The relationship between real-time and discrete-generation models of epidemic spread

Lorenzo Pellis*, Neil M. Ferguson, Christophe Fraser

Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London W2 1PG, United Kingdom

ARTICLE INFO

Article history:
Received 17 January 2008
Received in revised form 23 June 2008
Accepted 8 August 2008
Available online 26 August 2008

Keywords:
Final size
Chain binomial model
Reed–Frost model

ABSTRACT

Many important results in stochastic epidemic modelling are based on the Reed–Frost model or on other similar models that are characterised by unrealistic temporal dynamics. Nevertheless, they can be extended to many other more realistic models thanks to an argument first provided by Ludwig (final size distributions for epidemics, Math. Biosci. 23 (1975) 33–46), that states that, for a disease leading to permanent immunity after recovery, under suitable conditions, a continuous-time infectious process has the same final size distribution as another more tractable discrete-generation contact process; in other words, the temporal dynamics of the epidemic can be neglected without affecting the final size distribution. Despite the importance of such an argument, its presence behind many results is often not clearly stated or hidden in references to previous results. In this paper, we reanalyse Ludwig’s result, highlighting some of the conditions under which it does not hold and providing a general framework to examine the differences between the continuous-time and the discrete-generation process.

1. Introduction

The importance of the role played by mathematical models in infectious disease epidemiology is widely recognised. They represent invaluable tools in understanding the mechanisms behind the spread of diseases [1,2,17], highlighting the most important factors that drive the infection process (e.g. see [21]), investigating the effectiveness of different control policies [34,19], assessing the efficacy of vaccines and prophylaxis [24], designing more efficient observational studies [14,45], planning mass vaccination campaigns [22] and exploring future what-if scenarios [20].

After relatively few pioneering studies in the first half of the last century (among which the one that probably received most attention is that by Kermack and McKendrick [29]), the size of the literature on epidemiological modelling increased substantially throughout the second half of the century [18,5,1,25,17]. A wide variety of models has been proposed and analysed and, despite the analytical complexity due to the intrinsic non-linear behaviour of the infection process, many remarkable results have been obtained [10,7,33,16,18,17,2,25].

Some of these results are derived by exploiting ‘tricks’ that allow the transformation of the actual infection process into another one, which does not correspond to what can be observed in reality, but that has the advantage of being analytically tractable. Some examples are the Sellke construction [42] and the Scalia-Tomba imbedded representation [39,41], based on the concept of infection pressure, and the possibility to ignore the temporal dynamics of the epidemic process under certain suitable conditions. To our knowledge, this was first suggested by Ludwig (in [36], but already appearing in [35]).

On the one hand, the Sellke construction and the imbedded representation by Scalia-Tomba has been well referenced and has been often reported and explained when used (for example in [2] or [8]). On the other hand, the idea proposed by Ludwig is often not acknowledged and its presence behind many important results is hidden by chains of subsequent references; there are of course some exceptions, when authors have clearly stated its importance [26,32,34,38,3]. The main reason for this lack of referencing to Ludwig’s idea is that the result is well known among stochastic modellers. However, the result has other important epidemiological implications which may be of interest to a broader audience as well.

Naively speaking, Ludwig’s ‘trick’ consists in observing that, for a disease leading to permanent immunity after recovering, it is possible to associate to an epidemic process, occurring in continuous time, another (fictitious) contact process in discrete-time steps that has the same final size distribution [36]. In other words, as far as the final size distribution is concerned, the absolute times at which various events occur can be neglected.

Such a new contact process is often referred in the literature as ‘generation-based description of an epidemic’. However, as we will emphasise below, the word ‘generation’ may be misleading.
this reason, throughout the paper we will refer to the concept of \textit{rank} of an infective, as introduced by Ludwig in his paper: initial infectives are assigned rank zero; all those individuals that have an infectious contact with an initial infective (independently of whether they have already been infected or not by somebody else) are assigned rank one. Analogously, individuals are assigned rank \( n + 1 \) (\( n \geq 1 \)) if they avoid infectious contacts with infectives of rank less than \( n \) and have an infectious contact (not necessarily an effective infection) with an infective of rank \( n \). Ludwig argues that this rank-based process has the same final size distribution as the real epidemic process, since ‘an infective is an infective regardless of how his rank is assigned’ [36].

