THE UNIVERSITY
of EDINBURGH

Dynamics of simultaneous epidemics on complex graphs

Denys Zachary Alexander Janes

Doctor of Philosophy

The University of Edinburgh

2017

Abstract

The subject of this thesis is the study of a system of multiple simultaneously spreading diseases, or strains of diseases, in a structured host population. The disease spread is modelled using the well-studied SEIR compartmental model; host population structure is imposed through the use of random graphs, in which each host individual is explicitly connected to a predetermined set of other individuals. Two different graph structures are used: Zipf power-law distributed graphs, in which individuals vary greatly in their number of contacts; and Poisson distributed graphs, in which there is very little variation in the number of contacts. Three separate explorations are undertaken.

In the first, the extent to which two SEIR processes will overlap due to chance is examined in the case where they do not affect each other's ability to spread. The overlap is found to increase with increased heterogeneity in the number of contacts, all things equal. Introducing differences in infection probability or a delay between introducing the two strains produces more complex dynamics.

I then extend the model to allow strains to modify each other's transmissibility. This is found to lead to modest changes in the size of the outbreak of affected strains, and larger effects on the size of the overlap. The extent of the effect is found to depend strongly on the order in which the strains are introduced to the population. Zipf graphs experience somewhat larger reductions in outbreak size and less reduction of overlap size, but overall the two graphs experience similar effects. This is due to the reduced effect of modification in key high-degree vertices in the Zipf graph being offset by higher local clustering.

Finally, I introduce recombination and competition by replacement into the model from the first project. The number of recombinant strains that arise is found to be either very low or very high, with chance governing which occurs. Recombinant strains in Zipf distributed graphs have a significant chance of failing to spread, but not in Poisson distributed graphs. Replacement competition in the presence of a growing number of strains is found to both increase the chance of a strain failing to spread, and to reduce the overall size of outbreaks. This effect is equal in both graph types.

Lay summary

The geographical and social structure of human populations plays an important role in the spread of infectious diseases within them. Transmission of disease between two individuals is restricted by the rate of contact between them, and so differences in contact rates are important. These are often modelled using networks in which each contact is explicitly modeled. Coinfection, in which one individual is infected with two or more diseases or strains of a disease at the same time, is a major public health concern, because coinfected individuals experience more severe symptoms, and because new strains can arise through recombination, the genetic mixture of two strains to form a third. Such recombinant strains can inherit resistance to different drugs from each parent and be multidrug-resistant, or they can be inrecognizable to the immune system, leading to large unexpected epidemics, as was the case with swine-origin influenza in 2009.

This thesis describes the way that host population structure affects the interplay of two or more different disease or strains of disease spreading simultaneously through the same contact network. I use one type of network where contacts rates vary very little, and one where they vary a lot. First, I determine how many individuals are likely to become coinfected, and whether an individual's number of contacts affects its chance of coinfection. In general, the more variation there is in the number of contacts, the more coinfection, because very connected individuals contract and spread both diseases at a high rate.

Second, I examine how the extent of this overlap between different outbreaks changes when one of the diseases can either enhance or hinder the transmission of the other. Population structure affects the impact of these enhancements or hinderances in several ways, but these largely cancel out, so that the patterns are similar in both networks.

Finally, I examine how population structure affects the rate recombination in coinfected hosts, and how this depends on the rate of competition for resources among strains within the host. When contact rates vary, very poorly connected individuals are less likely to be the source of new strains, because they are less often coinfected, and strains that do arise are less likely to spread to others. Competition among strains strongly suppresses the rate that new strains arise, and reduce the size of their outbreak when they do. This does not depend on the variability in contact rates.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

D. Zachary A. Janes

August 2017

Acknowledgements

First, I would like to thanks my supervisors, Dr Nick Savill and Professor Andrew Rambaut for their help, guidance, advice and support. I would like to give a special thanks to Nick, who has allowed me the freedom to conduct this project as I felt it should be, while giving me all the support I could wish for, and making sure I didn't go astray. I would like to thank the MRC for the funding that supported me through this process. A special thanks to my examiners, Dr Thomas House and Dr Samantha Lycett for their very helpful comments, and in particular Dr Lycett for her support and advice during the corrections process.

A great many people have helped me in so many ways during this process, and I cannot adequately thank them all. I would like to extend special thanks to Dr Daniel Cornforth for many illuminating conversations, and for asking the questions that inspired the work in presented in this thesis; Dr Diarmuid Lloyd and Dr Pippa Stone for helpful discussions and for keeping me sane in the office; Swie Joo Liem, Alice Helliwell, Yusuke Onishi, Morgan Trigg, Paul Skelding, Dave Orr, Dr Peter Gibbons and the Edinburgh University Shukokai Karate Club for keeping me fit, and for providing me with a diverting avenue of learning that had nothing whatever to do with the problems I wrestled with in the office; David Stone and Cameron Fox-Clarke for all their friendship and support in so many ways; the staff at the Black Medicine Coffee Co. and the Auld Hoose, in which substantial parts of this work was undertaken, for their, generosity and patience; and the McElroys and Dr Simon Mayo and Dr Mark Kermode for keeping me going when I thought I could not.

Very special thanks to Ger O'Dea for being a counselor, teacher and inspiration; to Joe Farrimond and Petter Solhaug for being the best and most inspiring friends a person could hope to have; and to my father, Denys H. Janes, for putting my feet on this path, and for everything he taught me.

I would like to thank my mother Cecilie Hansen and my sister Caroline Janes for everything they have done for me throughout this process, and for the love and support they have given.

Finally, I want to thank my husband, Jody Rae, for being there, for always putting up with my frustration and every up and down of this process, and for his constant love and support, without which I could never have finished this. I love you.

Contents

List of Figures

List of Tables

Chapter 1

Thesis outline

The epidemiology of infectious disease is one of the largest areas of study in the science of public health. Understanding the dynamics of outbreaks, what determines the number of individuals infected, how the disease persists, and how it will respond to different interventions is necessary to developing effective policies for dealing with emerging outbreaks.

Because of the nature of the subject, empirical research is largely limited to analysis of data collected from outbreaks occurring "in the wild". Experimentation is very limited, and so elucidating the underlying mechanisms of disease spread can be difficult. Theoretical work can help us understand what the underlying mechanisms might be. Mathematical modelling comprises a large portion of the theoretical work, but because the systems being studied are complex, mathematical models are often limited by what is analytically tractable. Computater simulation allows the study of systems that are not easily defined and explored mathematically.

The study of host population structure is crucial to understanding the dynamics of diseases spreading through that population. One of the major tools currently used to study the interaction of host population structure and disease dynamics is contact networks, or *graphs*. Epidemic processes on graphs have been extensively studied, and many complex disease mechanisms are well understood in a range of host population structures.

The work presented in this thesis adds to the growing literature on the dynamics of multiple strains spreading simultaneously on a graph. The work consists of three simulation projects conducted on the same system, allowing increasingly complicated interaction dynamics between the strains. The structure of the thesis is as follows:

Chapter 2 provides a brief overview of the relevant literature, to allow the reader to put the work in context.

chapter 3 is a description of the simulation methods used in the thesis, and a comparison of their performance to other methods used in the literature.

In **chapter 4** I present the first project, in which I measure the size and structure of the overlap between two diseases spread in the same graph but without any effect on each other.

Extending this model, in **chapter 5** I present the second project, in which infection with one disease or strain alters the rate of transmission of a second disease or strain.

Also extending chapter 4, in **chapter 6** I present the third project, in which two diseases spread simultaneously without the cooperative or competitive dynamics of chapter 5, but where coinfection can give rise to new recombinant strains, and there is a cap on the number of strains that can infect a host simultaneously.

Finally, in **chapter 7** I discuss the results of the three projects as a whole, and suggest some promising areas of further work.

Chapter 2

Background

In this section I briefly present those parts of the literature on epidemic diseases that are most directly relevant to the present work.

2.1 Basic epidemiological theory

The vast majority of mathematical and computational modelling of the spread of infectious diseases that spread through direct contact between hosts is carried out using *compartmental* models. Much of the literature follows Kermack and McKendrick [101], who developed one of the earliest such models. Each individual in the host population is assumed to be in one of several *compartments*. A small fraction of the population starts in the *Infected* compartment (I), and the rest in the *Susceptible* compartment (S). Hosts are assumed to interact at random in a *well-mixing* fashion, so that each two hosts have an equal chance of interacting, and according to a *mass-action law*, so that the rate at which hosts go from S to I is given by the term βSI, where the constant β is the transmission parameter. The term βI is called the *force of infection* [98]. Hosts spend some time in the I compartment, before transitioning to the R compartment, which is variously called the *Resistant* or *Removed* compartment, depending on the fate of the hosts. Note that mathematically it does not matter whether hosts recover with complete immunity, die or are removed from the population through quarantine, so long as hosts in the R compartment cannot infect or become infected.

The time hosts spend in the I compartment is assumed to be exponentially distributed with mean γ, known as the *recovery rate* [29]. The three compartments give this model its name, the SIR model. If we assume that there is no birth, death or migration in or out of the host population, so that the population always consists of the same N hosts, we can express the distribution of the host population among the three compartments over time by the following system of ordinary differential equations:

$$\begin{aligned} \frac{dS}{dt} &= -\beta IS \\ \frac{dI}{dt} &= \beta IS - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{2.1}$$

Notice that as $N = S + I + R$, $R = N - (S + I)$, and so we can omit the equation for $\frac{dR}{dt}$. The SIR model is illustrated in figure 2.1, along with three other compartmental models that I shall discuss.

Solving this system of equations would give the functions $S(t)$, $I(t)$ and $R(t)$, which would tell us the state of the system at any time t. However, solving such systems is not always easy, and many major results have been derived by other means. We can obtain approximate solutions of the system by numerical simulation methods [59]. Figure 2.2 shows the output of such a simulation.

Comparing equation 1.1 with figure 2.2, several results are immediately apparent. Clealy $\frac{dI}{dt} = 0$ either when $I = 0$ or $\beta S = \gamma$. The first case makes intuitive sense - the disease cannot spread if it is not present in the population. The second case occurs when the rate at which hosts are becoming infected exactly equals the rate at which infected hosts are recovering. This point occurs at the peak of the I curve in figure 2.2 . I increases when $\beta S > \gamma$, and decreases when $\beta S < \gamma$. Since $\frac{dS}{dt} < 0$, βS decreases over time. and so regardless of how fast it initially grows, I will eventually start shrinking, and will continue to do so until the disease is extinct.

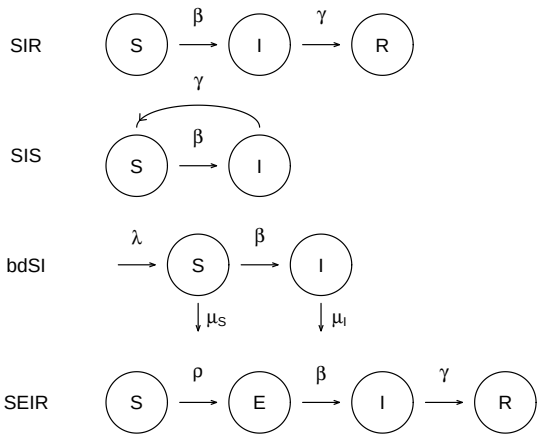


Figure 2.1: Diagram of the SIR, SIS, bdSI (SI with birth and death rates), and SEIR compartmental models. The equations for the bdSI model given in the text assume that $\lambda = \mu_S = \mu_I$, so that the term for $\frac{dS}{dt}$ can be simplified from $\lambda S - \beta SI - \mu S$ to $\mu I - \beta SI$.

We can rewrite the condition $\beta S = \gamma$ as $\frac{\beta S}{\gamma} = 1$. This function is usually called $R^*(t)$. In practice we usually consider the quantity

$$R_0 = R^*(0) = \frac{\beta S(0)}{\gamma} \tag{2.2}$$

called the *reproductive number*. R_0 can be thought of as the number of infections a single infected individual would cause directly if introduced into an entirely susceptible population. If $R_0 > 1$ the disease will initially experience a period of expansion before eventually dying out. If $R_0 < 1$ the disease will die out extremely fast, and no large outbreak is possible. Figure 2.3 shows the I curve from figure 2.2 scaled by the maximum value of I, I_{max}, and with the function R^* plotted, and shows that R^* is always decreaing, and that I peaks when $R^* = 1$.

One other major observation for this system is that, although we cannot easily calculate S_∞, the number of hosts who were never infected, we can show that $S_\infty > 0$, so that the disease never reaches every host. Approximating $S(t)$ and $I(t)$ with a simple numerical simulation (figure 2.4)

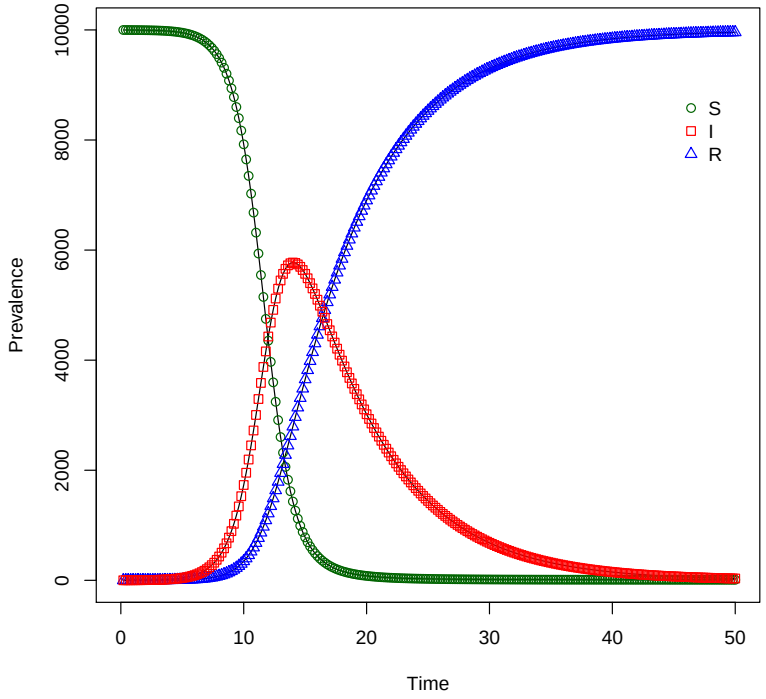


Figure 2.2: Simple Euler numerical simulation of the SIR ODE system. $N = 10000$, $S_o = 1$, $\beta = 0.0001$, $\gamma = 0.15$, $\tau = 0.2$

confirms that this is the case, but also shows that although $S_\infty > 0$, it may be so close to 0 that when the model is scaled to a real population $S_\infty < 1$.

At this point, I is worth noting that the formulation of the system given in equations 2.1, while standard, is not the only way to describe the system, and involves several implicit and rather strong assumptions about the way hosts interact and disease spreads. If the quantities S, I and R are taken to represent the number of individuals in each compartment, then the equations must be rewritten with these terms replaced by $\frac{S}{N}$, $\frac{I}{N}$ and $\frac{R}{N}$. Instead, it is common to take S, I and R to represent the density of individuals in each compartment in the host population [123]. This strictly implies that the population is taken to occupy a fixed spaceso that increasing the number of hosts increases the density [48]. This is usually assumed to be a unit area, so that we do not have to

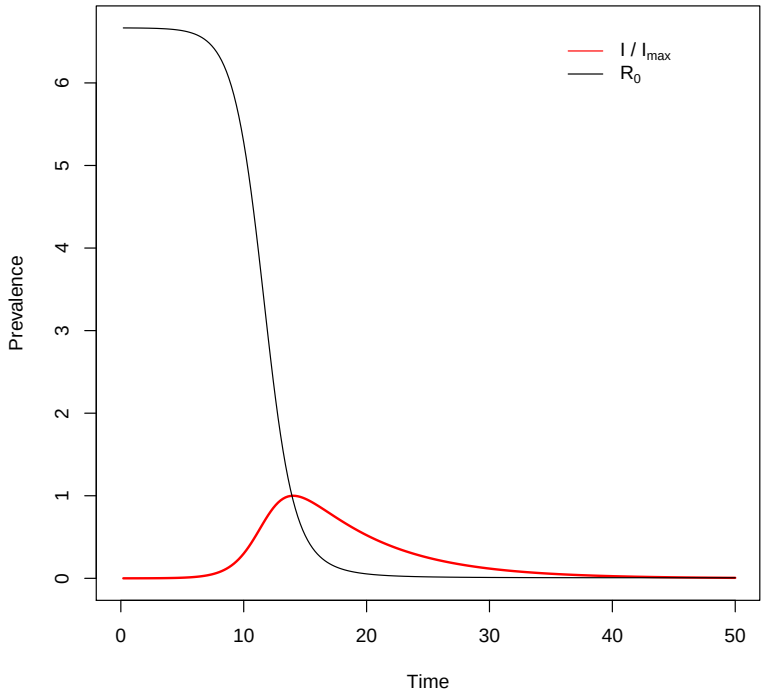


Figure 2.3: $I(t)$ in the simple SIR model, scaled by the peak prevalence I_{max}, and R^*. $R_0 = 6.66$.

explicitly include the size of the area in the calculation of density.

A common alternative to density dependence is *frequency dependence*. In this version of the equations, we do not assume the hosts occupy a fixed area, and so the rate of infection depends on the frequency with which interactions are with an infected individual, giving $\beta S \frac{I}{N}$.

In the density dependent model, we can use R_0 to derive a minimum necessary population size below which the host population is resistant to invasion by a pathogen (R_0 will not exceed 1) [4–7]. However, in the frequency dependent version the term for population size vanishes and there is no critical population size [48]

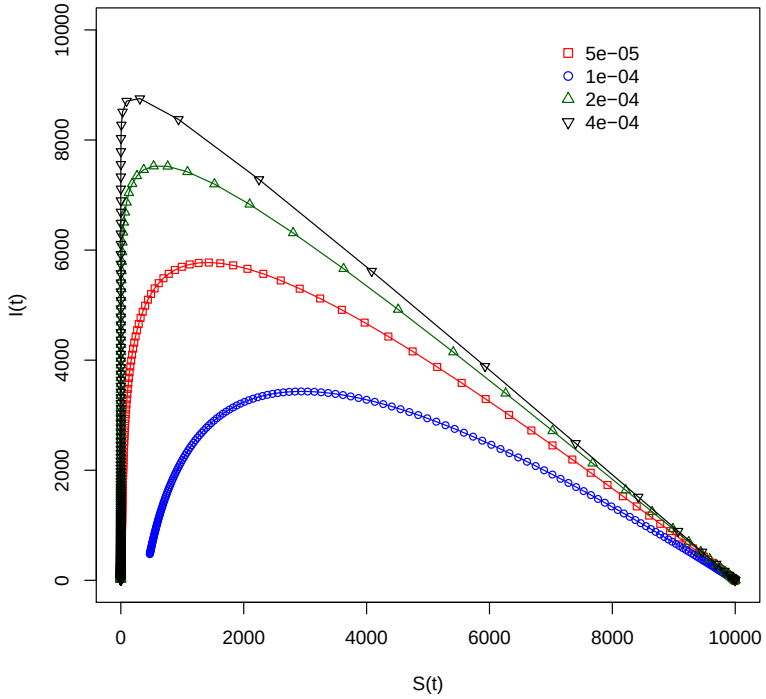


Figure 2.4: Relationship between $I(t)$ and $S(t)$ in the SIR model, for four different values of β.

Many authors do not carefully state which version of this model they are using [123]. In the study of many human disease this does not necessarily matter that much, because mortality rates are comparatively low, so that N does not change much over the course of the disease[1] and issues that arise from how we model dependence on N are minor at worst. In the study of animal diseases, where mortality can be very high indeed, it makes a much larger difference [123].

