By coupling the two time-inhomogeneous dynamical systems one then obtains, through a fixed point argument, an autonomous nonlinear dynamical system. If time scale differences exist (for instance, trait dynamics might be fast relative to the epidemic time scale) these should be exploited to facilitate the analysis!

Admittedly the above is a rather sketchy description of a huge class of models. It is intended as a stimulus, as an invitation, and not as a solid exposition.