Two different problems initially motivated this study. The first one refers to the fact that, since generations overlap during a real-time epidemic, an individual experiences different forces of infection over time in the real-time process and in the fictitious rank-based process, and therefore it is not at all intuitive why the two processes should lead to the same final size distribution [17, pp. 27–28]. Both processes are commonly considered in the literature (e.g. [17]), but the reasons and the conditions that allow one to switch between the two are often not clearly stated or lost in references to previous results and in mathematical details.

The second problem concerns an epidemic in a household where individuals are also allowed to be infected from outside. Ball et al. [8] obtained a remarkable result for the average final size of an epidemic in an infinite population partitioned into households, where individuals mix at two different rates within the household and in the population at large. The result is based upon the following:

\textbf{Theorem 1.} Consider a group consisting of \( n \) susceptibles and an initial infective in which the susceptibles can be infected both within and from outside the group. Assume that each susceptible escapes infection from outside independently of each other and with probability \( \pi \). Then the final size distribution is the same as that of a group with no infections from outside, consisting of \( n - Y \) susceptibles and \( a + Y \) initial infectives, where \( Y \) is a realization of a binomial r.v. on \( n \) trials and probability of success \( 1 - \pi \).

The theorem is even less intuitive than the previous problem, as it may well happen that a new individual is infected from outside when a first epidemic wave has already finished sweeping through the household. In this case, the number of new infections generated by the infectious case depends on how many susceptibles escaped the previous epidemic. Furthermore, concepts like that of ‘generations of infectives’ require more careful consideration in this setting. Although apparently different, both the problems can be understood in terms of Ludwig’s result, as already suggested in [32], where a situation similar to that of the theorem is first considered (see Section 2).

The paper is organised as follows: in Section 2 Ludwig’s result is reviewed and its usefulness is highlighted. Section 3 aims at clarifying the confusion that may arise from the common use of ambiguous expressions regarding ‘generations of infection’ in the literature. Section 4 then extends Ludwig’s result for different models, highlighting the hypotheses on which it is based. Section 5 offers a simple graphical representation that distinguishes between real infections and infection contacts that do not cause infection, which also provides an intuitive idea of why the result should hold. In Section 6 a thorough discussion about the conditions under which the result does not hold is presented. Finally, Section 7 is devoted to some concluding comments.

\textbf{2. Consequences of Ludwig’s result}

Although the argument that Ludwig presents in his paper is simple to understand and might even be considered ‘obvious’ [38], the results based on it have grown more and more complex and, if considered alone, they are far from being intuitive. Some of them are described here.

A first direct consequence of Ludwig’s argument is that the presence of a latent period of any form does not affect the final size distribution of an epidemic, despite having an impact on the system dynamics. For this reason, most stochastic epidemic models considered in the literature are of the SIR type, i.e. assume that individuals are immediately infectious after infection without experiencing a latent stage (e.g. see [7,21]).

A second important consequence of Ludwig’s argument is that final size results obtained for the Reed–Frost model can be extended to a wide variety of other models. The Reed–Frost model is one of the simplest possible stochastic models to describe the spread of a disease in a closed population and was firstly introduced by Reed and Frost in a series of lectures and discussions in 1928 [5, pp. 12–13]. Despite not having been formally presented in a publication, it is widely known and has been discussed by many authors (e.g. [6,39,40,21]).

The Reed–Frost model assumes that an individual, after infection, experiences a latent period of one unit of time, followed by an infectious period concentrated in a single point in time. Each susceptible escapes infection from each infective independently with fixed probability \( q \). After the instantaneous infectious period, the infective is removed and no longer participates to the spread of the infection. The infectious process thus develops at discrete times and the different generations of infectives are clearly separated.

The Reed–Frost model is based on three main questionable assumptions. First, the infectious-to-susceptible escaping probability \( q \) is the same for every pair. Therefore, many different extensions have been considered: the multitype Reed–Frost model, where different types of individuals have different levels of infectivity and susceptibility (e.g. see [40]), the randomised Reed–Frost model, where the infectivity of individuals is randomly drawn from a specified distribution (e.g. see [38,31]) and the multitype randomised Reed–Frost model, which allows for random infectivity levels drawn from various distributions (e.g. see [37]). A further generalisation of the randomised Reed–Frost model is the so-called collective Reed–Frost model [38] and a very general approach has also been considered by Scalia-Tomba [41].

The second questionable hypothesis of the Reed–Frost model is that the assumed temporal dynamics are in general not realistic, though it might represent a good approximation of the actual dynamics in the case of a small population and a disease with long latent period and relatively short infectious period [12], e.g. the spread of measles in a household [27,4].