In this chapter I will generally be using the density-dependent model, and assume that S, I and R represent the densities of the hosts, rather than counts. I will also assume that the area they occupy is of fixed unit size.

[1] At least not due to infection

There are many possible extensions and modifications to this model that attempt to deal with several fairly unrealistic assumptions that the model relies on. I will not discuss them all here, as mostly they do not bear directly on the work presented in this thesis. I will, however, present some of the more fundamental variations. Much more thorough discussions can be found in any number of text books [8, 30].

The simple SIR model implies that all disease outbreaks die out, and usually fairly quickly. This is clearly not the case we observe in reality. Some diseases, such as HIV, produce a sustained epidemic over very long time-scales [162]. Others, for example influenza and measles viruses, cause recurrent outbreaks every year, or every few years [22, 66, 89]. Some adjustment of the model is clearly in order. One simple extension is to relax the assumption that the disease has the progression $S \to I \to R$. Many viruses, such as the Herpes simplex viruses, cause permanent infection that is non-fatal in the vast majority of cases [28] In this case there is no R compartment, giving the simpler SI model:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI \end{aligned} \tag{2.3}$$

It is apparent by inspection that there are two possible long-term states of the SI system. Either the disease is never introduced to the population ($I = 0$ for all t), or the entire population is eventually infected ($S = 0$ after some time). In either case $\frac{dS}{dt} = \frac{dS}{dt} = 0$ and so the system will remain in that state indefinitely. The case $I = 0$ is called the *disease-free equilibrium*, while the case where disease persists indefinitely is called the *endemic equilibrium*. In the SIR system there is only one long-term state with respect to I: the disease always goes extinct, and the the focus of analysis is on determining the size of the outbreak before its inevitable extinction. In reality, many diseases exist long-term in the host population without either going extinct or infecting everyone. The simple SI and SIR models clearly do not accurately capture their dynamics.

One reason for this is the assumption that there is no birth, death or migration in the host population. The simplest way to introduce these into either model is to assume that births and deaths happen at rates λ and μ respectively, both proportional to N. We make the simplifying

assumptions that the disease does not cause deaths so that μ is equal in all compartments, and that new arrivals are always in state S. Finally, we assume that $\lambda = \mu$, to obtain the following in the SI case:

$$\begin{aligned} \frac{dS}{dt} &= \mu I - \beta SI \\ \frac{dI}{dt} &= \beta SI - \mu I \end{aligned} \tag{2.4}$$

Now the endemic equilibrium shifts from $S = 0$ to $S = \frac{\mu}{\beta} > 0$, and so the disease persists over the long term in a fraction of the population, because new hosts are entering the S compartment to replace those that become infected. In this model, $R^* = \frac{\beta S}{\mu}$. $R_0 < 1$ still means that the outbreak goes extinct very fast, but now $R^* = 1$ indicates that the system has reached the endemic equilibrium, and $R_0 > 1$ implies that the system will eventually reach the endemic equilibrium.

Many pathogens cause infection that is eventually cleared by the host, but does not result in immunity. Two examples are gonorrhoea and chlamydia [186]. In this case there is no R compartment, but hosts that are in the I compartment eventually return to the S compartment, giving the SIS model:

$$\begin{aligned} \frac{dS}{dt} &= \gamma I - \beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \end{aligned} \tag{2.5}$$

This is the same system as equation 1.4, with $\mu = \gamma$, and so the same dynamics. Figure 2.5 compares the prevalence curves for the SI, SIR and SIS models with the same parameter values and initial conditions, and shows their different final states.

None of these simple models differentiate between being *infected* and *infectious* — once a host becomes infected it is immediately capable of passing infection on to others. In reality many diseases have an *incubation* period after infection, in which the newly infected host is not yet infectious. These can vary in length, on the order of days or weeks for respiratory viruses such as influenza (less than a day) or measles (almost two weeks) [107] to several years for leprosy, a bacterial infection [148]. To accomodate this into the model we introduce a fourth compartment. This compartment is usually called the *Exposed* compartment (E). The resulting SEIR model is:

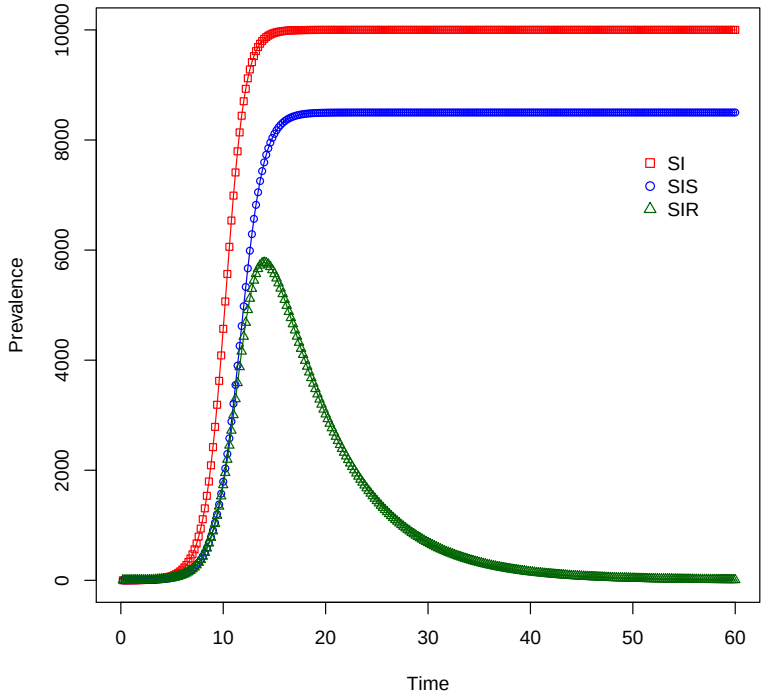


Figure 2.5: $I(t)$ for three basic comparmental models — SI, SIR and SIS — with $N = 10000$, $I_0 = 1$. In all three models $\beta = 0.0001$, in the SIS and SIR models $\gamma = 0.15$. Note the three different final states: $I_\infty = 0$ (SIR), $I_\infty = N$ (SI) and $I_\infty = \frac{\mu}{\beta}$ (SIS).

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dE}{dt} &= \beta SI - \rho E \\ \frac{dI}{dt} &= \rho E - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{2.6}$$

This simple SEIR model is straightforward to handle. In has very similar behaviour to the SIR model, since the difference is simply that the I compartment has been split into two parts. Although

the final size of the infection is still $I_\infty = 0$, the maximum size of I is smaller and occurs later, since hosts must first pass through the E compartment, and a fraction of the infected hosts are in the E compartment when I reaches its peak size. Figure 2.6 shows the prevalence curves for SIR and SEIR models with the same β and γ and initial conditions. The "total prevalence" $Y = E + I$ is also shown for the SEIR model, showing that the net effect of adding the E compartment is to slow the outbreak down significantly.

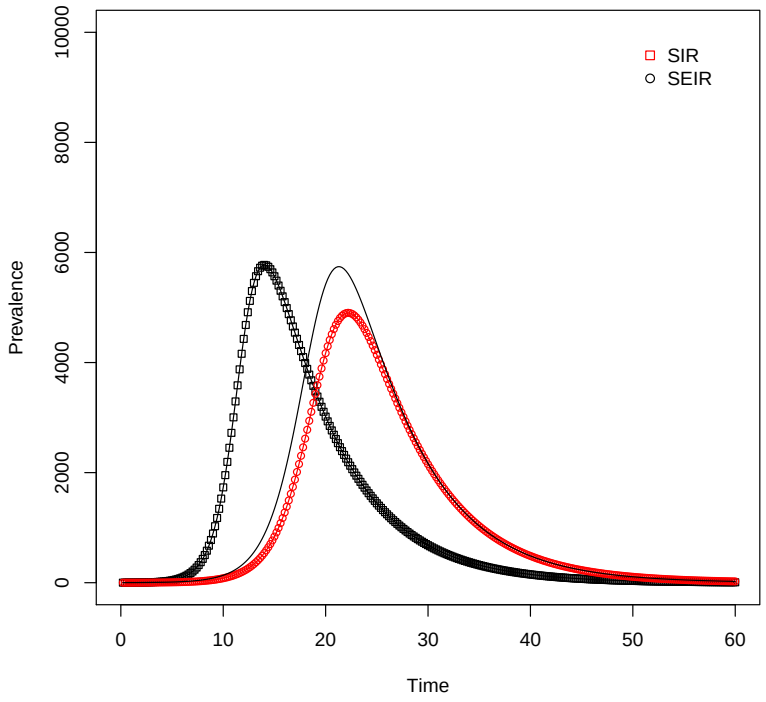


Figure 2.6: $I(t)$ for SIR and SEIR models with $\beta = 0.0001$, $\gamma = 0.15$, $N = 10000$ and $I_0 = 1$. The thin black line shows $Y = E + I$ for the SEIR model, that is all individuals who are *infected*, whether they are infectious or not.

The term describing recovery from infection in all the models discussed so far – γI – assumes that the duration of infection is exponentially distributed. This is somewhat implausible, because it implies that recovery can happen at any point after infection, including virtually instantaneously. In

reality there is much less variation in the duration of infection. Several classical studies of childhood diseases suggested that infection duration is often much more clustered around the mean [88, 185]. Although they are often less tractable, mathematical models that assume distributions that are more clustered around the mean have been studied (e.g. [112]).

The models discussed so far is a continuous-time model - the times as which individuals transition between compartments take real-number values according to the distributions, usually exponential, that govern each state transition. While this is probably an good model of the real world, in many situations it can lead to complexities in calculation and simulation. We can make the simplifying assumption that all infections caused by a given individual happen simultaneously at the end of the infectious period. Such models are called *discrete-time* models, because the times between infections are of even size and so are usually sscaled to an integer value. The first and simplest discrete-time model was presented by Reed and Frost in a series of lectures in 1928. They did not publish their model, which was first published by Abbey [1]. Like the Kermack-McKendrick model, it assumes a well-mixing population. Using the same notation as for the Kermack-McKendrick model, the expected number of infected individuals at time $t+1$ in the basic Reed-Frost SIR model is given by

$$\mathbb{E}(I(t+1) = S(t)(1 - q^{I(t)}) \tag{2.7}$$

where p is the probability that infected host u will infect susceptible host v, and $q = 1 - p$ [54]. Using an approximation for the early steps when $I(t)$ is very small, one can recover the R_0 threshold condition from the Kermack-McKendrick model described above, but now the behaviour when $R_0 > 1$ splits into two regions, one in which there is a large but finite epidemic (like the Kermack-McKendric model), and one in which the disease keeps spreading [31].

Unlike the Kermack-McKendrick model the Reed-Frost model is stochastic - it models infection as an event that happens in a given time-step with probability p and does not with probability q, which is why equation 2.1 gives the expected value of $I(t+1)$ rather than just $I(t+1)$ This is a better assumption for very small populations or very early in the outbreak, when there are very few infected individuals [31]. Stochastic epidemiological models are a large area of research, but as I will be considering a continuous-time model here, I will not be covering these in any

depths here. It is however worth mentioning the concept briefly, as discrete-time models are often easier to model computationally, and many computer simulation models are based on discrete-time models [81, 86, 115].

The final assumption in the basic model that will be relaxed in the model I study in this thesis is the assumption that the population is *well-mixing*. This is a necessary assumption for the infection term βSI in all the above models, and implies that *every* pair of individuals in the host population has the same chance of interacting, so that interactions happen completely at random. Most host populations exhibit some sort of structure, in which some individuals interact more with each other than with others in the population. The structure can be induced by geographical distance — two people living in the same town may be more likely to interact than two people living on opposite sides of the country. It can also be socially induced — you interact more often with friends and family than with any particular stranger.

Several approaches have been taken to incorporating population structure into epidemiological models [144]. Perhaps the most widely studied approach is *graph* epidemiology, in which the individuals in the host population are considered as nodes that are connected pairwise by discrete links.

2.2 Definitions from graph theory

Before discussing the use of graphs to model host population structure, it is necessary to introduce a number of definitions and terms from the mathematical theory of graphs. This section is intended for the reader to look up any terms that are unfamiliar, rather than as a primer in graph theory, and is as compressed as is practicably possible. To aid understanding, most of the definitions are illustrated in figure 2.7. This section summarises standard definitions, which unless otherwise stated are taken from *Networks: An Introduction*, by Mark Newman [139].

A *graph* $G(V, E)$ or G is a set V of *vertices* and a collection E of *edges*, where each edge in E is a pair (u, v) of vertices in V. An alternative nomenclature is to say that a *network* (graph) is made up of a set of *nodes* (vertices) [98]. If $(u, v) \in E$, we say that u and v are *neighbours*. The

collection of all edges $(v, u_i) \in E$ is called the *neighbourhood* of v. The size of the neighbourhood is known as the *degree* of v, k_v.

The edges in E can be ordered tuples, in which case we say that G is a *directed* graph; or unordered pairs, in which case G is *undirected*. An edge (u, u) connecting a vertex to itself is called a *self-edge* or *self-loop*. If two vertices u and v are connected by more than one edge in E, these edges are collectively referred to as a *multi-edge* or *parallel edge*. A graph that contain no self-edges or parallel-edges is called a *simple* graph.

Many calculations on graphs are made considerably easier by representing the graph as an *adjecncy matrix* $\mathbf{A}$, in which the elements $\mathbf{A}_{uv} = 1$ if vertices u and v are neighbours, and 0 otherwise[2].

There is a *path* between vertices v and w if there exists a set of n vertices $\{u_1, u_2, \ldots, u_n\} \in G$ with edges $\{(v, u_1), (u_2, u_3), \ldots, (u_n, w)\} \in E$, that is if we can travel vertex-by-vertex from v to w along these edges. The shortest path between two vertices is called the *geodesic path* between the two vertices. A graph in which there exists a path between every pair of vertices is called a *connected* graph. If a graph is not connected then it must consist of several *components*, each of which is connected[3].

A graph of N vertices in which every pair of vertices are neighbours is called the *complete* graph of size N, denoted K_N.

Given a graph $G(V, E)$, a second graph $G_2(V_2, E_2)$ is a *subgraph* of G if $V_2 \subseteq V$ and $E_2 \subseteq E$. That is, G_2 is a subgraph of G if it consists of only vertices and edges from G. A subgraph $G_2(V_2, E_2)$ of $G(V, E)$ is called an *induced* subgraph of G if $(u, v) \in E \implies (u, v) \in E_2$, that is G_2 consists of vertices from G and *all* edges that connect those vertices to each other in G. Every simple graph of size N is a subgraph of K_N.

We can measure and classify graphs in a number of ways. Many of these differ between directed and undirected graphs. Since in this thesis I will only be studying undirected graphs, I only give the

[2]The entries can take other values, if the edges of the graph are *weighted*. I will not be discussing weighted graphs further in this thesis, although they do have uses in epidemiology

[3]Each component could be a single vertex

definitions in the undirected graphs, which are generally simpler. I only mention the most standard measures that are commonly used in epidemiology on graphs, and that I will be using.

The *size* N of the graph is the size of V, $|V|$. The *density* of the graph is given by

$$\frac{|E|}{|V|(|V|-1)} \tag{2.8}$$

where $|V|(|V|-1)$ is the number of edges in K_N. That is, the density is the fraction of all possible edges that are in fact present. If $D(G) = 0$, $\mathrm{E} = \emptyset$; if $D(G) = 1$ then $G = K_N$. The *diameter* of the graph is the longest geodesic path in the graph[4].

The *degree distribution* of a graph is the frequency at which each degree occurs in the graph. The *triangle density* of the graph is the probability $P((u,w) \in E \mid (u,v),(w,v) \in E)$, that is, the probability that the two neighbours u and w of v are themselves neighbours. The triangle density is sometimes called the the *transitivity* of the graph (since if it is 1, then the property of being neighbours is transitive), or the *clustering coefficient* of the graph, although there are also other graph measurements that are referred to as clustering. The triangle density can be calculated as:

$$\frac{trace(\mathbf{A}^3)}{\|\mathbf{A}^2\| - trace(\mathbf{A}^2)} \tag{2.9}$$

The *degree correlation* or *assortativity*[5] of a graph measures the tendency of neighbours in the graph to have similar degree. If the assortativity is positive we say that the graph is *assortative*, if it is negative we say that the graph is *disassortative*. If it is zero we say the graph is *uncorrelated*. The assortativity ($A(G)$) can be calculated as:

$$A(G) = \frac{\sum_i k_i \sum_{ij} k_i k_j \mathbf{A}_{ij} - \left(\sum_i k_i^2\right)^2}{\sum_i k_i \sum_i k_i^3 - \left(\sum_i k_i^2\right)^2} \tag{2.10}$$

[4] A related measure is the *mean geodesic path length*, which is sometimes also called the diameter of the graph.

[5] Assortativity can be defined more generally as the correlation of any numerical property of nighbouring vertices in a graph, but the only property I will consider here is the degree of the vertex.

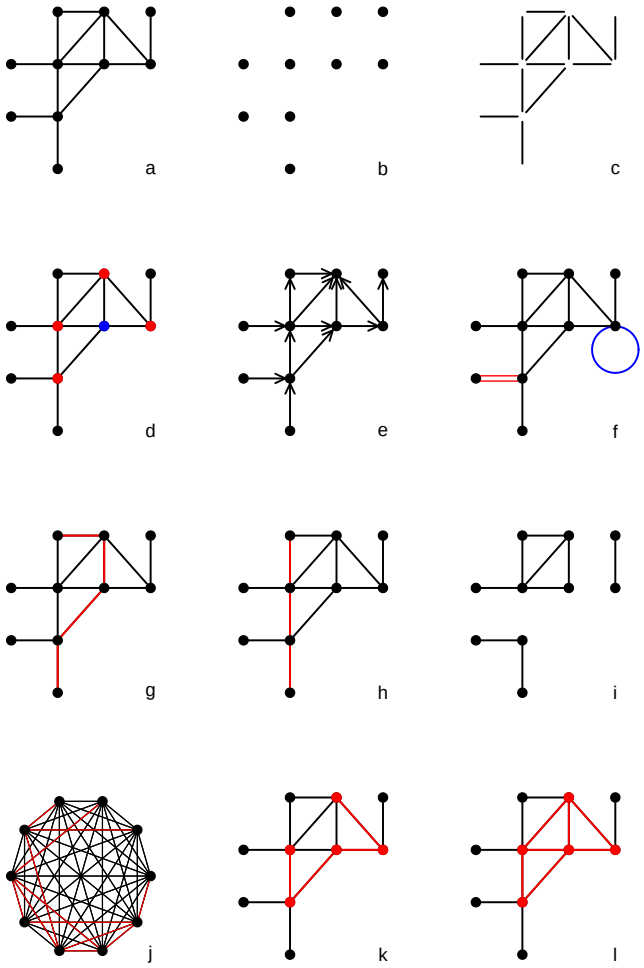


Figure 2.7: a) a graph $G(V, E)$, b) V, c) E, d) the neighbours (red) of a vertex (blue), e) a directed version of G, f) G with a self-edge (blue) and a two parallel edges (red) added, g) a path in G, h) the geodesic path in between the same vertices in G, i) three unconnected graphs, all subgraphs of G, j) G as a subgraph of K_9, k) a subgraph of G, l) the induced subgraph of the same vertices in G.

2.3 Generating random graphs

In epidemiolgy, graphs are commonly used to model the structure of a population, with the vertices representing the individuals in the population, and an edge between two vertices denoting that the two are capable of transmitting infection between them. The goal is to produce graphs that include all contacts that are relevant to the spread of disease, and no others. There are several problems. While considerable effort has gone into quantifying the contact structure of human populations with some success (see for example [71,110,151]) a completely accurate picture is very difficult to obtain. Even if one did have a perfectly accurate contact graph for a particular point in time, humans change their contacts over time [144]. People travel, move homes, and change jobs. Friendships and romantic relationships are broken and formed, and people are born and die. Although these various processes produce a lot of variability on the microscopic level of individual contacts, it is often possible to discern large-scale patterns in the structure [139].