Other models with variable and often more realistic temporal dynamics have been proposed. The most common examples include models in which the infection process is assumed to occur according to a homogeneous Poisson process (the general epidemic model [5], when the duration of the infectious period is exponentially distributed, and the generalised epidemic model [30,31] or standard SIR model [2]), if it is distributed according to any specified distribution) and the so-called time–since–infection models, where the contact process is a non-homogeneous Poisson process with infection rate that changes as a function of the time elapsed since the infection of the individual (see [29] for a discrete time version, [17] for continuous). However, the argument given by Ludwig leads to the conclusion that the final size distributions of these models can be obtained by studying a correspondent Reed–Frost model (or its extension to variable susceptibility and infectivity) and this justifies the great attention it received in the literature.

The third unrealistic hypothesis is the homogeneous mixing assumption. Although variation in susceptibility and infectivity already represents a form of heterogeneity [44], other forms related
to the social structure of human populations have been considered as well [33,13,8,9]. For example, it is reasonable to assume that individuals are more likely to have closer contacts in their household than outside; for this reason the last 25 years have seen an increasing attention paid in explicitly modelling the presence of households and the individuals’ different behaviour within and between households.

Longini and Koopman [32] proposed a model that has been referred to as the independent-households model [15]. The attention is focused on a single household and the epidemic within the household is modelled using a Reed–Frost model enriched by the possibility that individuals are also infected from outside (community transmission). The probability of infection from outside is the same for every susceptible and is constant in time, i.e. epidemics in different households develop independently of one another. This assumption might be realistic only in a very large population of households [15].

The model has been widely used in estimating the secondary attack rate (probability of transmission within the household) [33], the probability of community transmission [33], the efficacy of vaccine and prophylaxis in reducing susceptibility, infectivity and development of symptoms after infection [23,24]. It has also been used in comparing different designs for vaccine efficacy studies [14,45].

Since the independent-household model is basically of Reed–Frost type, all the conclusions drawn from it would have required to be taken with extreme caution because highly dependent on the often unrealistic temporal dynamics of the Reed–Frost model; in particular, results could have been reasonable for measles but not for influenza, because of the short latent period and the trend of different generations to quickly overlap [27]. Ludwig’s argument drastically increases the robustness of these results.

A more realistic model allowing the probability of infection from outside to vary throughout the epidemic is usually referred to as the model with two levels of mixing [8]. Individuals mix homogeneously within their households with a certain rate and, with a different rate (usually smaller), within the population at large. The model is mathematically complex: nevertheless, important results have been obtained, among which the one already referred to in Section 1 concerning the average final size in the limit of an infinite number of households [8]. Again, the result heavily exploits the possibility to ignore the temporal dynamics of the epidemic within a household even when multiple infections from outside occur at different times throughout the household epidemic.

### 3. Problem description and terminology

For a disease leading to permanent immunity after recovery, instead of the epidemic process itself, Ludwig [36] suggested to study a different contact process. This new process is based on the concept of infectious contact and on the notion of rank of an infective. An infectious contact is defined as a contact from an infectious individual towards any other individual, that would lead to an infection if the second individual is susceptible at that time. The rank of an infective is constructed in the following fashion: initial infectives are assigned rank zero; all individuals that experience an infectious contact from an initial infective during their infectious period are assigned rank one. Note that some of these infectious contacts result in an effective infection, while others may not, namely those directed towards individuals who have already been infected by some other non-initial infectives before experiencing their infectious contact from an initial one. In Fig. 1, for example, individual 1 is the only initial infective and therefore has rank 0, individual 2 has rank 1; finally, individual 3 has rank 1, despite having been actually infected by individual 2, because of the infectious contact from individual 1. Analogously to the assignment of rank 1, an individual is assigned rank \( n + 1 \) (\( n \geq 1 \)) if he (she) avoids infectious contacts from infectives of rank less than \( n \) and has an infectious contact (not necessarily an effective infection) from an infective of rank \( n \).

Note that, when an individual experiences more than one infectious contact, only the first one leads to an infection: in Fig. 1, the infectious contact between individuals 1 and 3 can only have occurred after the infectious contact with individual 2, otherwise the real infection would have come from individual 1. Therefore, the real infection process is highly dependent on the times at which events occur. On the other hand, the rank-based construction depends only on the presence or absence of infectious contacts and not on their times of occurrence.