In order to capture both the microscopic variability and the macroscopic structure of the host population, it is common to use *random graphs*, in which edges between vertices are added to the graph according to some rule that incorporates some element of chance. In the rest of this section I discuss several methods of generating random graphs and some of the properties of the graphs they produce.

The study of random graphs began almost simultaneously in two different disciplines: graph theory, a branch of mathematics, and percolation theory, a branch of physics. Percolation theory begins with the work of Broadbent and Hammersley [32]. Consider a regular square lattice in which each vertex (ignoring the boundaries) has four neighbours. We now begin removing edges from the graph independently with probability $1-p$, so that each edge is present with probability p. Figure 2.8 shows the result of this process in two lattices of 1000 vertices, with $p = 0.15$ and $p = 0.52$. We then ask the question, is there still a path connecting vertices on opposide boundaries of the lattice? This formulation of the problem is called *bond percolation*. An alternative model is called *site* percolation, where with probability $1-p$ we remove a *vertex* and its four edges from the graph.

Clearly, if $p = 0$ there is no infinite cluster, and if $p = 1$ there is. For an infinite lattice, one can show that the probability of there being an infinite cluster is either 0 or 1 [139] for *any* value of

p. It can be showsn that there is a threshold value of p below which there is no infinite cluster, and above which there is. This critical value is called the *percolation threshold*. The percolation threshold depends on the type of lattice (how each vertex is assigned neighbours), and is usually very difficult to calculate exactly. In the case of the two-dimensional lattice where each vertex has 4 neighbours it is known to be exactly $\frac{1}{2}$ [102]

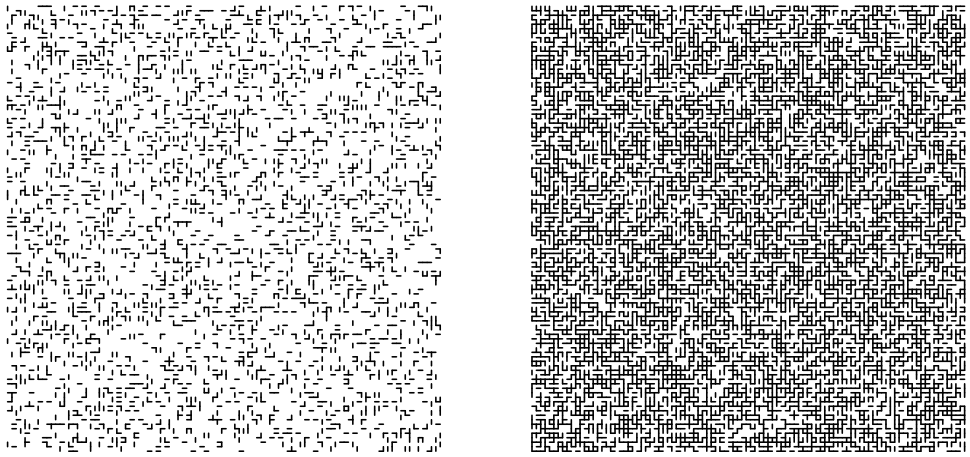


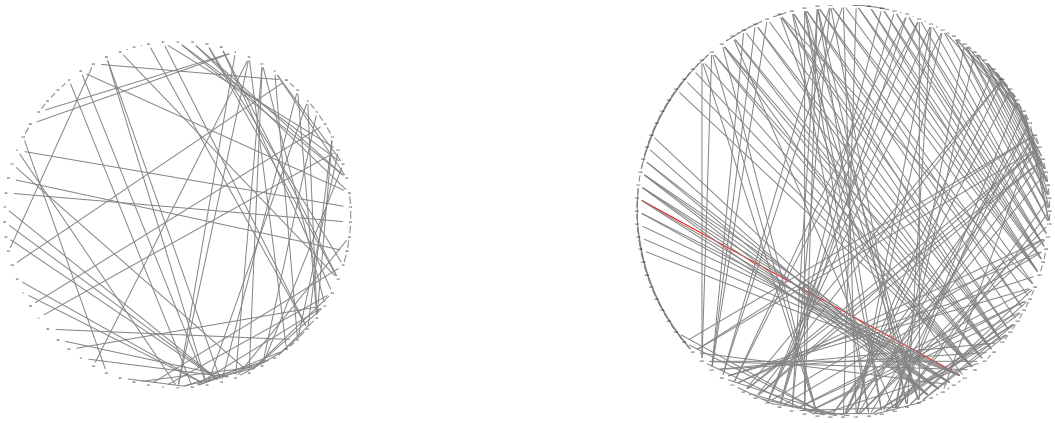
Figure 2.8: Bond percolation on two lattices with $N = 1000$, with $p = 0.15$ (left) and $p = 0.52$ (right). When $p = 0.52$, a giant component appears, connecting all four boundaries of the lattice.

Gilbert [74] studied what amounts to the bond percolation process on the complete graph: starting with the graph K_N, remove each edge independently with probability $1 - p$. Erdös and Rényi [61–63] studied a very similar process, in which for a given number of vertices N and number of edges M, from among all the possible combinations of M edges among the N vertices we choose one uniformly at random. The graphs generated by both of these processes are commonly

referred to Erdös-Rényi (ER) graphs. The graphs produced using Gilbert's method are denoted $G(n, p)$ [139]. These graphs are often used in simulation studies of random graphs, because the process of adding edges independently can easily be implemented computationally.

Like the lattice, in the limit of inifinite size, the ER graph has a threshold value of p below which there are only small components, and above which the majority of the graph is connected in one very large component. This component is commonly called the *giant component*, because it is usually very much larger than the next largest component [26]. A giant (or at least largest) component comprising most of the graph usually occurs even in quite small finite graphs, and the threshold is very low: in $G(1000, 0.0002)$, when the mean degree is 2, the largest component comprises $\approx 0.8N$. The degrees of ER graphs follow a Poisson distribution (see figure 3.1). Interestingly, when there is a giant component the graph has very small diameter, which grows as the logarithm of the size of the graph $log(N)$. The density of the ER graph is p, and its degrees are Poisson distributed with mean and variance $p(|V| - 1)$. The triangle density of the ER graph is generally quite low [139]. Figure 2.9a shows an ER graph with $N = 100$, $p = 0.02$.

Unlike many population contact structures, ER graphs are not very clustered, in the sense that their triangle density is usually very low. Watts and Strogatz [183] introduced a model for generating graphs that were both clustered and had low diameter. They begin with a regular graph or lattice with high clustering, such as a ring of vertices where each vertex is connected to the two closest vertices on either side. Such a graph is highly clustered, but since the only way to go from one "side" of the ring to the other is to travel along the circumference, it has very large diameter. To rectify this, a fraction of all edges are removed uniformly at random from the graph, and replaced with edges chosen as in the ER model. These "rewirings" dramatically reduce the diameter, because the new edges often cross the interior of the ring, and so one can take a short-cut along these edges from one side of the graph to the other [183]. The combination of the two properties of high triangle density and low diameter is referred to as the *small-world* property, and has been observed in many real-world networks [139]. Figure 2.9b shows a Watts-Strogatz graph with $N = 100$, a ring lattice structure where each vertex has has 6 neighbours (3 on each side), and a rewiring fraction $p = 0.025$.



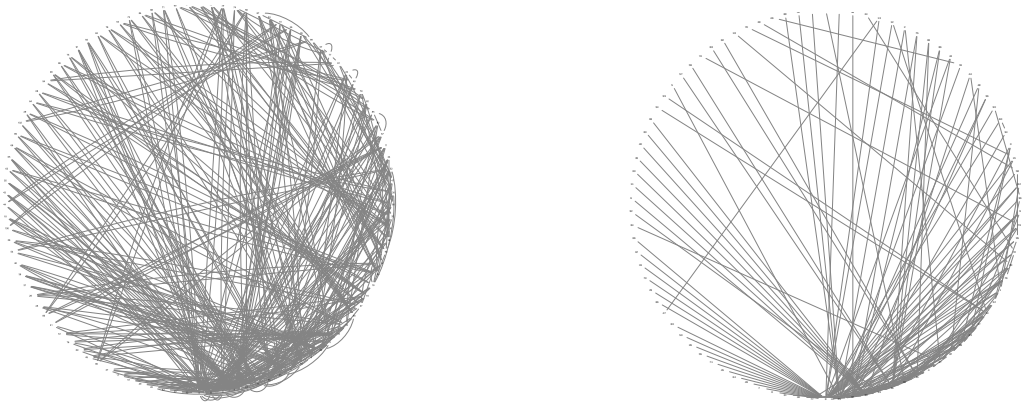
(a) (b)

Figure 2.9: (a) $G(n, p)$ Erdös-Rényi graph with $n = 100$ and $p = 0.02$; (b) Watts-Strogatz (WS) small-world graph with $N = 100$, with an underlying ring lattice with neighbourhood size 6, and a rewiring fraction 0.025.

Barabási and Albert [16] studied a method of constructing random graphs in which one begins with a small graph of size m and adds vertices one at a time, connecting each new vertex to m randomly chosen pre-existing vertices. If the m neighbours of the new vertex are not chosen uniformly, but rather with probability proportional to their degree, this approach is referred to as a *preferential attachment* model, and the resulting graphs as Barabási-Albert (BA) graphs. Like Watts-Strogatz (WS) graphs, BA graphs have the small-world property. Unlike either the ER or WS graphs, the degree distribution of BA graphs is heavy-tailed, in that while the majority of vertices have degree m or slightly higher, a small but non-zero fraction have much higher degree, which in the infinite graph has no theoretical upper limit, and even in finite graphs is usually several orders of magnitude higher than m. In particular, the degree distribution of BA graphs follows a *power-law*, so that the probability of a vertex having degree k is proportional to $\frac{1}{k^\alpha}$. The value of α depends on the value of m [16,139]. Figure 2.10a shows a BA graph with $N = 100$ and $m = 2$.

Power-law distributed graphs, and more generally graphs with heavy-tailed distributions, are of particular interest in epidemiology, because many studies have found that the contact graphs for various diseases have heavy-tailed distributions [71, 110, 151].

While the dynamics of outbreaks on ER, WS and BA graphs have all been extensively studied, these three generating algorithms are limited, in that each produces graphs with very specific statistical properties. While many graphs found in nature appear to be power-law distributed, it can be difficult to accurately distinguish between power-law distributions and other heavy-tailed distributions [139]. A model that provides greater control over the resulting graph, and in particular the degree distribution, is the *configuration-model* [133]. The configuration model is similar to the $G(n, p)$ ER model, but before edges are generated, each vertex is assigned a degree drawn at random from a specified discrete degree distribution. Edges are then assigned at random as in the ER model, but with the limitation that every vertex must end up with its pre-specified number of vertices. Because of its considerably flexibility, the configuration model has been widely adopted for simulation-based studies of outbreaks on random graphs [94]. Figure 2.10b shows a graph with $N = 100$ vertices and degrees following a truncated power-law distribution so that $\alpha = 1.5$, and degrees range from 1 to 20.



(a) (b)

Figure 2.10: Bárabasi-Albert (BA) graph with $N = 100$, $m = 2$ (a); Configuration-model truncated power-law distributed graph with $\alpha = 1.5$, minimum degree 1, maximum degree 20. Note that while both these graphs follow a power-law distribution, the configuration-model used here imposes a maximum degree, leading to wuite different-looking graphs.

2.4 Epidemiology on random graphs

Random graphs have been used in variety of ways to capture the structure of a host population. Vertices can be used to represent groups of hosts, such as all the individuals in a household (*household-network models* [14, 15, 146]), or a larger unit, such a as a town or city (*metapopulation models* [42, 82]). Here, I focus on models in which each vertex represents a single host.

In the basic compartmental model of outbreaks, the only thing that differentiates individuals is which compartment they are in. The system is therefore completely described by the number of individuals in each compartment at a given time, and a rule that determines how these numbers will change in time. When we introduce structure to the population by assigning each individual to a vertex in a graph, the individuals are no longer identical - they now differ in whom they can interact with and this needs to be taken into account in mathematical models of the system.

If we keep the assumption from the basic model that both infection and recovery (and more generally all transitions between compartments) are Poisson processes, then the system is a Markov chain [156], with q^N states, where N is the size of the graph, and q the number of compartments each vertex could be in (two for the SIS model, three for the SIR, four for the SEIR, etc) [176]. We can now calculate how the system transitions between states, and in theoty this gives us complete knowledge of the system [144]. In practice, because the number of equations to be solved grows as $q^N \times q^N$, this becomes computationally intractable for large N. Although some work has been done using this approach (see [39, 46, 165, 177], the focus of mathematical epidemiology on graphs has been on deriving and exploring more easily tractable models.

Pastor-Satorras *et al* provide a comprehensive review of the major methods [144]. Even more recently, Wang *et al* provided a comparison of the seven most commonly used models [182]. As the work presented in this thesis is entirely simulation-based, I will not be going into deep mathematical detail, but rather provide a brief overview of the main mathematical models, and some of the major results that have been derived.

An obvious place to start when trying to incorporate graph structure into the compartmental model of epidemic spreading is to note that the force of infection changes. In the well mixing case

the force of infection is given simply by βS. When infection can only pass along the edges in the graph, the force of infection experienced by any one vertex is limited by the number of neighbours it has. If we make the assumption that all vertices have the same degree - which is accurate in regular graphs and lattices but is otherwise a very strong assumption, this gives the *mean-field model* [182]:

$$\frac{dS}{dt} = \gamma I - \beta\langle k\rangle SI$$
$$\frac{dI}{dt} = \beta\langle k\rangle SI - \gamma I \tag{2.11}$$
$$\tag{2.12}$$

Note that the only aspect of graph structure this model incorporates is the mean degree - all other information, such as degree distribution, clustering and degree correlations, is lost. This simple model gives the infection threshold λ_c at

$$\lambda_c = \frac{1}{\langle k\rangle} \tag{2.13}$$

rather than 1 [182]. It is a fairly good predictive model for very regular graphs, such as the constant random graph in which each vertex has the same degree and all vertices are randomly connected [134], but for more heterogeneous graphs it is less accurate [65], and so more sophisticated methods are necessary.

The mean-field model considers every vertex to be equivalent, which ignores almost every aspect of graph structure. The *heterogeneous* or *degree-based* mean-field model (HMF) instead divides the graph into classes, where all members of a class have the same degree, and all members of a class are considered equivalent [143]. We then calculate ρ_k^i, the fraction of the degree class k that is in compartment i. We can then derive a general condition for the epidemic threshold

$$\lambda_c = \frac{1}{\Lambda_M} \tag{2.14}$$

where Λ_M is the largest eigenvalue of a *connectivity matrix* with elements $C_{kl} = k\mathbb{P}(l|k)$ and $\mathbb{P}(l|k)$ is the probability that a vertex with degree l has a neighbour with degree k [25]. This threshold holds both for the SIS model and the SIR model [120, 134]. In the case of uncorrelated networks the threshold can be further refined to

$$\lambda_c = \frac{\langle k\rangle}{\langle k^2\rangle} \tag{2.15}$$

For ER graphs, the threshold becomes $\frac{1}{\langle k\rangle+1}$ [182]. For power-law distributed graphs equation 2.4 implies that when the exponent $2 < \alpha < 3$ the threshold vanishes, since for exponentis in this range, $\langle k\rangle$ tends to infinity in inifinite graphs. This is distinctly different from the prediction of simpler models, which predict that there is a finite non-zero epidemic threshold.

The HMF model incorporates the degree distribution and degree correlation (assortativity) of graphs, but infection can pass between any two vertices, so that the specific edges of the graph are in effect removed. On way to think of this is that each vertex has a a number of "stubs" corresponding to its degree, and that infection is transmitted by two vertices connecting stubs briefly. This model is therefore exact for *dynamic graphs* in which edges are removed and replaced over time, and in which this process happens much faster than the epidemic process [181], but is an approximation for static graphs.

The *individual-based mean-field* or *quenched mean-field* (QMF) model [40, 78] incorporates much more of the graph structure. The state of each vertex is tracked, and by using the adjacency matrix of the graph, disease only passes along edges. If one assumes that the fates of neighbours are independent (that is, that $\mathbb{P}(v \in i, u \in j) = \mathbb{P}(u \in i)\mathbb{P}(u \in j)$ where v, u are vertices and i, j compartments) then one can obtain closed form expressions for the probability of infection for each vertex as a function of time [159–161, 176]. In the SIS model, the epidemic threshold can be calculated as

$$\frac{1}{\Lambda_1} \tag{2.16}$$

where Λ_1 is the largest eigenvalue of the adjacency matrix of the graph. In the SIR case the exact same threshold has been derived [189], but also shown to be a poor approximation [37].

The QMF approach assumes that the states of neighbours are independent, which is a strong assumption. An approach that takes the correlation between neighbours into account explicitly is the *pair-approximation* (PA) method [18, 58, 97, 99]. The equations tracking the number of individuals in each compartment in the basic well-mixing model are replaced by equations tracking the number of pairs of neighbours in each possible combination of compartments (SS, SI, IS, II, etc) [99]. By incorporating the adjacency matrix, the PA method preserves the structure of the graph [118].

Consider $[SI]$, the number of pairs of neighbours where one is susceptible and the other infected. This number shrinks whenever a) the infected neighbour in a pair recovers, or when the susceptible neighbour becomes infected, either b) by the infected neighbour, or c) by another neighbour that is also infected. $[SI]$ grows if d) one neighbour of a pair in the state $[SS]$ becomes infected. Situations c) and d) both involve a third vertex, connected to at least one of the pair, transmitting infection. Therefore, in order to calculate the change in $[SI]$ one must know the number of triples $[SSI]$ and $[SII]$. Clearly this argument can be extended to show that to calculate the change in these triple-states one must know number of quadruples, etc.

Although it is possible to obtain equations tracking all these classes [175], this is in practice very inefficient, and it is reasonable to assume that the correlations among sets of four or more vertices play a small role, compared to those of pairs and triples[6]. It is common therefore to close the equations by using an approximating term for the number of triples. Several different approximations are possible [38, 97, 99, 118]. The epidemic thresholds in the PA model depend on the approximation chosen.

These three methods, heterogeneous mean-field, quenched mean-field and pair-approximations, are the three most extensively studied approaches to epdiemic spreading in random graphs, but many others have been considered too. For single-wave models such as the SI and SIR, the epidemic spreading can be mapped exactly to a bond-percolation on the graph, so that standard techniques from percolation theory can be used to obtain thresholds [137].

Volz [179] considered the probability p_I that given a susceptible vertex a specific neighbour of that vertex is infectious, and using the probability generating function of the graph degree distribution developed a set of four nonlinear ODEs describing the progression of an SIR epidemic over time. This model can be related very directly to the pair-approximation method of Eames and Keeling [91], and is less general. However, it is significantly less computationally intensive, comprising only four equations, while the number of equations in the more general model grows with the maximum degree of the graph [144]. Volz' model has been further simplified by Miller [126] to a single equation.

[6] Although some models have been developed to efficiently incorporate higher-order correlations, and find that they play a singificant role in determining the epidemic threshold in some graph types [23]

The final approach I will mention is the message-passing approach of Karrer and Newman [94]. On graphs with no loops, they derive a set of exact equations giving the probability that a vertex u is in each state in terms of of the probability that it has not been infected by a particular neighbour u. For the more common case of graphs with loops, they redefine the graph in terms of a two-way directed transmission graph, in which each edge replaced by two directed edges which are each pre-assigned a time-to-infection. and infection is transmitted from vertex u to v only if u becomes infected and the time assigned to the edge is earlier than the time at which u recovers, in effect building the infection process into the graph. By considering a vertex in this graph that has no outwards edges, so that it may be infected but may not infect others, they then recover an upper bound on the extent of the infection.

The message-passing approach was originally applied to the SIR model, but has since been extended to the SIS model [164], and been further refined to allow more accurate estimates on graphs with loops [149]. Although the method in only approximate in the presence of loops in the graph, and is computationally intensive, it can be used to study epidemic models in which transition times between compartments are not exponentially distributed, which is a generally a requirement in other models and, as I have discussed above, is biologically unrealistic for many diseases.

The algorithm I use to generate graphs in this thesis is a slight modification of the standard configuration model. For the case of a single disease spreading on a configuration graph, considerable analytical results have been found, both for the proportion of the population that becomes infected [126, 179] and for the early-stage behaviour of the outbreak [12, 80]. However, to my knowledge these results have not yet been extended systematically to the case of multiple diseases spreading simultaneously, either independently of one another or with any form of inter-strain interaction.

R_0

Perhaps the most important result that arises from the basic compartmental model is the existence of an *epidemic threshold* at $R_0 = 1$. If $R_0 > 1$, then an outbreak will affect a large finite fraction of the population (SIR, SEIR and other single-outbreak models), or reach a steady state in which a constant fraction of the population is infected (SIS, SIRS, SIR with births and deaths, etc). On the other hand, if $R_0 < 1$, any outbreak dies out exponentially fast, and reaches a vanishingly small

fraction of the population.

Early estimates of R_0 for emerging epidemics that assume a well-mixing population sometimes give predictions of outbreak size that are wrong by many orders of magnitude [187], which has been attributed to, among other factors, the models not incorporating the structure of the population [125]. Because of the importance of determining R_0, and the unreliability of well-mixing estimates, a major focus of graph epidemiology has been to determine whether a threshold still exists, and if so, deriving expressions to calculate it. In each of the models I have discussed above, critical thresholds can be determined. Since the models are generally approximations of the system, the thresholds are estimates, and the thresholds given by different models to not always agree [144].

A very interesting early result of graph epidemiological models was the fact that the critical threshold generally depends on the reciprocal of the second moment of the degree distribution, $\frac{1}{\langle k^2 \rangle}$. Because this moment tends to infinity for power-law distributions with parameter $2 < \alpha < 3$, it implies that for these graphs, the epidemic threshold is zero [145]. Although estimates from different models differ in the values of α for which there is no treshold, this phenomenon occurs in most estimates [144]. However, it depends on the fact that there is no upper limit on the maximum possible degree in the graph, which in turn clearly depends on the graph being infinite, so that in real-world graphs, the non-existent threshold result does not hold perfectely [120]. Moreover, while may real-world graphs have heavy-tailed degree distributions, there are often constraints that impose a maximum degree. In general, if the transmission of a disease requires a substantial interaction between hosts (close proximity, skin touch, sexual interaction, for example), then the maximum degree is at the very least limited by the number of interactions an individual can manage to have during the day. Therefore the epidemic threshold generally does exist, but it is often extrremely low [182].

Calculating R_0, defined as the number of secondary infections caused by an infectious individual introduced into an entirely susceptible population, at first seems straightforward on graphs: simply multiply the per-edge transmissibility T by the mean degree $\langle k \rangle$. This definition is not accurate if clustering is relatively high, since if the neighbours of the index vertex are themselves neighbours, the index vertex may be prevented from causing some secondary infections because those neighbours have already been tertially infected. It is also not a good estimate if, as is the case in power-

law distributed graphs - most vertices have degree much lower than $\langle k \rangle$, so that in many cases the effective R_0 will be much lower. This illustrates that unlike in the well-mixing model, in heterogeneous graphs R_0 will be a necessary but not sufficient condition for an epidemic to spread significantly.

Alternative definitions have been derived, that compensate for this problem and generally recover the critical threshold property at $R' = 1$ (where I use R' to denote any definition of R^* or R_0 that is different from the definition in words in the previous paragraph). For example, Volz and Meyers [181], considering dynamics graphs, use an alteriative defimition that assumes that the vertex for which R' is calculated is a typical vertex very early in the outbreak, but not the first vertex. House and Keeling [90] use a similar assumption to derive an alternative definition of R'.

In this thesis, I will not generally be calculating any version of R_0 for my epidemics. There are two main reasons for this. First, the work here concerns the interaction of multiple circulating strains. For this reason I generally study parameters ranges for which there is always a major outbreak[7], and so R_0 is generally well in excess of 1. Secondly, my models are parametrised by the per-edge probability of infection (transmissibility, see Chapter 3) and R_0 depends on this, but also on a range of other parameters. I therefore feel it is clearer to discuss the effects of the parameters I explicitly vary in terms of the parameters themselves, rather than in terms of a compound parameter R_0.

2.5 Simulating outbreaks on random graphs

In this section I give a brief introduction to some approaches to simulating outbreaks on graphs. Generally, most mathematical theory that is developed is presented along with simulation-based results that verify the accuracy of the mathematical approximations. A comprehensive review is difficult to accomplish, because many authors do not specify the simulation methods they use, simply giving the results of their simulations.

There are two broad approaches to simulating disease spread on graphs: simultaneous update and continuous update models. In simultaneous update models at each time step the states of

[7]Some simulations still contain strains that failed to spread, because early in the outbreak "bad luck" can mean a strain did not cause any secondary infections.

all vertices (that require updating) are updated. The models are therefore discrete-time models. Simultaneous update models are generally simple to implement and give good results for prevalence and incidence rates [81]. However, the speed of the outbreak is to some extent determined by the size of the time step. Additionally, although in theory all state changes at a given time step are handled simultaneously, the fact that computers must process them one at a time imposes an order. This order is not usually determined directly by the algorithm, but rather by incidental aspects of the implementation. For example, if the vertices are assigned integers as identifiers and are stored in an array in the position corresponding to their identifier, they will usually be handled in the order of the identifiers; if they are stored in a hash-map keyed on their identifiers they will be handled in the order of their hashed identifiers, etc. This does not matter if we are only interested in how many vertices are in each compartment at each time step, but if for example we are interested in who infected whom, it is unhelpful that the order of processing the updates is not under our direct control or governed by a process that is a part of the model being studied.

Continuous update approaches implement a priority-queue model [144]. In this model, each state change has an assigned time at which it takes place. The state changes are placed in a queue sorted by their assigned time and processed in this order. This means that the order in which events are processed is random, and determined entirely by the algorithm. It also removes the somewhat unrealistic simultaneous update of the entire system.

Continuous update models can be used to simulate discrete-time models [86], but their real strength lies in the fact that they can simulate continuous-time processes as well. A very common approach is the Gillespie algorithm [75], a framework developed for simulating chemical reactions. The basic algorithm as applied to the SIR model on a graph is as follows:

- Choose an index vertex to become infected;
- for each of its neighbours, generate a random exponentially distributed time at which it will become infected. Place all these events in a queue ordered by their time;
- generate an event for the index case recovering, and place this in the queue;
- repeat this process until either no events remain in the queue, or some pre-determined limit (number of steps processed, size of outbreak, nubmer of recovered individuals) has been

reached.

This approach allows considerable flexibility and is statistically exact [144], but can be computationally demanding. Several faster approximate algorithms exist [76], as well as adaptations which are faster and exact [24].

A large class of important simulation models in epidemiology are *agent-based models*. In these models, each individual in a population is modelled explicitly, and can be assigned any number of behaviours. Such models are able to simulate almost any level of detail and can incorporate almost any features whose impact on the progression of an outbreak we wish to examine [54]. Many large agent-based simulations have been developed [17, 21, 36, 41, 51, 54, 56, 64, 131, 142]

These models can be useful in trying to evaluate the likely effeciveness of public health strategies to mitigate or prevent outbreaks [52,53,131,188], but are less useful for drawing general conclusions about underlying mechanisms, since the wealth of factors make it difficult to disentagle the relative role of each. For this purpose smaller, more specific simulations are generally used [86, 115, 179]

An important question when designing simulations is the parameters to consider. One of the advantages of mathematical models over computational ones is that, assuming they are analytically tractable, they allow one to understand exactly the effects of the factors considered. In constrast, in a computational model, as with experiments in general, one only ever precisely knows what the observed behaviour under the parameter combinations one has actually simulated, and one is forced to intra- or extrapolate to other parameter combinations.

Broadly, there are two approaches that are taken to choosing ones parameters. One can attempt a broad sweep of the parameter space, to get a picure of overall variability of behaviour. The larger the number of parameter combinations used, the better the picture [35, 95]. The other approach is to attempt to model a specific disease as closely as possible using available empirical estimates, and vary the parameters around these in a narrow range, so that the parameters considered can be understood to be biologically relevant [85, 111].

2.6 Epidemiology of coinfection on random graphs

One of the main reasons to study cocirculating epidemics is that diseases may interact and modfiy each others' outbreaks. There are many ways in which they can do this - by altering a host's immune responses, by altering a host's behaviour, or by changing the probabiltiy of severe complications of infection.

Immune modification interactions

Perhaps the most well-studied class of strain interactions is immune modification, in which infection with one strain alters a host's susceptibility to subsequent infection with another strain. Such interactions can be broadly classified into *competitive* interactions, in which coinfection reduces the ability of one or both diseases to spread; and *cooperative* interactions, in which infection with one disease enhances the spread of the other in some fashion. Examples of competitive interactions include *cross-immunity*, in which infection with one strain confers complete or partial immunity to other strains, as is the case with Influenza A [27, 47]; and resource-competition within the host, which has been documented between different strain of the malarial parasite *Plasmodium chabaudi* [49, 50]. An example of a cooperative interaction is infection with one strain or disease reducing a host's resistance to infection with another. A classical example is the Influenza A H1N1 andemic of 1918-1919, in which it has been hypothesised that much of the observed mortality was due to the viral infection facilitating bacterial pneumonias [34]. In the modern day, HIV can be a driver of the spread of other diseases by increasing hosts' susceptibility. An example that is of major public health conern is the increased prevalence of tuberculosis among HIV positive people in large parts of the world [33, 43].

Competitive interactions

Competitive interactions have been studied in several different frameworks. A generalised framework exists for the SEIR system in the well-mixing case [178], but I restrict the discussion here to work on multiple epidemics on graphs.

The majority of the literature assumes that the two diseases are spreading in a *multiplex graph*, that is in several graphs that share vertex sets but have distinct edge sets. Several models exist,

and much work has been done (*e.g.* [70, 115]). I summarise the most significant below.

Funk and Jansen [70] developed an analytic model for this case that allows the study of the spread of two pathogens where infection with one causes complete or partial immunity to the other. They found that the efficacy of immunisation depended both on the degree of heterogeneity in the number of contacts each host has, and the correlation of the number of contacts each host has in each graph - when the two graphs are highly correlated an immunity-granting first strain dramatically reduces the potential spread of a subsequent strain, largely through the effect of immunisation of crucial high-degree vertices. When the two graphs are not strongly correlated this heterogeneity interferes with the effect of the immunity-granting strain, because well-connected vertices in one graph are likely to become immunised, but are not likely to be well-connected in the other graph and so have lower impact on the spread of the second strain.

A separate analytic model was developed by Marceau *et al* [115]. This model considers a discrete-time SIR model, and shows broadly similar results. They apply their model to the case of uncorrelated graphs, and both partial and complete immunity, but in this case granted by a strain b which begins to spread *after* the other strain. They obtain similar results to Funk and Jansen [70]

The case of multiple outbreaks spreading through the *same* graph has also been considered. The first significant work in this regard was carried out by Newman [138], who derived analytic results for two outbreaks spreading in the same network, but where one outbreak had been driven extinct prior to the introduction of the second, and where the first granted complete immunity to the second. The key finding was that the effective removal of a large portion of the graph by the action of the first strain introduced a *coinfection threshold* in the infectiousness of the second strain, above which it would still be able to spread successfully in the partially immunised graph. This threshold is considerably higher than the classical threshold for outbreak size seen in the well-mixing case and for some graph types.

Karrer and Newman [95] developed a framework based on earlier work by Newman [138], again assuming that infection with one strain granted complete immunity to the other and using a discrete-time epidemic model. They showed that in most cases one disease came to dominate the graph

with the other only infecting a tiny fraction of vertices, but that for certain infection rates the two strains could coexist.

Miller [127] developed an analytic model largely analogous to that of Karrer and Newman [95], but based on a conceptually much simpler framework the author developed earlier [128].

Cooperative interactions

Cooperative dynamics have been less exhaustively studied, although there are good biologcal reason to want to study this system - a number of diseases are known to "cooperate" in that infection with one can make infection with the second more likely [69,114,158,174]. Models assuming well-mixing have been developed, finding complex coexistance and oscillatory dynamics [117,178]. Cooperation between pathogens has also been confirmed in *in vivo* experiments [171]

Although several mathematical models have been developed, to my knowledge almost all assume a multiplex graph, in which each disease spreads along edges unique to that disease (see for example [10,130,157]). The model of Newman and Ferrario [140] remains the only one to study cooperative coinfection in a single graph. They studied a system of two diseases on a single graph, in which prior infection with one disease was a requirement for infection with a second. They found that in ER-type graphs there were two disctinct epidemic thresholds of the first disease, a lower threshold for that disease to cause an outbreak, and a higher threshold for the second disease to spread through the outbreak of the first. In contrast, the epidemic treshold in a scale-free network was zero for both diseases, indicating that an outbreak of both always occured.

Grassberger *et al* [81] recently conducted an extensive series of simulations to examine the behaviour or existence and extinction thresholds in a system of two cooperating strains on a range of graph structures.

Transmission modification interactions

Another broad class of interactions between co-circulating strains or diseases is *transmission modification*, in which infection with one strain does not necessarily affects the host's susceptibility to subsequent infection with another strain, but the ability of the second strain to be passed on to

other hosts is altered. Such interactions can be due to direct effects of coinfection, such as increased HIV plasma viral loads in people who are also infected with malaria [87,132], or due to behavioural changes. For example, infection with a disease that causes an otherwise highly gregarious individual to be bedridden will reduce that person's ability to transmit a second milder infection contracted at roughly the same time. Conversely, hospitalisation due to a severe infection may increase the ability of a non-gregarious individual both to contract and transmit a second disease.

As a separate mode of strain interaction, transmission modification has received far less attention than immune modification, partly because immune modification implicitly affect the ability of a host to transmit infection: if the host does not become infected it cannot transmit, and so reducing the chance of infection will also reduce the chance of transmission. However, it worth considering transmission modification on its own, so that its effects can be disentangled from other epidemiological consequences of immune modification. To my knowledge there are no models which explicitly consider transmission modification on its own and not as a consequence of immune modification. I consider this class of interactions in chapter 5

2.7 Recombination

One of the main reasons to study the dynamics of coinfections is the potential of coinfection to complicate the prevention and treatment of infectious diseases. Different forms of competition among co-circulating strains has been shown to be a potential driver of increased virulence through mutation [3,121,122]. Co-infection is a necessary requirement for recombination, the other major source of genetic diversity in pathogens. Here, I define recombination very generally as any mechanism by which a novel genetic sequence is produced through the mixture of elements from two *parent* sequences. Under this very broad definition many major human pathogens undergo recombination.

Both theoretical models [155] and empirical studies [100] imply that recombination can cause the rise of multiresistant strains in HIV, although some models predict that recombination will instead prevent the rise of such strains [68], or that recombination will occur too infrequenctly to have much effect [105].

Recombination can also lead to the rise of new strains to which the host population has little immunity, and which consequently cause large epidemics or even pandemics. A prominent example of the latter is the 2009 H1N1 influenza pandemic, which was caused by a novel strain that arose as the result of multiple recombination events in humans and pigs [166]

Recombining mechanisms can differ dramatically between different types of pathogens. Here, I will briefly describe recombination in HIV and reassortment in influenza.

In HIV, each virion contains two strands of RNA, and when virions are formed in cells infected with two different strains, these can become packaged into the same virion [104]. The subsequent reverse transcription stage will then generate sequences composed of segments of each sequence [153], becaues it can switch which strain is being used as the template [147]. This template switching can occur up to 12 times per replication cycle [190]. Recombination tends to occur more in specific "hot spots" along the genome [9, 167, 190]. The recombination rates also depend on the host cell type the virus is replicating in [45]. Figure 2.11 illustrates this process.

Reassortment, the recombining mechanism in influenza, is very different. The genome of influenza A consists of eight separate strands, each encoding for one or two proteins [173]. During the assembly of new virions, if the genomes of more than one virus strain are present in the cell, *reassortant* virions containing segments from each strain can arise [136]. Figure 2.12 illustrates this. Between two strains there are thus $2^8 = 256$ possible reassortants. Not all reassortants are viable [169].

Unlike HIV, where infection is permanent, so that recombination has a long time window in which to occur, influenza infections are short-lived, so reassortment can only happen if two infections happen close in time. When two influenza strains were allowed to infect the same Guinea pig but with a delay between the infections, a delay of more than 18 hours was sufficient to prevent any reassortment, and the optimal delay was between 6 and 12 hours [169]. A comparable window was also found in cell culture experiments [116]

Although recombination has been modelled extensively in well-mixing populations, to my knowledge there has only been one study, that of Buckee *et al* [35] that explicitly models recombination between strains on graphs. They found that, in a system with complex dynamics regulating strain

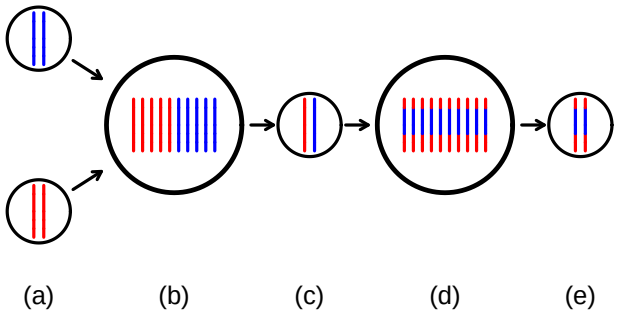


Figure 2.11: Recombination in HIV. (a-b) Two strains infect a cell; (c) one strain from each strain become packaged in to the same virion;(d) new recombinant strain is generated during the reverse transcription phase in a subsequent infection, (e) resulting in a recombinant virion. Figure based on Gasunov *et al* [72].

diversity, recombination was not a primary driver of strain diversity. This was because hosts that had been infected with a given strain would then be partially resistant to all recombinants of that strain. This causes recombinant strains to become extinct almost immediately, due to the tight clustering of parent and recombinant strains in the random graphs under study. It is perhaps because of this lack of a significant effect that so far modelling work on this problem has been limited. However, as I will show in chapter 6, when recombinant strains are not recognised by vertices that have been infected by the parent strains, complex dynamics arise.

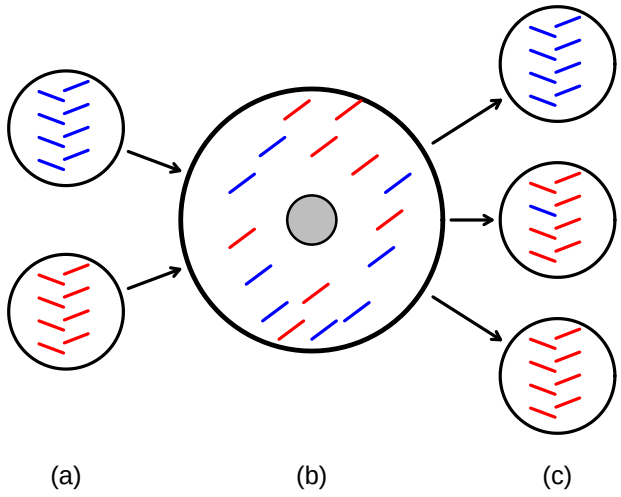


Figure 2.12: Reassortment in influenza. (a) Two influenza strains infect the same cell, (b) the segments are replicated, (c) segments are packaged into new virions, potentially mixing segments from both lineages. Figure based on Trifonov *et al* [173].