It is worth at this point defining some terminology used in the remainder of the paper: in general, the word process will be used to refer to a set of events and their times of occurrence; other words like network, tree, etc. will not contain any information about times of occurrence. A infectious contact process is the process describing all the infectious contacts between individuals and their times of occurrence: some of them, namely those from infectious individuals towards susceptible ones, will lead to an infection, but others may not. The infection process is the reduction of the infectious contact process only to those contacts that cause infection and their time of occurrence. The infection tree is what is obtained from the infection process once the times of infection are ignored. Infected individuals can then be gathered in generations. All these concepts are related to what can be observed in a real-time epidemic.

If, from the infectious contact process, individuals are assigned a rank according to Ludwig’s construction and only infectious contacts that are relevant for the rank assignment and their times of occurrence are considered, we obtain what will be called the rank-based process. Ignoring times will generate a rank-based contact tree. Once this tree is defined, the individuals in it can be gathered in ‘generations’. Note that these ‘generations’ may in general be different from those obtained from the infection process (see Fig. 1). For the sake of clarity, we prefer to reserve the term generations for the real generations of infectives, i.e. those obtained from the infection tree, and use the term rank while referring to Ludwig’s construction.

Such precise definitions are necessary because the infection tree and the rank-based contact tree may be topologically different; therefore, individuals with the same rank are not necessarily in the same generation. However, in the literature, the rank-based construction is almost always referred as the ‘generation-based description of the epidemic’. Analogously, sentences like ‘studying an epidemic in generations’ are ambiguous, since they probably refer to the ranks obtained from Ludwig’s construction and not to the trivial process of grouping in generations (in the real-time sense) the individuals infected in the real-time infection process. Furthermore, when discrete-generation models like the Reed–Frost one are used, the two constructions coincide and distinctions in terminology are not needed.
4. Models for the infectious contact process

In this section, we present various models for the infectious contact process between individuals. Common assumptions to all these models are the following: the disease leads to permanent immunity after recovery; the population consists of $n$ individuals, labelled with numbers from 1 to $n$; individual 1 starts being infectious at time $t = 0$ and we do not investigate how he acquired infection.

4.1. Standard SIR epidemic model

In the standard SIR epidemic model (see [2]), individuals mix homogeneously and, upon infection, they experience infectious periods that are independent and identically distributed according to a random variable $I$. During the infectious period, each infective makes infectious contact with any other given individual at the points of a homogeneous Poisson process with rate $\beta$. After the infection period has terminated, the individual recovers and does not participate further to the epidemic spread. The model has also been referred to in the literature as the generalised epidemic model [30,31].

Note that, according to this definition of the model (other authors have proposed slightly different ones), the individuals’ behaviour is described only during their infection period, since this is the only part of individuals’ behaviour that is relevant for the spread of the epidemic. However, it is clear that they interact with each other also before being infected and (assuming the disease does not cause death) after recovery.

Assume for the moment that the disease is sufficiently mild that it does not influence the social interactions of an individual, i.e. infected individuals do not behave differently from susceptible or recovered ones. For each pair of individuals $(i,j)$, we could then construct a sequence $(t_{ij}^k)_{k=0}^\infty$ of all the times of occurrence of the contacts from $i$ to $j$ after time $t = 0$. If $i$ is infected at time $T_i$ and his infectious period lasts for a time interval $I_i$, we can associate to $i$ a function $p(t)$ with a constant value $p$ for $t \in [T_i, T_i + I_i]$ and 0 otherwise, which describes the probability that the infection is transmitted across a contact.

From the set of all the times of contacts, the values of $I_i$’s associated to each individual and a sequence of realisations of independent Bernoulli trials with success probability $p$ (used to describe which contacts would be able to transmit infection) it is possible to follow the epidemic in real time, observe $\forall i$ if $i$ acquires infection and, in this case, store the time $T_i$ of infection.

Note that, so far, the development of the epidemic is observed in calendar time and that no assumption has been specified to describe how contacts occur in time. We now assume that, for each pair $(i,j)$, contacts from $i$ to $j$ occur at the points of a Poisson process with rate $c$. Poisson processes relative to other pairs are mutually independent by assumption and independent of the sequence of Bernoulli trials and of the length of the individual’s infectious periods. This assumption allows great simplification, thanks to the fact that the intervals between subsequent contacts from $i$ to $j$ are exponentially distributed and therefore satisfy the memoryless property: for this reason, the probability that a contact occur at a time $\tau$ after the infection of $i$ is exponentially distributed with parameter $c$, independently of the absolute time $T_i$ of his infection. This property allows us to switch from a calendar time perspective to a time since infection one.