Chapter 3

The model

3.1 Overview

In this chapter I describe the model I used for all simulations. In the first section I describe the algorithm I used to build graphs, in the second the results of some tests for correctness and a comparison to a different software package implementing a similar algorithm.

In the third section I describe the algorithm used to model the spread of disease on the graph. I compare the output of the model to that of another simulation package and to expected values from the mathematical literature. As some of my model choices are uncommon in the literature, in the fourth section I explore the spread of a single strain with the exact model I will use in subsequent chapters, to allow easier comparison with the multiple-strain case. Finally, in the fifth section I summarise the model used, consider its applicability to modeling specifid real-world diseases, and present a reference table of the parameters of the model and the values and value ranges I use.

I make no attempt to calibrate the time-scale of this model to any data from the real world. Instead, the timing is given relative to the mean duration of infection, which is set as 1 time unit ($t.u.$). Wherever the timing of events in the epidemic simulations are discussed, times are given in $t.u.$

All the experiments I describe in this thesis were done using a simulation package written by me. The package is written entirely in Clojure, a Lisp dialect that runs on the Java Virtual Machine (JVM). I used Clojure 1.6.0, the last stable version at the time the code was written. Random numbers were generated using the Uncommons Maths library by Daniel Dyer [57], which ports the Mersenne Twister algorithm of Matsumoto and Nishimura [119] to Java. Priority queues were handled using the Clojure *priority-map* package, by Mark Engelberg [60].

To compare the output of my code, both for graph generation and disease simulation, to similar software, I used EpiFire by Hladish et al [86].

For all algorithms used, a pseudocode version of the algorithm is provided at the end of the chapter.

3.2 Host population structure model - random graphs

I model the host population as a random undirected simple graph, with hosts occupying the vertices and infection passing along the edges connecting them. I generate the graphs using a variation of the configuation model of Molloy and Reed [133]. Like the standard configuration model, I begin by assigning to each vertex v_i a degree k_i drawn from a specified degree distribution.

The standard configuration model then follows the Erdös-Rényi model [61] in adding edges between random pairs of vertices to the graph one at a time, until each vertex v_i has k_i neighbours. In contrast, I proceed by choosing one vertex v_i and assigning to it k_i randomly selected neighbours. I then repeat with the remaining vertices. I call my method a *vertex-focused* method, because it proceeds one vertex at a time, in contrast to the *edge-focus* of the standard configuration model. The vertex-focused method produces very similar graphs to the standard method, but allows for an optimisation which the edge-focused method does not, which will be discussed briefly below. Figure 3.16 gives my modified configuration algorithm in pseudocode, with the standard configuration algorithm for comparison given in Figure 3.17.
The standard configuration model makes no guarantees about any structural aspects of the resulting graphs beyond the distribution of its degrees. In particular, the resulting graphs frequently contain a small number of self-edges and parallel edges, and so are not simple graphs. They also often

contain more than one component, and so are not connected graphs. The configuration is a bond percolation the set of vertices, and for all cases I shall consider here there is a giant component that consists of over 90% of the vertices.

The size of the giant component is usually so great, and the number of self- and parallel edges so small, that one can simply use the resultant graph [129]. If it is desirable to have a strictly simple graph, several modifications of the standard model exist. One approach is to generate the graph according to the standard model, and then remove any non-simple edges. The resulting graph can then be used directly, or a number of edges can be added, equal to the number removed. This final step is done to partially correct the slight change in degree distribution that results from removing non-simple edges.

The second approach that can be taken is to ensure that self edges and multiple edges are never added to the graph by drawing the two ends of an edge without replacement, and by checking that an edge is not a duplicate of an existing edge before adding it to the graph. This produces and equivalent deviation in degree distribution to removing non-simple edges at the end without replacing them. It has the advantage that it is a single-pass method - once the final vertex has been generated the process is complete. I have chosen the latter approach.

Degree heterogeneity has been shown to have a major effect on the spread of infectious disease in graph. In this thesis I compare the behaviour of outbreaks on two types of graph with very different degree heterogeneity. The degrees in the homogeneous graphs are Poisson distributed, while in the heterogeneous graphs they are Zipf distributed. The Poisson distribution takes one parameter, λ, and its probability mass function is given by

$$P(x = k) = \frac{\lambda^k e^{\lambda}}{k!} \tag{3.1}$$

Like the normal distribution the Poisson distribution gives a "bell-curve" probability mass function (Figure 3.1) that is symmetric and quite closely clustered about the mean value λ, because for the Poisson distribution the variance σ^2 is equal to the mean λ, so that in a graph with this degree distribution the degree of most vertices is very close to the mean degree.
The Zipf distribution is a truncated discrete power-law distribution with support on the set $\{1, \ldots, k_{max}\}$. It takes two parameters, the exponent parameter α and the maximum value k_{max}. It's probability

mass function is given by

$$P(x = k) = \frac{1}{H_{\alpha,k_{max}} k^{\alpha}} \tag{3.2}$$

where the normalising constant $H_{\alpha,k_{max}}$ is the k_{max}th generalised harmonic number of α, given by

$$H_{\alpha,M} = \sum_{i=1}^{k_{max}} \frac{1}{i^{\alpha}} \tag{3.3}$$

The Zipf distribution is a heavy-tailed distribution, in which the probability of generating values of k close to k_{max} is fairly high (see fig 3.2a). In contrast to Poisson distributed graphs, in Zipf distributed graphs most vertices have very low degree, but there exist a few vertices with degree close to k_{max}, and a small class in between. Zipf distributions are often plotted as a double logarithm plot, in which case their distributions become approximately a straight line (Figure 3.2b). In all the graphs I study in this thesis, $k_{max} = 150$. This value was chosen so that the graphs would have some vertices with degree an order of magnitude greater than the mean, while still producing very sparse graphs. It also lies in the range of estimates of Dunbar's number - an estimate on the upper limit of social contacts humans are capable of maintaining [55], whiche has received some empirical support [79], so that for the purposes of this rather theoretical work, I shall take it as a reasonable upper bound on the number of contacts. It seems in any case a more plausible model than a power-law distribution with no finite upper limit - there must be a limit on how many contacts an individual can have, if for no other reason than that they must sometimes sleep.

Since the two degree distributions have different parameters, comparing them is not necessarily straightforward. Perhaps the most obvious approach is to compare graphs that have the same mean degree $\langle k \rangle$. I take this approach here, and for the remainder of this thesis, graphs will be referred to by the distribution type (Poisson or Zipf) and $\langle k \rangle$.

A note on optimising graph generating for small outbreaks

Why use the vertex-focused method at all, if the edge-focused method produces equivalent graphs and is already well established? The vertex-focussed algorithm allows for *lazy* or on-the-fly graph building. Notice that the spread of infectious disease is a process that starts at one vertex, then spreads to some of its neighbours, then spreads to some of *their* neighbours, and continues like this until it has reached all vertices it will reach. Suppose a vertex is never part of the outbreak.

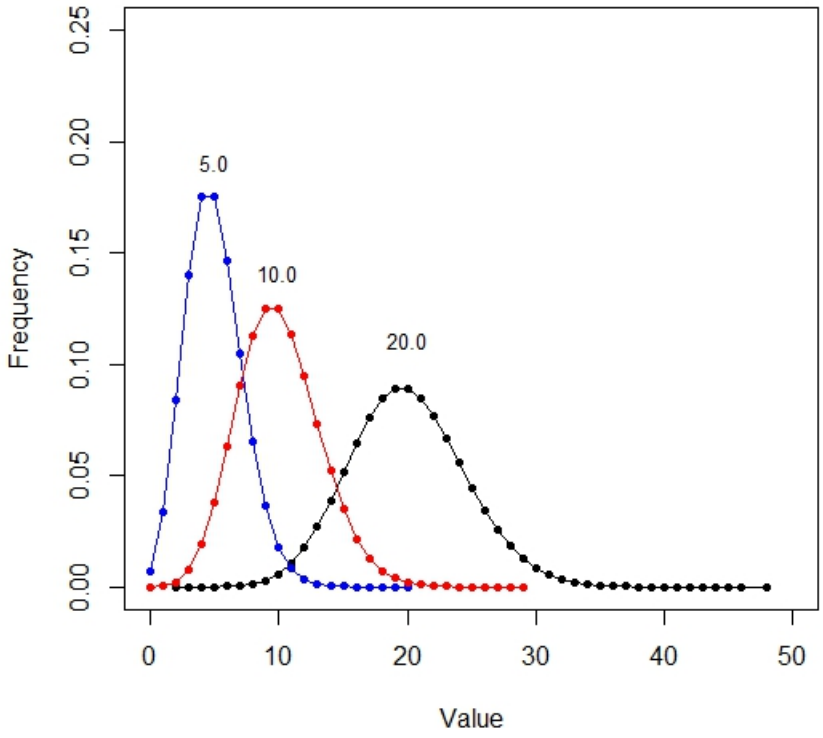
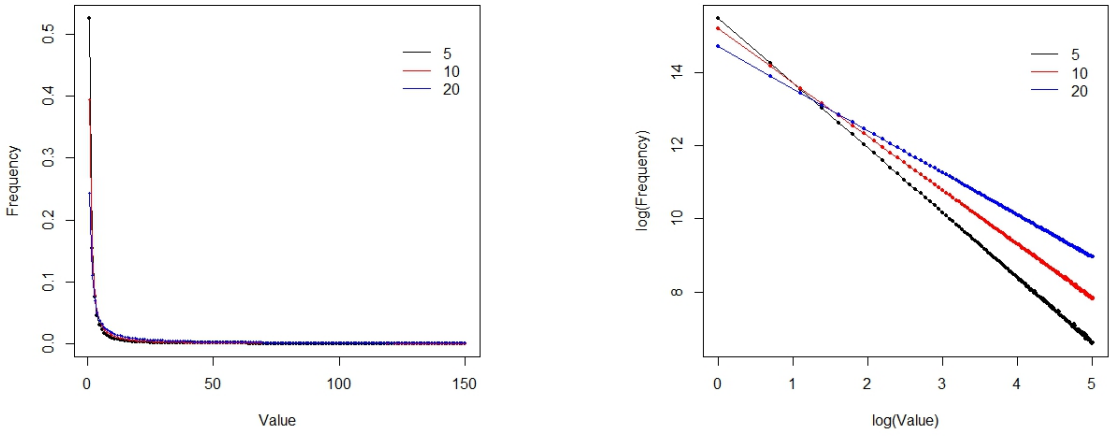


Figure 3.1: Poisson distributions with $\lambda = 5.0$ (black), 10.0 (red) and 20.0 blue.

Is there then any reason to generate that vertex? If we can avoid generating vertices that will remain uninfected the graph generating process can be speeded up. The problem is that because the simulations are stochastic we do not know at the start of the outbreak exactly which vertices will be infected.

By modifying the vertex-focussed algorithm slightly we can get around this problem. Instead of adding vertices to the graph in random order, we choose one random vertex and assign a degree d and d neighbours, and assign a degree to each neighbour. We then initialise the spread of the disease at this vertex, and determine which of its neighbours will become infected. These then are assigned neighbours, which are assigned degrees. We continue in this fashion until the disease has



(a) (b)

Figure 3.2: Zipf distributions ($k_{max} = 150$, $\alpha = 1.77$ (black), 1.47 (red), and 1.15 (blue), giving means of 5, 10 and 20 respectively. Distribution plot (a) and log-log plot (b). Note that while the frequency of higher values is low in (a), it is not zero.

run its course. Only edges with at least one end in the outbreak have been generated, and only those vertices that were either infected or had an infected neighbour were assigned degrees.

Is this useful? Consider the somewhat daunting problem of modelling the population of the UK as a graph with approximately 65 million vertices. If we are simulating the spread of HIV, which affects less than 1% of the population [103], this method can reduce the size of the generated graph by a factor of 100. Provided the resulting graph exhibits the same statistical properties, this method would represent a very significant speed-up.

This approach is not novel. The idea of evaluating parts of large data structures only as they are explicitly needed is widely used in many branches of computer science, and is usually known as *call-by-need* or *lazy* evaluation. Ball and Neal used the lazy evaluation strategy in developing a deterministic model of epidemic spread on a random graph with optional additional global spread outwith the graph [13]. Marceau *et al* develop a similar framework for the case of two graphs sharing vertices but with independently generated random edge sets [115]. The same group implement a lazy graph evaluating algorithm and test its performance compared to an *eager* algorithm that generates the entire network ahead of time [141]. They report very close correspondence in results of the two algorithms, and as one might expect the lazy algorithm is significantly faster. Despite this, lazy graph generating algorithms do not appear to have been widely adopted for simulations.

In this thesis I do not use lazy network generation, since the majority of outbreaks I consider reach a very large fraction of the population. The graphs considered are also not very large (N = 10000), and so the potential speed-up is limited. Lazy evaluation is the main reason to prefer vertex-focused network generation. Neither the standard edge-focussed configuration model, the Barabási-Albert, Erdös-Rényi or Watts-Strogatz generating algorithms can be optimised in this way, because while the graph is being generated the degrees of the vertices are not guaranteed to be the same as they will be in the final graph. Thus an outbreak spreading on an incomplete graph generated using these algorithms is spreading on a subgraph, but not an induced subgraph.

3.3 Testing the random graphs

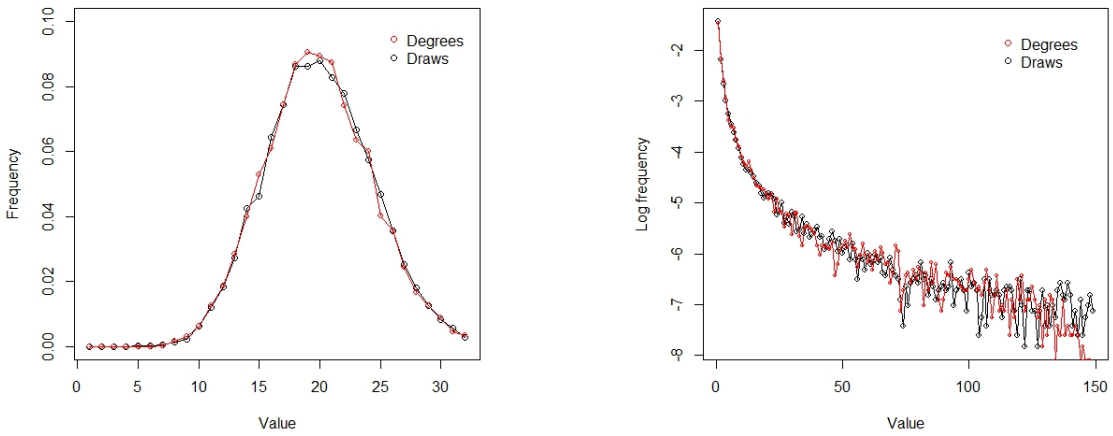
Both the vertex-focused algorithm and the approach of avoiding adding non-simple edges are uncommon choices in the literature. I therefore carried out several analyses to measure the deviation of the resulting graphs from the theoretical ideal. Since the algorithm is designed to produce graphs with a specific degree distribution it is important that any deviation from that distribution is not large.

Since the graphs are intended to be used to simulate the spread of infectious disease, it is also important that if the graph is not connected there is a giant component and that this component is large, since an outbreak that starts in one component will only infect that component.

Figure 3.3 compares the normalised distribution of 10000 draws from a Zipf distribution ($\alpha = 1.15$, $k_{max} = 150$) and a Poisson distribution ($\lambda = 20$) to the normalised distributions of two graphs with 10000 vertices with degrees generated from the same distributions. The visible deviations are relatively small, and the results of a Kolmogorov-Smirnov two-sample test give poor support to the hypothesis that they are drawn from different distributions (Zipf case D=0.0738, p=0.8115; Poisson case D=0.0938, p=0.999). The deviation from specified degree distribution appears to be fairly minor.

Figure 3.4a shows the mean size of the giant component in Poisson and Zipf graphs, and Figure 3.4b the mean number of components. The giant component is larger in the Poisson graph because in the Zipf graph many vertices have degree 1 or 2, and if these vertices are connect to each other by chance there are no edges remaining with which to connect them to the giant component. The mean size of the small components is almost exactly 2 in all cases. In the Poisson graph there are far fewer vertices with very low degree, and when $\lambda = 20$ there are almost none, and so no small components form.

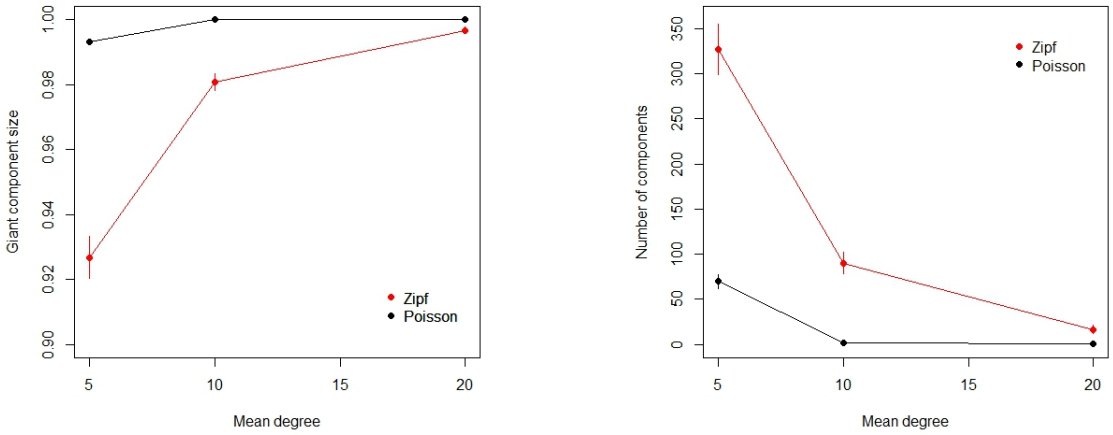
The smallest giant components are produced in Zipf distributed graphs with low mean degree. Even in the case of mean degree 5, which is the smallest I will be considering, the giant component comprises on average 93% of the graph, which is sufficient for a very large epidemic to break out.



(a) (b)

Figure 3.3: Log of normalised degree distribution for the Zipf distribution (a) and degree distribution of the Poisson distribution (b). The black line shows the results of 10000 draws from the random number generator for each distribution, the red shows the degrees of a graph with N=10000 generated from distributions with the same parameters.

In table 3.1 I compare several metrics of graphs generated by the vertex-focussed configuration algorithm to graphs generated by the edge-focussed configuration algorithm. To generate graphs using the edge-focussed algorithm I used the EpiFire epidemic simulation package by Hladish *et al* [86]. The two algorithms generate graphs with very similar mean degree, giant component size and number of components. Although neither algorithm makes any guarantees about the triangle density or assortativity of the resulting graphs, since both properties affect the spread of disease, I measure these too. For all tested graphs except Zipf distributed graphs with mean degree 20 the graphs generated with the vertex-focussed algorithm have higher triangle density. There is no consistent pattern to the difference in assortativity between the two algorithms, but all graphs generated are disassortative, and the Zipf distributed graphs are much more strongly disassortative than Poisson distributed graphs.



(a) (b)

Figure 3.4: Mean size of the giant component (a) and mean number of components (b) in Zipf (red) and Poisson graphs (black), as a function of mean degree. The size is given as a fraction of total graph size. All data points are the mean of 100 graphs. In the case of the Poisson graph with mean degree 20 the giant component is always the entire graph, so the graph is connected.