The standard SIR model is now a consequence of the above more elementary assumptions since, during his infectious period, $i$ makes infectious contacts towards $j$ at the point of a homogeneous Poisson process with rate $\beta = pc$.

Note that, if more than one infectious contact occur from $i$ to $j$ during $i$’s infectious period, only the first one may be relevant: all the subsequent ones are superfluous, since $j$ has been already infected by the first contact from $i$ or was already infected before. Therefore, the full epidemic process can be described by the set \( \{ t_{ij}, i,j = 1, \ldots, n, i \neq j \} \), where $t_{ij} > 0$ represent the time of the first infectious contact from $i$ to $j$ since the (hypothetical) infection if $i$ and are drawn from an exponential distribution with parameter $\beta$. If $j$ ultimately escapes infection, $T_i$ is not defined and all the times $t_{ij}, j = 1, \ldots, n, j \neq i$, are superfluous.

It may be convenient to represent all these times in a $n \times n$ matrix with entries in $\mathbb{R}^+ \cup \{ \ast \}$, where $\mathbb{R}^+$ represents the strictly positive real numbers and the symbol $\ast$ indicates an element that has no meaning for the epidemic process, for example all the elements on the main diagonal. Once a matrix $(\tau_{ij})$ and the vector $(l_i)$ representing the duration of the (hypothetical) infectious periods are specified, all the elements $\tau_{ij}$ such that $\tau_{ij} > l_i$ can be substituted with $\ast$ because they refer to infectious contacts from $i$ occurring after he has recovered.

The matrix so obtained represents the full infectious contact process (in the terminology of Section 3). Once it is specified, we can construct both the real time infection process and Ludwig’s rank-based process and, from these, the infection tree and the rank-based contact tree. Alternatively, a graphical representation of the full infectious contact network can be obtained by drawing an arrow for every non-$\ast$ element in the matrix $(\tau_{ij})$: from it, the rank-based contact tree (but not the infection tree any longer) can then be constructed. Fig. 2 exploits a graphical representation of both the infection tree (in bold arrows) and the infectious contact network (in dashed arrows) on the same set of nodes in order to provide, through an example, an intuitive explanation of Ludwig’s result.

4.2. Time-dependent infection rate

The standard SIR epidemic model assumes a constant infectivity during the infection period but in the literature models with infection rate that varies over time are often considered [17]. In the simplest case, all individuals mix homogeneously and, at time $\tau$ after infection, they make infectious contact with any other given individual at a rate $\beta(\tau)$, where the function $\beta(\tau)$ is the same for each infective. Similarly to the previous section, the social interaction is not modelled before infection or after recovery.

However, as in the previous section, we can assume that contacts between individuals are still made at the points of a homogeneous Poisson process with rate $c$, while now the probability of transmission across a contact changes as a function $p(\tau)$ of time since infection, equal for each infective. As before, thanks to the memoryless property of the inter-contact times, it is possible to switch from an absolute time perspective to a time-since-infection one and the usual time-since-infection model is recovered by defining $\beta(\tau) = cp(\tau)$.

The main difference from the standard SIR model is that $\beta(\tau)$ may be non-zero $\forall \tau > 0$. Strictly speaking, in this case the infectious never recover. However, for each pair $(i,j)$, there still is a positive probability $q = e^{-\int_{0}^{T_i} p(\tau) d\tau}$ that $j$ escapes infectious contacts from $i$. The same matrix $(\tau_{ij})$ of the previous section can be defined: $\forall i, t_{ij} = \ast$ and $\forall i, \forall j, t_{ij} = \ast$ with probability $q$ and with probability $p = 1 - q$ it is a real positive number drawn from the density distribution

$$
\frac{\beta(\tau) e^{-\int_{0}^{\tau} p(\tau) d\tau}}{1 - q}
\int_{0}^{\tau} \beta(\tau) e^{-\int_{0}^{\tau} p(\tau) d\tau} d\tau
$$

i.e. from the normalised unconditional probability density of $j$ experiencing an infectious contact from $i$ at time $\tau$ after $i$’s infection.
Note that the assumption that the disease is mild enough not to interfere with the contact process can be relaxed: in fact, we could consider a contact rate \( c(t) \) that changes in time, therefore making the contact process non-homogeneous. However, since the contact rate changes only as a function of time since infection, the switching between an absolute time perspective and the time-since-infection one is still possible since, after infection of \( i \), the first infectious contact towards \( j \) occurs at the first point of a non-homogeneous Poisson process with rate \( \beta(t) = p(t)c(t) \).

### 4.3. Variable susceptibility and infectivity

Various forms of heterogeneities are often considered to make an epidemic model more realistic and generally applicable. These heterogeneities can be of two different types: they can refer to the interaction between the pathogen and the individual, thus affecting the probability of transmitting the infection from the ‘donor’, the probability of acquiring from the ‘receiver’ or both, or they can involve the individual–individual interaction, thus generating different contact patterns in different pairs of individuals.

The approach adopted in the previous section can be extended to allow for most forms of heterogeneity by setting, for each pair \( (i,j) \), a function of time since infection \( \beta_{ij}(t) \), being it deterministic or randomly drawn from a suitable set of non-negative functions.

Such a generalisation allows Ludwig’s result to be applicable to most of the models for diseases leading to permanent immunity after recovery appearing in the literature of stochastic modelling. Firstly, it includes all the models where the infectivity and susceptibility depend on some static characteristics of the individuals: in this case, the population is partitioned into a finite number of classes, grouping all the individuals of the same type. Individuals of type \( k \) are characterized by a susceptibility \( a_k \) and an infectious profile \( b_k(t) \). Note that, since the type of an individual is independent of the infection process, in particular the infectivity of an individual is independent of the time of infection and the identity of the infector.

A second class of models that can be considered is that characterised by randomly drawn susceptibility and infectivity. More formally, this is the case when each individual \( i \) has a susceptibility \( a_i \), drawn from a set \( A \subseteq \mathbb{R}^+ \), and an infectivity profile since infection \( b_i(t) \), drawn from a set of possible disease histories \( B \) (non-negative integrable functions on \( \mathbb{R}^+ \)). Ludwig’s result holds whenever it is possible to specify \( a_i \) and \( b_i(t) \) a priori: therefore, for all \( i \), \( a_i \) and \( b_i(t) \) are required to be independent of calendar time and \( \forall i \neq j \), \( b_i(t) \) and \( b_j(t) \) must be independent of each other. Since, in general, susceptibility and infectivity within the same individual may be correlated, the most natural assumption is independence between individuals, i.e. between \( b_i(t) \) and \( a_i \) for \( i \neq j \); a sufficient condition for all these assumptions is that \( \beta_{ij}(t) = a_i b_j(t) \). This has been referred to as separable mixing in [17].

A third kind of models is obtained when the infection rate between individuals is determined by the underlying (static) social structure (e.g. households models, overlapping group models, etc.). The simplest example is that of a households model where, denoting by \( \beta_{ij}(t) \) the infectious profile from an infective towards any susceptible in the population (global infectivity) and \( \beta_{ij}(t) \) the additional infectivity towards a susceptible within the same household (household infectivity)

\[
\beta_{ij}(t) = \begin{cases} 
\beta_{ij}(t) + \beta_{kij}(t) & \text{ if } i \text{ and } j \text{ share the same household,} \\
\beta_{ij}(t) & \text{ otherwise.}
\end{cases}
\]

A forth very general type of models is obtain by combining all the previous cases, i.e. by allowing the sets \( A \) and \( B \) from which susceptibility and infectivity are randomly drawn to depend on the individuals involved in the infection process. In this case, \( \beta_{ij}(t) = a_i b_j(t) \), where \( a_i \in A_i \) and \( b_j(t) \in B_j \). This case includes the models with two levels of mixing [8] or the more general overlapping groups models [9]. The households model described above is a particular case where \( A_i = \{1\} \) and

\[
B_q = \begin{cases} 
\{\beta_{ij}(t) + \beta_{kij}(t)\} & \text{ if } i \text{ and } j \text{ share the same household,} \\
\{\beta_{ij}(t)\} & \text{ otherwise.}
\end{cases}
\]

For all these models, analogously to what has been done in the previous section, we first draw the functions \( \beta_{ij}(t) \). Finally, we define \( q_i = \int_0^\tau \beta_{ij}(t) \, dt \), and finally we set \( \forall i, \tau_i = \cdots = \tau_i \) and \( \forall i, \forall j, \forall \ell \), \( q_i = \cdots = q_i \) with probability \( q_i \) and with probability \( p_i = 1 - q_i \) we randomly draw from the density distribution

\[
\frac{\beta_{ij}(\tau) e^{-\int_0^\tau \beta_{ij}(s) \, ds}}{\int_0^\tau \beta_{ij}(\tau) e^{-\int_0^\tau \beta_{ij}(s) \, ds} \, d\tau} = \frac{\beta_{ij}(\tau) e^{-\int_0^\tau \beta_{ij}(s) \, ds}}{1 - q_i}.
\]
5. A particular case

A simple example in a population of size \( n = 3 \) is considered in this section to remark that both the real-time and the rank-based process have the same final size distribution but that the rank-based process is much easier to study. Although we have in mind the infectious contact model described in Section 4.1, any consideration, apart from the technical details about how the probability of various epidemics is computed, holds for all the models of Section 4.