Algorithm	Degree distro	Specified mean	Actual mean	Triangle density	Assortativity	Giant component	Number of components
Edge	Poisson	5	5.004	0.000499	-0.08299	9928	72
Vertex	Poisson	5	4.995	0.00848	-0.00348	9930	69
Edge	Poisson	10	10.025	0.001029	-0.00094	10000	1
Vertex	Poisson	10	9.988	0.001494	-0.00301	10000	1
Edge	Poisson	20	19.985	0.001986	-0.00301	10000	1
Vertex	Poisson	20	19.973	0.002798	-0.0005	10000	1
Edge	Zipf	5	4.743	0.0255	-0.086	9348	326
Vertex	Zipf	5	4.906	0.0582	-0.179	9267	327
Edge	Zipf	10	9.468	0.0493	-0.085	9383	80
Vertex	Zipf	10	9.639	0.0619	-0.131	9807	90
Edge	Zipf	20	20.617	0.0583	-0.082	9963	17
Vertex	Zipf	20	19.537	0.0458	-0.072	9967	16

Table 3.1: Comparison of graph metrics for graphs generated using the standard configuration model with non-simple edges removed implemented in EpiFire (Edge) and graphs generated using the vertex-focused configuration model as implemented by my code (Vertex). Poisson graphs were generated with $\lambda = 5.0$, 10.0 and 20.0; Zipf graphs with $k_{max} = 150$, $\alpha = 1.77$, 1.47 and 1.15.

3.4 The disease model - priority queue compartmental models

The basic model

In this thesis I model the spread of disease using a continuous-time priority-queue, based on the Gillespie algorithm [75]. The basic algorithm for the case of an SIR model is:

- Choose an index vertex u at random. Add the event $u : S \Rightarrow I$ with time 0.0 to the queue
- Take the first item in the queue. The system time is now the time t_u of this event.
  - If it is an infection event $u : S \Rightarrow I$ check if u is in state S. If so
    - set the state of u to I.
    - For each neighbour w of u, draw a time of infection t_w from an exponential distribution with parameter λ. If t_w is less than the infection duration Γ, infection succeeds. Add the event $w : S \Rightarrow I$ to the queue with time $t_u + t_w$.
    - After infection events for all neighbours have been attempted, add the event $v : I \Rightarrow R$ to the queue with time $t_u + \Gamma$.
  - If it is a recovery event $u : I \Rightarrow R$, set the state of u to R.
- Repeat the previous step until the queue is empty, or some other pre-defined end condition is reached.

This procedure simulates an SIR model where infection passing between two vertices is a Poisson process λ, and where the infection duration is a constant Γ. The probability that a vertex v infects a neighbour u is then given by

$$1 - e^{-\lambda\Gamma} \tag{3.4}$$

Following Karrer and Newman [94], I will call this quantity the *transmissibility* T. Instead of setting the value of λ directly, we specify the desired transmissibility and calculate the required value of λ as

$$\lambda = \frac{ln(T)}{-\Gamma} \tag{3.5}$$

Optimisations and extensions

For a susceptible vertex u that is infected late in the outbreak, there may be up to k_u events added to the queue for infecting u. Of these, only the one with the earliest associated time will be successful. Since many outbreaks grow exponentially, and each vertex v that becomes infected may generate up to k_v infection events, the queue can become very large as the outbreak grows. This is further compounded when we study models with multiple strain spreading simultaneously. We can reduce the size of the queue however: whenever an event $u : S \Rightarrow I$ with time t_{u_1} is generated, we associate t_{u_1} with u in the graph (u has not yet been infected). If another event $u : S \Rightarrow I$ is generated with time t_{u_2}, we only add it to the queue if $t_{u_2} < t_{u_1}$, since otherwise we already know that the second event will fail to infect.

Each infection of a vertex v with k neighbours adds between one and k events to the queue (one recovery event and at most $k-1$ infection events), so that the total number of events per simulation is at most $m \sum k_i$, where m is the number of strains, and k_i is the degree of vertex $i \in \{1 \dots N\}$. Of these events, a fraction $1 - T$ will have an infection time after v recovers and not be added to the queue, so that the expected maximum number of events for m independently spreading strains becomes

$$\sum_i k_i \sum_j^m T_j(I) \tag{3.6}$$

This implies that two strains spreading independently with $T = 0.5$ for both strains would generate approximately 200000 events. By using the optimisation just descried to exlclude events that cannot succeed, the total queue lenght is dramatically smaller. When $T = 0.5$, each strain in practice adds approximately 25000 events, or approximately a four-fold reduction.

This optimisation is only logically consistent for models without return to the S state (SI, SIR, SEIR, etc). If a vertex can become susceptible again (SIS, SIRS, etc), infection events that were assumed doomed to fail may now be valid and so having failed to add them to the queue breaks the logical consistency of the queue. These events can be "recovered" in a statistically correct fashion, but this requires additional calculation and can add significantly to the runtime of the code. In this thesis I will not be considering models that allow return to the S state, so I do not implement this.

One consequence of using a continuous-time model is that the smallest possible time increment between events is the smallest value that can be used by the programme, in this case the smallest values that can be stored in a Java double. This means that the chance that any two events have exactly the same time assigned to them is vanishingly small, since it would require that they were randomly signed exactly the same double. However, because each vertex that is infected will attempt to infect all its neighbours, as the size of the outbreak grows the number of events being added to the queue grows, and so the time between events shortens. This can lead to extremely short periods between a vertex becoming infected and that vertex passing on the infection, on the order of 10^{-d}, where d is the number of digits stored in the double. As I show below this leads to very fast epidemic outbreaks. In most scenarios the model could be applied to there is a minimum time ϵ needed between a host becoming infected and the pathogen gaining a sufficient foothold for the host to become infectious[1]. Adding such a minimum time to the model makes the behaviour of the model more realistic. This effectively adds the exposed (E) step to the model, so that an SIR model becomes an SEIR model. In this thesis I will therefore be studying SEIR models.

To implement this in the model, we add infection events to queue with time $t + \epsilon$, and calculate λ using the equation

$$\lambda = \frac{ln(T)}{-(\Gamma + \epsilon)} \tag{3.7}$$

The focus of this thesis is on the dynamics of multiple strains spreading simultaneously on the same graph. Extending the basic queue model to handle multiple strains takes two steps. First, tag the state of each vertex with the strain it refers to. A vertex in state I_2R_1 is infected with strain 2 and has recovered from infection with 1. Second, to allow each strain to have its own disease dynamics, add a second set of parameters for the second strain, so that an SEIR system with two strains 1 and 2 spreading has the parameters Γ_1, Γ_2, ϵ_1, ϵ_2, λ_1 and λ_2. Figure 3.5 shows the compartment diagram of the model for two strains. The rates denoting transition from $E \Rightarrow I$ and $I \Rightarrow R$ have been omitted, as these are not varied in the simulations presented in this thesis.

In Chapters 5 and 6 I model more complicated systems in which strains infecting the same host can interact or generate novel strains. This requires extending the model further. These extensions

[1]This is not true if the incubation period is a vanishingly small fraction of the infectious period, such as Herpes simplex and HIV.

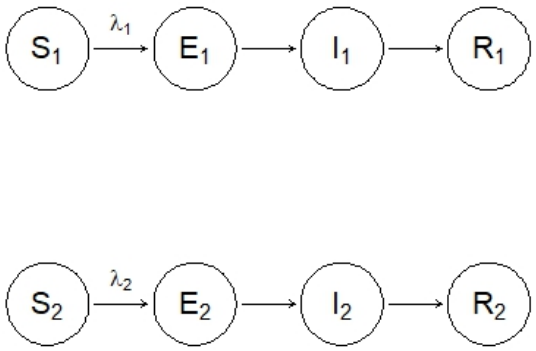


Figure 3.5: SEIR diagram for two independent strains. The upper line denotes the state of the vertex with respect to strain 1, the lower with respect to strain 2. The "rates" of transition $E \Rightarrow I$ and $I \Rightarrow R$ are left off.

are described in the relevant chapters.

In stochastic simulations, not all outbreaks succeed. If an outbreak begins in a vertex with few neighbours, it may by chance fail to infect any of its neighbours, and so the disease never spreads beyond the index case. The greater the degree of the index vertex, the smaller the chance of this happening. This means that if we are simulating epidemics in heterogeneous graphs, if we are interested only in the dynamics of large outbreaks and we want n repeat runs of a simulation, we must often run far more than n simulations in order to get enough in which an outbreak took place, and these "successful" simulations will mostly be those with high-degree index cases.

Since I am only interested in the dynamics of epidemic outbreaks, rather than the probability of an outbreak occurring, I modify the process for selecting an index vertex. The index vertex is selected at random from among all vertices whose degree is exactly the mean. In addition to

ensuring that the likelihood of the disease spreading beyond the index case is fairly good, it also largely prevents the index case from being in a small component, since almost without fail the small components consist of vertices with only one or two neighbours. This guarantees that the rate of failed outbreaks is fairly modest, although it does not altogether eliminate them, especially when both T and $\langle k \rangle$ are low.

Figure 3.18 shows the full algorithm for the model in the case of multiple strains as pseudocode. In order to keep the readability of the algorithm fairly good it does not include the modifications necessary for cross-immunity or recombination between strains.

3.5 Testing the disease model

To verify that the model behaves correctly I compare its output running a single SIR model to that of EpiFire, and to the output of a numerical simulation of mean-field approximation of the standard SIR model on a graph with constant degree.

Quantifying epidemics

There are many ways one can measure the extent and severity of an outbreak of infectious disease. One very common measure is R_0. For reasons I have outlined in the previous chapter, I do not directly measure R_0 in this thesis. The main metric of an outbreak I am interested in is the size and the speed of the outbreak.

I use two measures of the size of an outbreak of strain i: $I_i(t)$, the size at time t; and O_i, the total outbreak size. We have

$$O_i = \int_0^\infty I_i(t)dt = R_i(t), t \to \infty \tag{3.8}$$

For most simulations I will report O_i, as well as show the prevalence curves of the strains, that is the time series of $I_i(t)$. For legibility, subscripts denoting the strain are omitted whenever the strain in question is unambiguous.

There are many ways to measure the speed of an outbreak. A common measure is the *serial interval*, the average time between an individual becoming infected and transmitting the infection on. The serial interval is implicitly specified in my model, since for strain i is simply $\epsilon_i + \lambda_i^{-1}$. Another common measure is the *incidence* of the outbreak, the number of new infections in a given period of time. Although this would capture the speed of the outbreak, I have opted to measure the simpler measure of *peak outbreak time*, the time at which the outbreak is largest. Note that the peak time occurs at the maximum of the prevalence curve, at which point

$$\frac{dI}{dt} = 0 \tag{3.9}$$

In later chapters I will introduce other measures to capture other quantities as the simulations become more complex.

The tests

The EpiFire package provides several different algorithms for simulating infection spread. The most comparable to my algorithm is the Binomial Chain model. This is a discrete-time priority queue. The model takes two parameters: transmission probability T^{EF} per unit time, and infection duration Γ. Time to infection is drawn from a truncated geometric distribution with parameter T_E and support on the set of integers $\{1, \ldots, \Gamma + 1\}$. Transmission succeeds if the time drawn is less than $\Gamma + 1$. Thus, to compare my model to EpiFire I replace the exponential distribution to determine time to transmission with a truncated geometric distribution, and run the model with only one strain, and $\epsilon = 0$. Figure 3.20 shows the pseudocode of the EpiFire Binomial Chain algorithm, Figure 3.19 the modified discrete-time version of my model.

Figure 3.6 shows the results of my model running this discrete-time mode, with $T_E = 0.1$, $\Gamma = 5$. Individual runs are shown in grey, and the average of 100 runs is shown in red. The black line shows the average of 100 runs of EpiFire on the same graph, with $T = 0.1$, $\Gamma = 5$. The graph has a Poisson degree distribution with mean $\lambda = 20$. It is apparent that the two models give the same results. A K-S two-sample test (D=0.0995, p=0.8817) also suggests good agreement between the models.

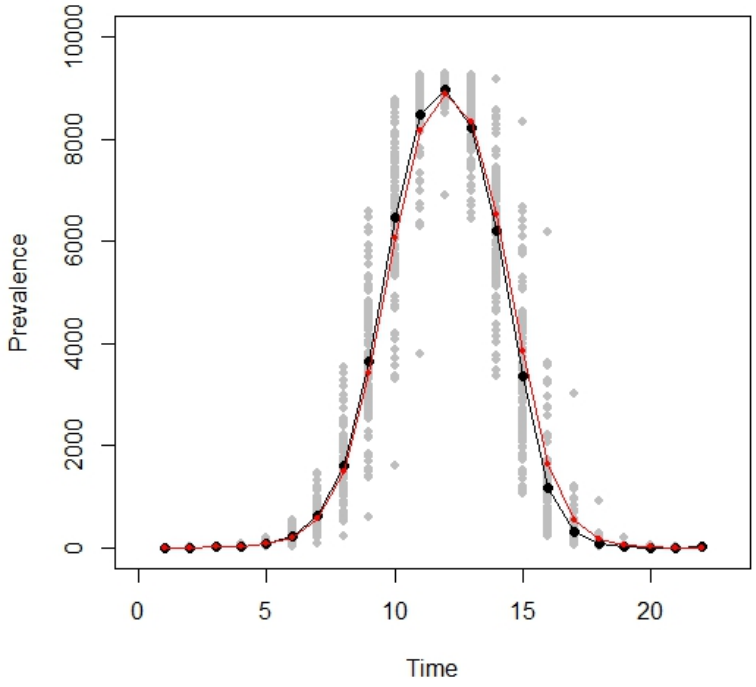


Figure 3.6: Prevalence curve of a discrete-time priority queue SIR model run on a Poisson distributed graph with $\langle k \rangle = 20$. 100 simulations run in my model with $T = 0.1$, $\Gamma = 5$ are shown in grey, and their average in red. The average of 100 runs of The BinomicalChain algorithm run in EpiFire on the same graph with the same parameter values is shown in black.

A well-mixing population of size N is equivalent to the complete graph with N vertices, and the standard SIR ODE system should be equivalent to the continuous time queue model, with the modification that infection duration is exponentially distributed, rather than constant, with $\lambda = \beta$ and $\Gamma = \gamma$.

However, my code stores graphs as hashmaps. The graphs I consider in this thesis are all very sparse, i.e. $\langle k \rangle$ is usually two orders of magnitude less than the size of the graph. For sparse graphs the hashmap approach is much more efficient than the alternative matrix approach. However, for very dense graphs it becomes very inefficient. The complete graph is the densest possible simple

graph, and for even reasonably small graph sizes the model runs slowly.

One partial solution to this is to run the simulation on a configuration model graph with constant degree k: vertices are attached at random such that each vertex has d neighbours. I compare this to the prediction from the simplest model that incorporates graph structure, the mean-field model [134]:

$$\frac{dS}{dt} = -\beta\frac{\langle k\rangle}{N}IS$$
$$\frac{dI}{dt} = \beta\frac{\langle k\rangle}{N}IS - \gamma I$$
$$R = N - S - I$$

Figure 3.7 shows the results of this comparison. The red line shows the mean of 200 simulations on a constant-degree configuration graph with $N = 10000$, $d = 100$; the black line is the results of the ODE model. The simulations were run with $\lambda = 0.005$ and $\Gamma = 0.06$. The ODEs were calculated with parameters $\beta = 0.005$, $\langle k\rangle = 100$, $\gamma = 0.06$ and initial conditions $S_0 = 9999$, $I_0 = 1$. The result from the ODE lies within the 1s.d. of the mean simulation result, but the result is clearly not identical. There are two reasons for this. This is due to the fact that the mean-field model does limit infections to vertices that the mean-field approximation does not account for clustering between vertices, which will reduce the number of susceptible vertices that each infected vertex can infect.

These two comparisons show that the software I have produced behaves comparably to other simulation pacakges and to mathematical theory. However, all the work presented in this thesis concerns an SEIR model with exponentially distributed time to infection, and constant infection and incubation durations. This is diffferent from both the basic SIR model and from that used in EpiFire's Binomial Chain simulator. In the rest of this chapter I illustrate how each of these differences affects the model, and then give a quick overview of how my model behaves in the single-strain case. Red lines in figures indicate the model being used for this thesis, and in the case of infection and incubation duration, which are not varied in the experiments in the main part of this thesis, the red line shows the parameter values used.

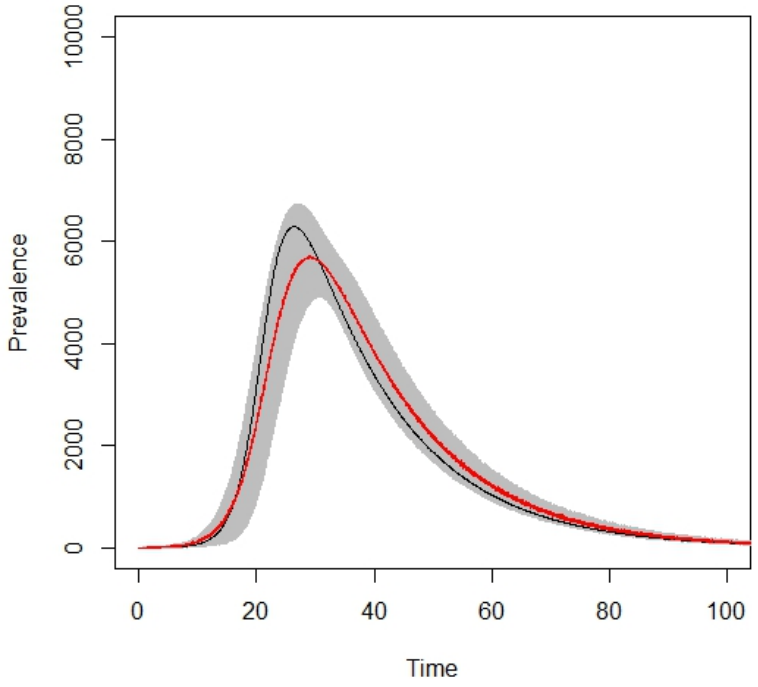


Figure 3.7: Results of 200 simulations of continuous-time priority-queue SIR ($\lambda = 0.005$, $\Gamma = 0.06$) on a constant-degree random graph (N=10000 and $\langle k \rangle$=100). The red line indicates the mean prevalence curve, and the grey area the 1s.d. range around the mean. The black line shows numerical simulation of the SIR ODE ($\beta = 0.005$, $\Gamma = 0.06$).

The simulation model used to generate the results shown in Figure 3.7 assumes that both the time to infection and the infection duration are exponentially distributed. The assumption that the infection duration is exponentially distributed is made for the basic SIR model, but it is known to be inaccurate. The infection duration of many diseases is typically fairly invariant, being clustered around a particular value [88, 185], suggesting that a normal distribution may be more realistic. Figure 3.8 compares simulations of three models with the same dynamics for determining time of infection, the same mean infection duration, and run on the same graph, which is Poisson distributed with $\langle k \rangle = 20$. Changing from an exponential distribution of infection duration (dotted black line)

to a normal distribution (solid black line) quite substantially changes both the spead and maximal size of the outbreak. Changing from a normal distribution with mean and standard deviation $1.0t.u$ to a constant infection duration of $1.0t.u$. (red line) has minimal effect on the outbreak dynamics. Since a constant infection duration will lead to faster simulations, I have chosen to use a constant infection time in this model as an approximation to a normal distribution.

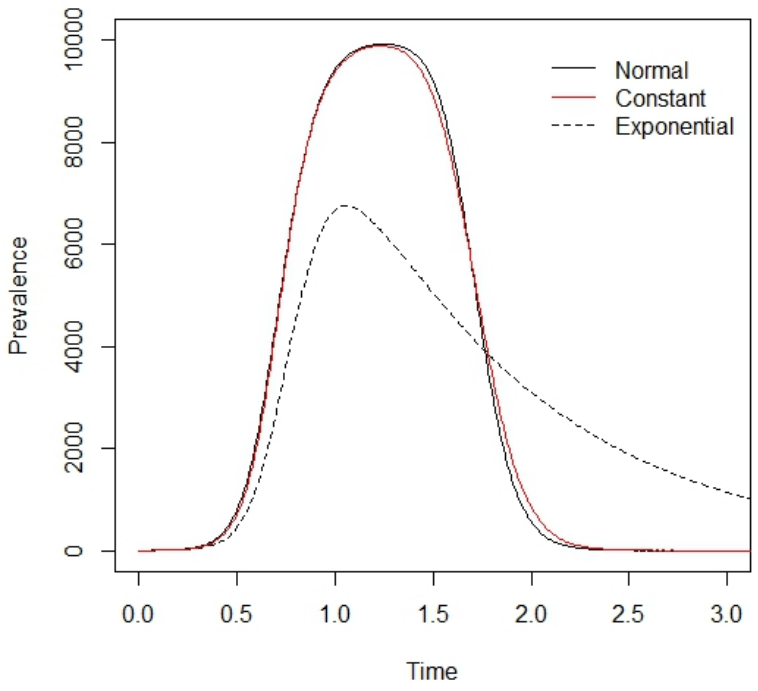


Figure 3.8: Three SIR models on a Poisson distributed graph with $\langle k \rangle = 20$. β was the same for all three simulations, as was Γ. Three different distributions of infection duration were used: exponentially distributed (mean = Γ, black dotted line); normally distributed ($\mu = \sigma = \Gamma$, solid black line), and constant Γ (red line).

The BinomialChain model in EpiFire is a discrete-time model. As discussed above, changing to a continuous-time model both speeds the growth of the outbreak and increases the maximum size. Figure 3.9 shows that the speed difference between continuous-time (solid black line) and

discrete-time (dotted black line) can be substantial, and that introducing an incubation period (red line) moderates this effect, removing the effect of cases with implausibly short incubation time. I will be using an SEIR model with incubation time $0.1t.u$ throughout.

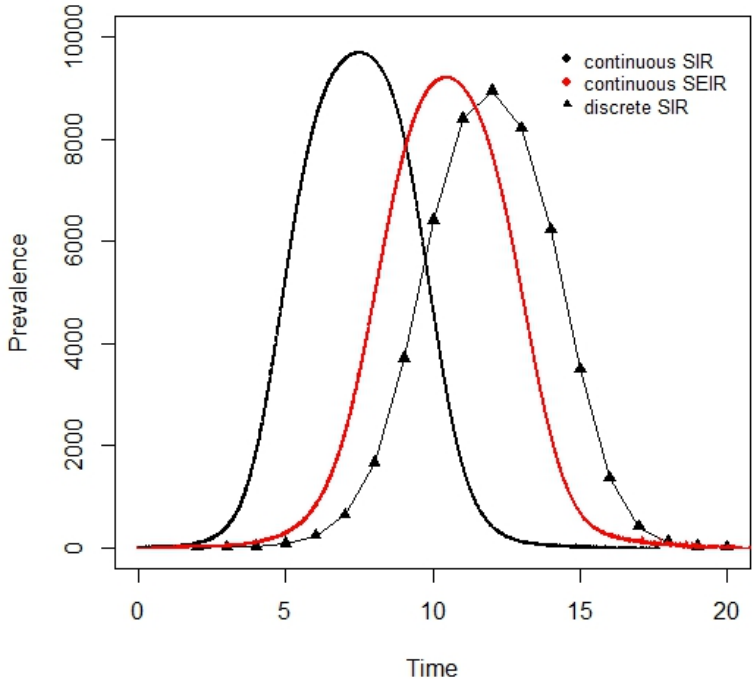


Figure 3.9: Discrete-time (dotted black line) and continuous-time (dotted black line) SIR simulations with identical parameters, compared to a continuous-time SEIR model (red line).

Figure 3.10 shows the effect of varying the transmissibility between 0.1 and 0.4 in both Poisson and Zipf graphs. As T increases the outbreak grows faster, peaks earlier and at a greater size, and reaches a larger proportion of the population. In a Poisson distributed graph with mean 20, over 90% of the population is infected for $T \geq 0.2$ (Figure 3.11)

Unsurprisingly, the outbreaks are much smaller in the Zipf graph. This is due to the large portion of the graph that is poorly connected - if a vertex with only one neighbour fails to be infected by

that neighbour, it will never become infected.

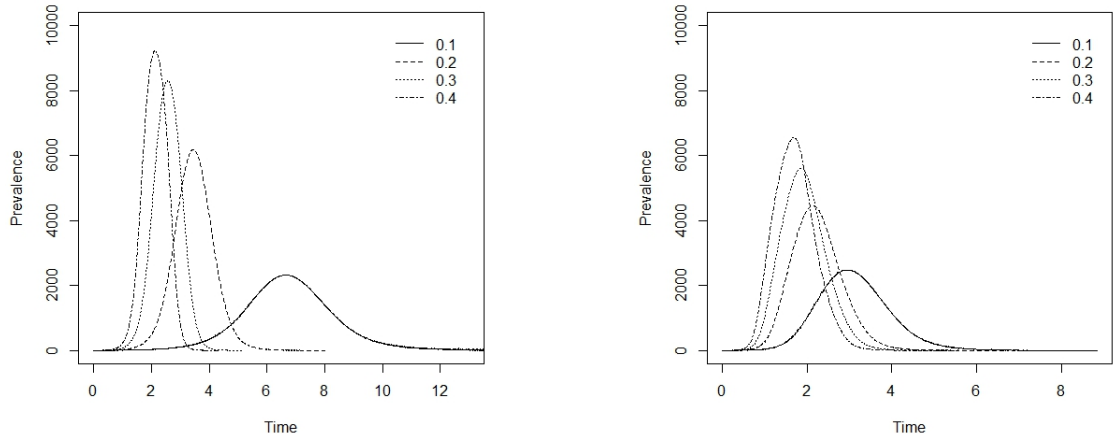


Figure 3.10: $I(t)$ of outbreaks with $\epsilon = 0.1t.u$ and $\Gamma = 1.0t.u.$, varying T (given in figure legend), in graphs with Poisson (a) and Zipf (b) degree distribution.

Figures 3.12 and 3.14 show that both longer infectious periods and longer incubation period slightly delay and reduce the peak outbreak. However, the overall size of the outbreak is not affected (Figures 3.13 and 3.15. This is the case in both Poisson and Zipf graphs. This is not a fundamental property of the underlying disease model being studied, but rather is because the infection duration and incubation duration are used to calculate T. This means that the λ parameter is adjusted to compensate for the changes in infection and incubation duration, and so the final outbreak size ought to remain largely unaffected. While the model could be adjusted to allow control of incubation and infection period independently of transmissibility, I have chosen not to vary the duration of the infectious or incubation period in the experiments presented in this thesis.

3.6 Summary

In summary, the model I will be using is an SEIR model, simulated using a continuous-time priority queue. Time to infection is drawn from an exponential distribution with parameter λ.

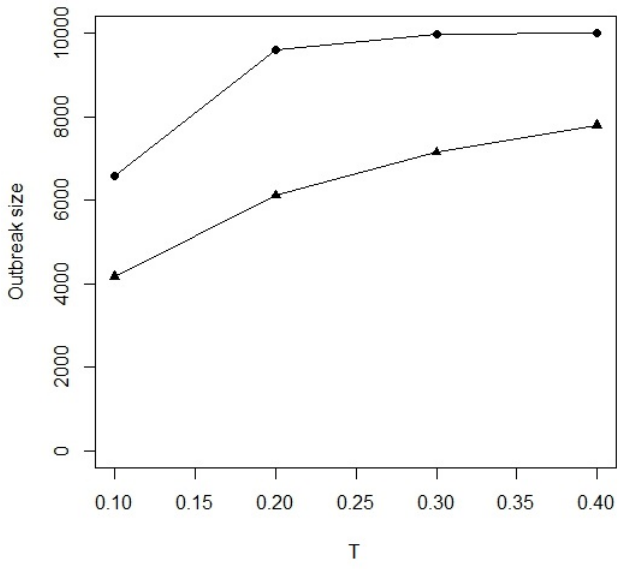


Figure 3.11: O of simulations with $\epsilon = 0.1t.u$ and $\Gamma = 1t.u.$, varying T, in graphs with Poisson (circles) and Zipf (triangles) degree distribution.

Infected vertices become infectious after an incubation period ϵ. Infection lasts for a set duration Γ. When multiple strains are spreading, each strain has its own exponential distribution to determine time to infection, but all strains have the same incubation and infection durations. The host population is simulated as N identical individuals located at the vertices of a graph. The graphs are generated using a vertex-focused configuration model and have either Poisson or Zipf distributed degrees, specified by mean degree $\langle k \rangle$. The parameters, along with the values and ranges of values I use are summarised in table 3.2. Additional parameters that are required for subsequent simulations are specified in the chapters detailing those simulations.

These parameters are chosen partly due to computational constraints, and partly out of biological plausibility. The maximum degree in the Zipf distributed graphs (k_{max}) was chosen to lie in the middle of the range of estimates for Dunbar's number [55]. The values for the mean degree ($\langle k \rangle$) were chosen to be one to two orders of magnitude lower, and spanning a fairly broad range. This lead to values of $\langle k \rangle$ somewhat higher than are often used for simulations validating mathematical

Parameter	Value (range)	
Mean degree	$\langle k \rangle$	5, 10, 20
Power-law exponent	α	1.77 ($\langle k \rangle = 5$), 1.47 (10), 1.15 (20)
Maximal degree	k_{max}	150
Transmission time parameter	λ	0.1 - 0.9
Infected period	Γ	$1.0t.u.$
Incubation period	ϵ	$0.2t.u.$
Index degree	N/A	$\langle k \rangle$

Table 3.2: Basic parameters of graph and epidemic models with values or value ranges, not including adjustments required for subsequent chapters.

models (*e.g.* [179]). The resulting graphs are therefore denser and more computationally intensive. This placed a practical upper limit on $\langle k \rangle$ of 20, and also on the size of the graphs ($N = 10^5$).

The population model I use in this thesis does not include any demographic dynamics (birth, death, migration). This means that it is a better fit for outbreaks that happen on sufficiently short time-scale that changes in the population size and composition are negligible [29], such as individual outbreaks of measles [83], influenza [191]. SEIR model is classically applied to measles [83], but if different strains are treated as separately spreading diseases, as is the case here, it can also be applied to influenza. The requirement is simply that the incubation period is an appreciable fraction of the duration of the infection, and that individuals cannot become infected with the same strain twice, either due to immunity or death [29]

In order to keep the number of parameter combinations manageable, both Γ and ϵ are fixed. Since Γ is defined to be $1.t.u$ the disease model is parameterised by λ and ϵ. The ratio $\epsilon/\Gamma = 0.2$. This values is far longer than for HIV, where the incubation period of 2-4 weeks is a tiny fraction of the untreated infection duration of up to ten years [2], but less than for measles or ebola, where the incubation period is as long or longer than the infectious period [67]. It is perhaps closer to that of influenza A, where the incubation period is up to two days [67], and the host remains infectious for around seven days [106]

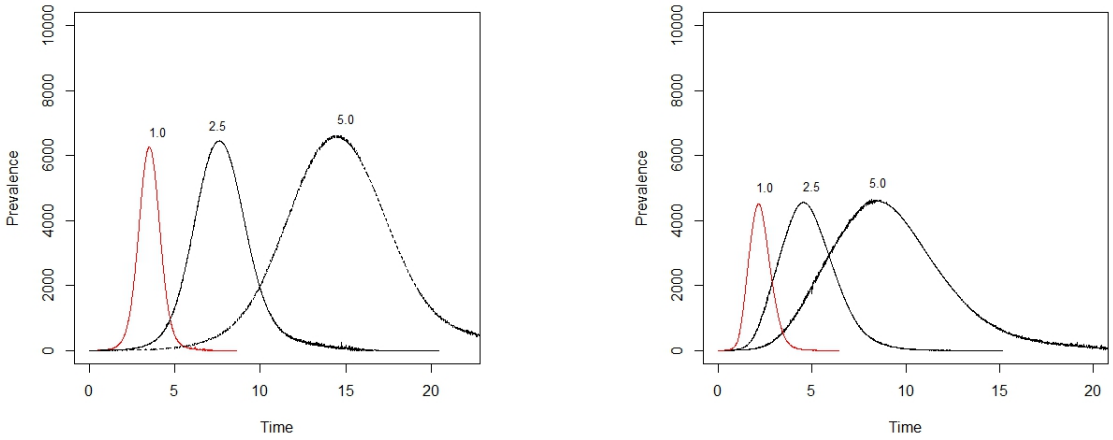


Figure 3.12: $I(t)$ of outbreaks with $T = 0.2$, $\epsilon = 0.1t.u$, varying Γ (labels on curves in figure), in graphs with Poisson (a) and Zipf (b) degree distribution.

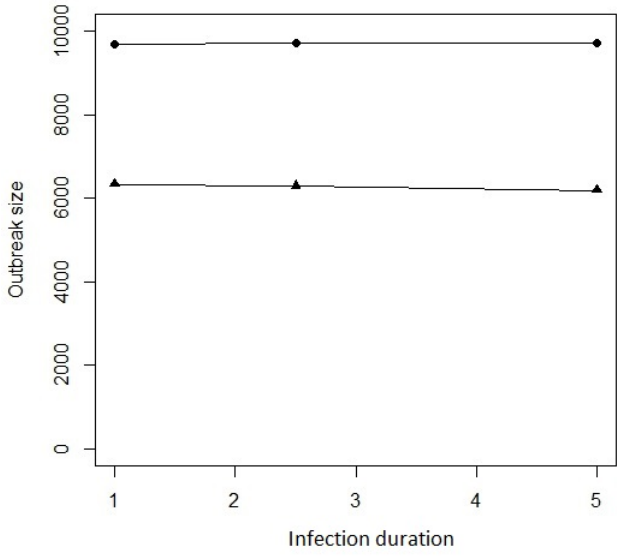


Figure 3.13: O of simulations with $T = 0.2$, $\epsilon = 0.1t.u$, varying Γ, in graphs with Poisson (circles) and Zipf (triangles) degree distribution.

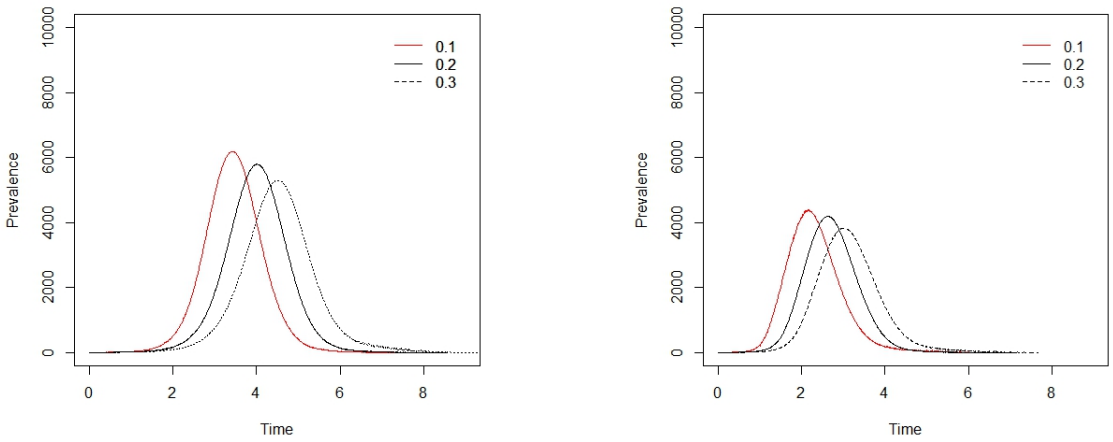


Figure 3.14: $I(t)$ of simulations with $T = 0.2$, $\Gamma = 1.0t.u$, varying ϵ (values in figure legend) in graphs with Poisson (a) and Zipgf (b) degree distribution.

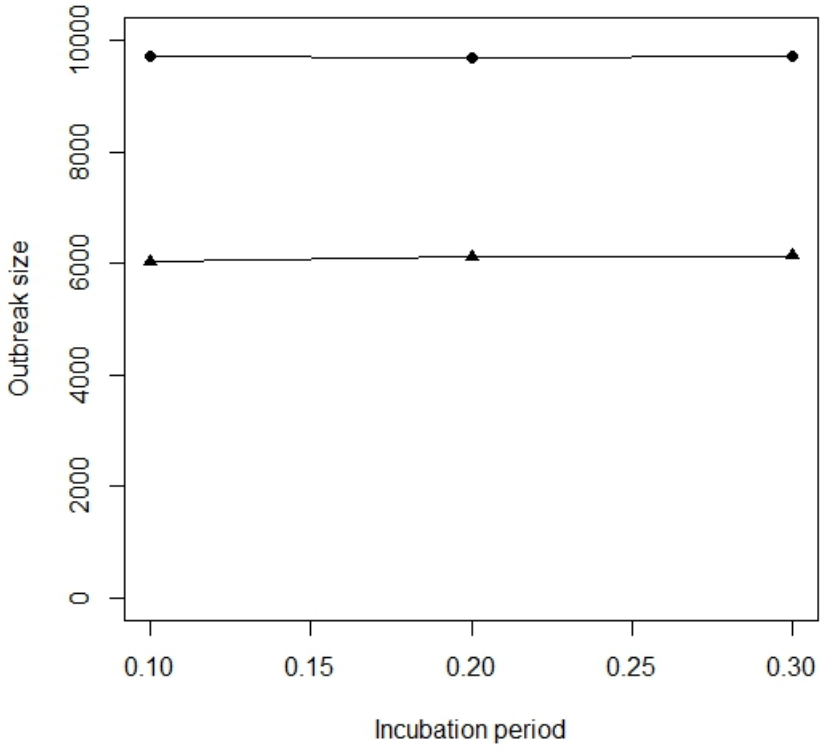


Figure 3.15: O of simulations with $T = 0.2$, $\Gamma = 1.0t.u$, varying ϵ in graphs with Poisson (circles) and Zipf (triangles) degree distribution.

```
Vertex-focused graph builder

Input Vertex-set V, degree-distribution D
Procedure
    for vertex v in V:
        degree(v) = next-integer(D)
    Vertex-set completed-vertices = ()
    while V not empty
        v = first(V)
        Targets = rest(V)
        do degree(v) times
            u = choose random element from Targets
            add u to neighbours(v)
            add v to neighbours(u)
            remove u from Targets
        for w in V
            if w has degree(w) neighbours
                add w to complete-vertices
                remove w from V
     return completed-vertices
```

Figure 3.16: Pseudocode rendition of the vertex-focused configuration algorithm used to generate graphs throughout this thesis.