In the graphical representation of Fig. 2, bold arrows represent real infections and dashed arrows represent infectious contacts that do not lead to an infection. Consider an epidemic with final size 3. On the graph with nodes 1, 2 and 3, two bold arrows must reach nodes 2 and 3 and the possible configurations are \([\{1,2\},\{1,3\}],\{1,2\},\{2,3\}\) and \([\{1,3\},\{2,3\}\) and \([\{1,3\},\{3,2\}\) For each combination of bold arrows, all the remaining four directed links between nodes can be occupied or not by a dashed arrow, giving a total of \(3 \cdot 4^3\) possible graphs. However, any dashed arrow towards node 1 is superfluous for the epidemic, and is therefore convenient to ignore the presence or absence of these arrows, thus reducing the number of graphs to 12, each of which actually represents four graphs. The 12 graphs are collected in Fig. 2. In the terminology of Section 3, they represent all the relevant infectious contact networks.

Assuming the infectious contact model of Section 4.1, it is possible to explicitly calculate the probability of each infectious contact network in Fig. 2 to occur. For example, conditioning on the values of \( I_1, I_2, I_3 \)

\[ P(\mathcal{F} | I_1, I_2, I_3) = P(\tau_{12} < I_1, \tau_{13} \geq I_1, \tau_{23} < I_2, \tau_{32} \geq I_3, \ \text{any value of} \ \tau_{21} \ \text{and} \ \tau_{31}, \ \tau_{12} + \tau_{23} \geq \tau_{13} | I_1, I_2, I_3) \]

Note, however, that the infectious contact network \( \mathcal{F} \) has probability

\[ P(\mathcal{F} | I_1, I_2, I_3) = P(\tau_{12} < I_1, \tau_{13} < I_1, \tau_{23} < I_2, \tau_{32} \geq I_3, \ \text{any} \ \tau_{21}, \ \text{any} \ \tau_{31} \ \text{and} \ \tau_{12} + \tau_{23} \geq \tau_{13} | I_1, I_2, I_3) \]

and therefore, by the law of total probabilities

\[ P(\mathcal{F} \cup \mathcal{F} | I_1, I_2, I_3) = P(\tau_{12} < I_1, \tau_{13} < I_1, \tau_{23} < I_2, \tau_{32} \geq I_3, \ \text{any} \ \tau_{21} \ \text{and} \ \text{any} \ \tau_{31}, \ \text{any} \ \tau_{12} + \tau_{23} \geq \tau_{13} | I_1, I_2, I_3) \]

which is easier to compute.

Analogously, the probabilities of \( \mathcal{F} \cup \mathcal{F} \) and of \( \mathcal{F} \cup \mathcal{F} \cup \mathcal{F} \) are easier to compute than the probabilities of the single infectious contact networks. However, the final size is the same regardless on how we study it.

6. Validity of the result

The general framework presented in this paper allows us to analyse Ludwig's result in depth, highlighting what is preserved and what is not when switching from the real-time to the rank-based process. Thanks to this approach, it is easily seen on which assumptions the result is based and hence for which models it does not hold.

First of all, we recall that the result is specified for models of diseases that lead to permanent immunity after recovery and therefore cases of waning or no immunity are not considered in this context. However, there are some exceptions also in the case of permanent immunity.

Strictly speaking, Ludwig's rank-based construction, as described in his paper, can always be performed retrospectively after an epidemic has finished, provided that the infectious contact process is somehow known: it is just sufficient to consider all the individuals that have been ultimately infected and attribute them the proper rank value. It is clear that, since by definition the rank is assigned only to individuals that have been infected, the final sizes of the real-time and the rank-based processes are the same; furthermore, since this attribution of the rank values can be done whatever the outcome of the epidemic, it is also clear that the probabilities of occurrence of each particular final size have to be the same for both processes. However, this observation is not useful if the final size distribution cannot be computed easily in either of the two processes. Since the real-time process is rarely analytically tractable, the hope lies on the tractability of the rank-based process.

We highlighted in Section 4 how this is the case when individuals' behaviour can be studied \textit{a priori}, i.e. before the epidemic is observed. We pointed out that the key idea is to switch from a calendar time perspective to a time-since-infection one and in particular that this operation can be performed under the assumption that contacts between individuals occur at the points of a homogeneous Poisson process or, alternatively, that the non-homogeneity of the Poisson process is only due to factors that influence it as a function of time since infection.

If the contact rate changes as a function of calendar time the simplification is not possible: this may be the case when social external factors affecting individuals' behaviour are taken into account. Some examples include: individuals mixing during the day but not during the night; social patterns changing between weekdays and weekends; individuals spending some time with some individuals (e.g. at home) and then some other time with others (e.g. outside or at work); the beginning of the winter season with children going to school and mixing more closely; the implementation of a control policy increasing social distancing and so on.

If we avoid modelling all the contacts and we define individuals' behaviour only after infection (as it is done in most cases), the pos-
sibility to refer to a time-since-infection perspective is not derived, but directly assumed. Even in this case, however, the real-time infection process and the rank-based contact process are not equivalent. Firstly, the two processes differ in the absolute time of infection of an individual (e.g., in Fig. 1, the real-time infection process sees the infection of 3 before what would occur by considering the rank-based one); therefore, the equivalence of the final size distribution for the two is again compromised when the infectivity of individuals is allowed to depend on calendar time, in addition to the time since infection.

Secondly, the two processes differ in who infects whom (what is referred as the bookkeeping in [17]) and therefore Ludwig's result in this paper cannot hold in general if there is a correlation between infector and infectee, i.e., if the infectivity of an individual is affected by some characteristics of the individual that infected him. This is the case, for example, when two or more strains of the same pathogen circulate in the same population and lead to different levels of infectivity and some sort of cross-immunity: in this case, the identity of the infector determines the type of strain (and thus the future infectivity) of the infectee. See [28,43,11] for some theoretical work on this 'infector-dependent'-type model.

Note that these objections are not minor limitations to the applicability of the result to practical situations: models in infectious disease epidemiology often require external forcing to reproduce observed patterns and prediction of the likely impact of different control policies requires modelling sudden changes in the external conditions. Furthermore, some degree of dose-dependence in transmission or evolutionary changes in the pathogen's strain structure under selection pressure may sometimes play an important role when considering specific diseases.

However, although in many practical contexts the assumptions of independence on calendar time and the uncorrelated infectivity between individuals are violated, when the population is relatively small or the duration of the epidemic is relatively short, the models described in Section 3 might provide good approximations of a real-world infection process.

7. Discussion

In addition to drawing attention on Ludwig's result, in the present paper we rederived it from a different perspective and we offered a new convenient graphical way to summarise enough information of the contact process to allow the derivation of both the real-time infection tree and to assign ranks to individuals according to Ludwig's construction (see figures). We furthermore pointed out some ambiguities in terminology that can represent a source of confusion in the literature and we commented and extended the insights that can be gained from Ludwig's result, highlighting the assumptions under which it does not hold. Such a general approach also allows the investigation of more complex situations than those originally considered by Ludwig, in particular that of Theorem 1 described in Section 1.

In fact, since the final size distribution can be studied by considering the full infectious contact networks, which do not depend on the absolute time at which events occur, even in the case of multiple infections of the same household from outside at different times, the probability distribution of the sets of individuals ultimately infected or of the final size distribution can still be computed combining the probabilities of various networks occurring with those of various individuals being infected from outside and taking the union of the outgoing connected components of these individuals. Furthermore, the assumption of constant probability $\pi$ of infection from outside is also justified in light of the possibility to ignore the temporal dynamics of the epidemic in the population at large. In particular, $\pi$ can be calculated from the total force of infection due to all the individuals ultimately infected in the population, independently of when they are infected.

This argument increases even further the relevance of Ludwig's result, which was already clear from the fact that it allows the results concerning the final size of the Reed–Frost model and its generalizations (multitype, randomised, collective Reed–Frost models, Reed–Frost models with infection from outside [32,38,37]) to be extended to many other models with more realistic temporal dynamics. It is therefore a fundamental result of robustness with direct and indirect consequences in the field of mathematical models for infectious disease epidemiology.

Acknowledgments

The Institute for Mathematical Sciences, Imperial College London, is gratefully acknowledged for funding the main author's Ph.D.; we also thank the two referees and Dr. Pete Dodd for having much improved the readability of this paper and Dr. Simon Cauchemez for useful comments and discussions.

References