Eedge-focused graph builder (configuration model)

```
Input Vertex-set V, degree-distribution D
Procedure
    sum-of-degrees = 0
    for vertex v in V:
        degree(v) = next-integer(D)
        sum-of-degrees += degree(v)
    edges = sum-of-degrees / 2
    Vertex-set Targets = V
   while edges ≠ 0
        v = choose random element from Targets
        u = choose random element from Targets
        add u to neighbours(v)
        add v to neighbours(u)
        if v has degree(v) neighbours
            remove v from Targets
        if u has degree(u) neighbours
            remove v from Targets
        edges -= 1
    return V
```

Figure 3.17: Pseudocode rendition of the standard edge-focused configuration algorithm in wide use. Included for comparison with Figure 3.16

Multi-strain continuous-time priority queue SEIR model

Input Graph G with mean μ, Parameters $\{\lambda_i\}$, $\{\Gamma_i\}$, $\{\epsilon_i\}$
Procedure:
    time = 0
    $\{D_i\}$ = exponential-distribution(λ_i)
    for all v in G **and** j in 1 to i
        state(v) = S
    Infections = empty priority-queue
index-cases = choose i random elements from G, each with degree μ
    for j in 1 to i
        add [time, index-cases[j], I, j] **to** Infections
        G, Infections = infect-vertex(G, Infections)
    while Infections not empty
        [event-time, u, state, strain] = first(Infections)
        time = event-time
        if state = I
            G, Infections = infect-vertex(G, Infections)
        else
            state(u, j) = R
            Infections = rest(Infections)

function *infect-vertex* (G, Infections)
    [time, v, state, σ] = first(Infections)
    if state(v, σ) $\neq S$
        return [G, Infections]
    else
        state(v, σ) = I
        add [time + Γ_σ, v, R, σ] **to** Infections
        for all u in neighbours(v)
            if state(u, σ) = S
                infection-time = next-real(D_σ)
                if infection-time + $\epsilon_\sigma \leq \Gamma_\sigma$
                    if infection-time + time + ϵ_σ < next-infection-time(u, σ)
                        next-infection-time(u, σ) = infection-time + time + ϵ_σ
                        add [infection-time + time + ϵ_σ, u, I, σ] **to** Infections
    return [G, Infections]

Figure 3.18: Pseudocode rendition of the continuous-time priority-queue implmentation of a multi-strain SEIR model used throughout this thesis. The adjustments required to allow interactions and recombination between strains is not shown. The optimisation of infection events is shown.

Discrete-time priority queue model (mine)

```
Input Graph G, Parameters T, Γ
Procedure:
    time = 0
    D = truncated-geometric-distribution(T)
    for all v in G
        state(v) = S
    Infections = empty priority-queue
index-case = choose random element from G
    add [time, index-cases, I] to Infections
        G, Infections = infect-vertex(G, Infections)
    while Infections not empty
        [event-time, u, state] = first(Infections)
        time = event-time
        if state = I
            G, Infections = infect-vertex(G, Infections)
        else
            state(u) = R
            Infections = rest(Infections)

function infect-vertex (G, Infections)
    [time, v, state] = first(Infections)
    if state(v) ≠ S
        return [G, Infections]
    else
        state(v) = I
        add [time + Γ, v, R] to Infections
        for all u in neighbours(v)
            if state(u) = S
                infection-time = next-real(D)
                if infection-time ≤ Γ
                    if infection-time + time < next-infection-time(u)
                        next-infection-time(u) = infection-time + time
                        add [infection-time + time, u, I] to Infections
    return [G, Infections]
```

Figure 3.19: Pseudocode rendition of a single-strain discrete-time SIR version of the continuous-time priority-queue model used in this thesis. Included for comparison with Figure 3.20

Discete-time priority queue model (EpiFire)

```
Input Graph G, Parameters T, Γ
Procedure
    time = 0
    D = truncated-geometric-distribution(T, Γ + 1)
    for all v in G
        state(v) = 0
    Infections = empty priority-queue
    index-case = choose random element from G
    add [time, index-case] to Infections
    G, Infections = infect-vertex(G, Infections)
    while Infections not empty
        [event-time, u] = first(Infections)
        if event-time ≤ time
            G, Infections = infect-vertex(G, Infections)
            Infections = rest(Infections)
            time = event-time
        else
        for all u in G
            increment state(u)
            if state(u) = Γ
                state(u) = -1

function infect-vertex (G, Infections)
    [time, v] = first(Infections)
    if state(v) ≠ 0
        return [G, Infections]
    else
        state(v) = 1
        for all u in neighbours(v)
            if state(u) = 0
                infection-time = next-integer(D)
                if infection-time ≤ G
                    add [infection-time + time, u] to Infections
    return [G, Infections]
```

Figure 3.20: Pseudocode rendition of the BinomialChain algorithm implemented in EpiFire. Included for comparison with Figure 3.19

Chapter 4

Two strains spreading independently

4.1 Introduction

The purpose of this chapter is to understand how host population structure affects the rate of coinfection when there are two strains circulating simultaneously and *independently*, as if the two strains were circulating in separate but identical graphs. I want to explore the behaviour of the overlap between the outbreaks in the absence of any interaction between strains. In later chapters I will be looking at different types of interactions between strains within the host, and the results of this chapter form a null case that the dynamics of these interactions can be compared to.

By *overlap* I mean those vertices that become infected with both strains, and where the infections overlap in time. That is, if a vertex becomes infected with strain 1 and subsequently with strain 2 before it recovers from infection with strain 1, it is counted as part of the overlap. Of course, the overlap can be defined more broadly as all vertices that become infected with both strains, whether they have both infections simultaneously or not. To distinguish the two, I use the term *intersection* to refer to the latter, broader definition.

Notation

Here it is useful to define the overlap function $\omega_{i,j}(t)$, which analogously with $I_i(t)$ gives the number of overlapping infections at time t. As with $I_i(t)$ I will usually present it as a prevalence curve (time series) of the infection. I also calculate the *total overlap* $\Omega_{i,j}$ of strains 1 and 2:

$$\Omega_{i,j} = \int_0^\infty \omega_{i,j}(t)dt \tag{4.1}$$

I will also have occasion to measure the intersection, for which we define the total intersection $\Upsilon_{i,j}$ analogously to $\Omega_{i,j}$.

In addition to $\omega_{i,j}$ and $\Omega_{i,j}$, I will measure $\langle k_O \rangle$ and $\langle k_\Omega \rangle$, the mean degrees of the vertices in O_i and $\Omega_{i,j}$ respectively, to determine which parts of the graph are infected by either of both outbreaks.

Throughout this chapter, whenever I compare Ω or Υ to O_i, I will usually only compare it to O_l, the larger of O_1 and O_2 in each simulations. One the occasions when the smaller outbreak size is also given, this will be referred to as O_s. Table 4.1 summarises all the symbols used in this chapter. As in the previous chapter, wherever they are not needed for disambiguatio, subscripts indicating the strain in question are omitted in order to make the text more readable.

Symbol	Meaning
λ_i	The rate parameter of the exponential distribution determining the time to infection.
Γ	The duration of the infected phase Y. $\Gamma = 1t.u$ throughout
ϵ	The duration of the incubation period E. $\epsilon = 0.2t.u.$ throughout.
N	The size of the host population
T_i	The transmissibility of strain i, i.e. the probability that v successfully transmits i to u with before v recovers
O_i, $i \in \{1, 2\}$	The number of vertices that have been infected with strain i by the end of the simulation.
O_l, O_s	The larger (respectively smller) of O_1 and O_2
I_i, I_l	Shortened form of $I_i(t)$, the number of vertices infected with strain i at time t. I_l refers to the strain with the larger O_i
Ω	The number of vertices that by the end of the simulation were at some point in time infected with both strains
ω	Shortened form of $\omega(t)$. The number of vertices that at time t were infected with both strains.
Υ	The number of vertices that by the end of the simulation had been infected with both strains, not necessarily at one time.
$\langle k \rangle$	The mean degree of the whole population graph
$\langle k_O \rangle$	The mean degree of the larger outbreak (k_O is short for k_{O_l}) when the simulation has ended
$\langle k_\Omega \rangle$	The mean degree of the total overlap when the simulation has ended.
k_ω, k_I	Corresponding quantities to $\langle \Omega \rangle$ and $\langle k_o \rangle$, but for vertices in ω and I_l respectively.
τ	The delay after strain 1 is introduced to the host population before strain 2 is introduced.
$\langle\, \rangle$	Used around any symbol, except those pertaining to mean degree, to indicate its average over all simulations with the same parameter values.

Table 4.1: Symbols used in this chapter

Simulations run

This chapter concerns four sets of simulations, which I will discuss in separate sections. All simulations are carried out in both Poisson and Zipf graphs with size $N = 10000$, and with two strains circulating. In the first section, both strains are equally infectious, and are introduced to the population simultaneously. This serves as the basic model, and I examine the effects of varying graph structure and T in some detail. Simulations were run on graphs with Zipf distributed degrees and graphs with Poisson distributed degrees. Graphs with three different $\langle k \rangle$ – 5, 10 and 20 – were used. T ranged from 0.125 to 0.875 in increments of 0.125. These values were chosen because when the transmissibility is very close to 0 the majority of outbreaks fail to spread, and when it is very close to 1 the outbreaks always infect $> 99\%$ of the giant component. For each combination of $\langle k \rangle$, degree distribution type and T, 100 simulations were conducted. In general, throughout this chapter all combinations of parameters explored were simulated 100 times.

In the second section the two strains have equal transmissibility but only one strain is introduced at time 0, and the second after a delay τ. I consider different $\langle k \rangle$, graph type and τ, but for this set of simulations there is no variation in T.

In the third section, the two strains have different T and are introduced to the population simultaneously. I consider both a low and a high degree graph for each graph type, and several different values of T.

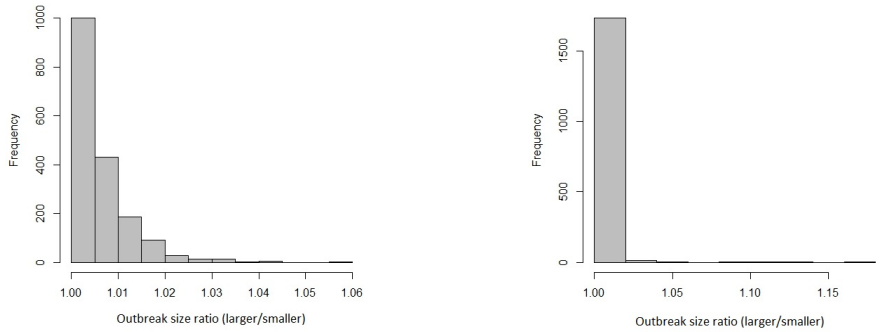
In the fourth section both T and τ are varied, but for only one value of $\langle k \rangle$ for each graph type.

4.2 Simultaneous introduction of strains with equal transmissibility

Figure 4.1 shows the ratio of the larger to the smaller outbreaks size, $\frac{O_l}{O_s}$. The vast majority differ by a relatively small amount. 121 simulations are not shown in figure 4.1. These are cases where only one of the strains successfully spread in the population, or where the smaller outbreak caused fewer than 50 infections. Despite the steps outlined in the previous chapter to prevent unsuccessful outbreaks, these cases make up 1.65% of all outbreaks. All these cases occurred in

simulations with $\langle k \rangle = 5$ or $T \leq 0.25$, or both. The size of the overlap (Ω) between smaller and larger outbreak in these simulations was in every case 0 or 1. These simulations have been removed from all subsequent analyses. This step will also be taken for the other three experiments outlined in this chapter, unless otherwise noted.

Although in a small fraction of the remaining simulations $\frac{O_l}{O_s} \leq 1.15$ in the Poisson case and 1.05 in the Zipf case, in all but 24 simulations in the Poisson graphs and 63 in the Zipf graphs $\frac{O_l}{O_s} < 1.02$.



Zipf Poisson

Figure 4.1: $\frac{O_l}{O_s}$ for Zipf (left) and Poisson graphs (right). Results fom 121 simulations are not shown. These cases have ratios between 86.23 and 9751. In all but 32 of the simulations not shown (16 in each graph type) $O_s = 1$, and the largest outbreak among them reached 46 vertices.

Overlap size

The first thing I am interested in is how Ω depends on the graph properties and epidemic parameters. Figure 4.2 shows the ratio $\frac{\langle \Omega \rangle}{\langle O_l \rangle}$ as a function of T. Increasing either $\langle k \rangle$ or T increases the ratio. For low $\langle k \rangle$ or low T, the ratio is much greater in Zipf graphs, despite the fact that the outbreaks themselves are always larger in the Poisson graphs when comparing outbreaks with equal T on graphs with equal $\langle k \rangle$ (figure 4.3). When both T and $\langle k \rangle$ are increased the ratio is greater in Poisson graphs.

Figure 4.4 shows the same points as figure 4.2 plotted as a function of $\langle O_l \rangle$. When T and $\langle k \rangle$ are high, all the points from Poisson distributed graphs cluster on the extreme right of the graph, because each outbreak is affecting nearly the entire population. Figure 4.4 shows that the overlap between two outbreaks of size O_i and O_j will be larger in a Zipf distributed graph than in a Poisson distributed graph, for all but the largest outbreaks.

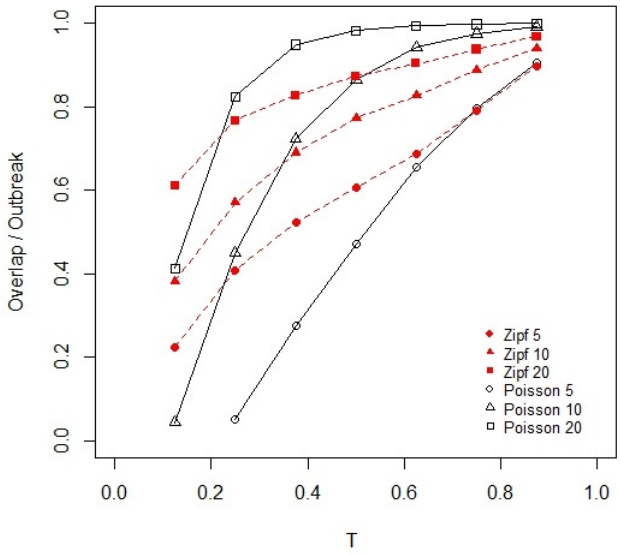


Figure 4.2: $\frac{\langle \Omega \rangle}{\langle O_l \rangle}$, as a function of T.

Comparing overlap and intersection size

Different graph types appear to produce overlaps of different size, both in absolute terms and relative to the size of the individual outbreaks. I would also like to know whether the overlaps in both graph types differ from the well-mixing case. That is, are the overlaps in Zipf graphs larger than the well-mixing population, or are the Poisson overlaps smaller? Or are both larger or both smaller?

I do not simulate well-mixing populations, and calculating $\langle \Omega \rangle$ even in the well-mixing case is not trivial, since the times of infections must be accounted for. However, the expected intersection

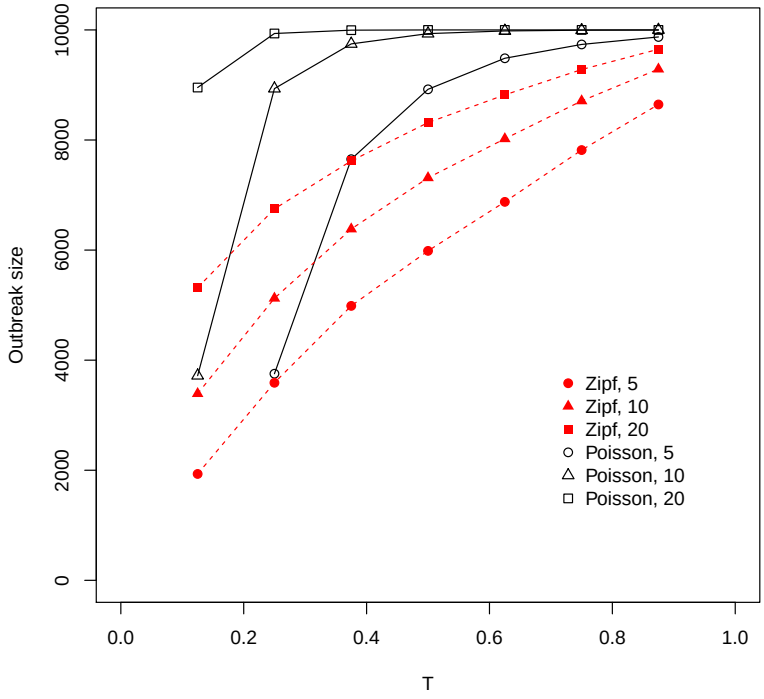


Figure 4.3: $\langle O \rangle$ as a function of T for different graph types and $\langle k \rangle$.

in the well-mixing case $\langle \Upsilon^M \rangle$ can be calculated. Since individuals in a well-mixing population are not linked, i.e. there is no correlation of the states of pairs of neighbours we can calculate $\langle \Upsilon^M \rangle$ in a population of size N by randomly drawing without replacement a set X of O_i integers and a set of Y of O_j integers from two separate sets $\{0...N\} \in \mathbb{Z}$ and calculating $S = X \cap Y$. If we repeat this process many times and take the average value of S,

$$\langle S \rangle \approx \langle \Upsilon^M \rangle \tag{4.2}$$

Figure 4.5 shows $\langle \Upsilon \rangle$ and $\langle \Omega \rangle$ from the same simulations as figure 4.4 for Zipf distributed graphs, along with $\langle \Upsilon^M \rangle$. Figure 4.6 shows the corresponding plot for Poisson distributed graphs. In Zipf distributed graphs we find that

$$\langle \Upsilon^M \rangle < \langle \Omega^{Zipf} \rangle < \langle \Upsilon^{Zipf} \rangle \tag{4.3}$$

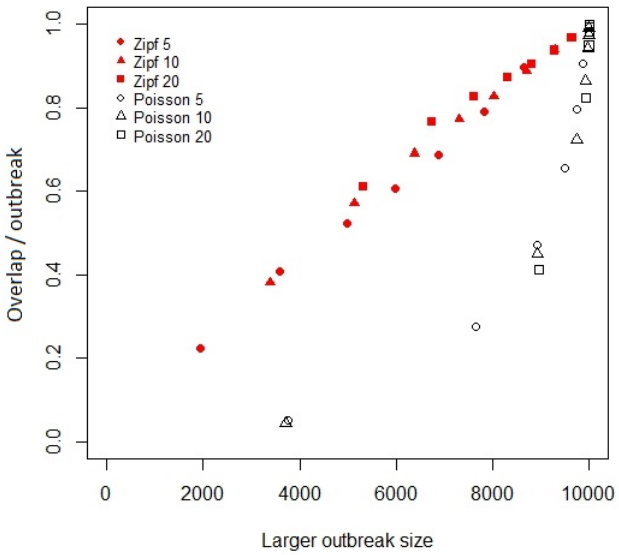


Figure 4.4: $\frac{\langle\Omega\rangle}{\langle O_l\rangle}$, as a function of $\langle O_l\rangle$. T rises going left to right.

Whereas, in Poisson distributed graphs

$$\langle\Omega^{Poisson}\rangle < \langle\Upsilon^{M}\rangle < \langle\Upsilon^{Poisson}\rangle \tag{4.4}$$

and when either T, $\langle k\rangle$ or both are low,

$$\langle\Omega^{Poisson}\rangle \ll \langle\Upsilon^{M}\rangle \tag{4.5}$$

In general, $\Omega \leq \Upsilon$, since if an infection contributes to the overlap, it contributes to the intersection by definition. Thus we can extend inequality 4.2:

$$\langle\Omega^{M}\rangle < \langle\Upsilon^{M}\rangle < \langle\Omega^{Zipf}\rangle < \langle\Upsilon^{Zipf}\rangle \tag{4.6}$$

So we see that in Zipf distributed graphs we get larger outbreaks and overlaps than in the well-mixing case. The same cannot be done for Poisson distributed graphs, and we cannot infer any

general relationship between Ω^M and $\Omega^{Poisson}$. However, for a population of size N

$$max(0, O_i + O_j - N) \leq \Upsilon \leq min(O_i, O_j) \tag{4.7}$$

In both Zipf and Poisson distributed graphs, for high T or $\langle k \rangle$, we see that O_i, $O_j \approx N$, and so $\frac{\Upsilon}{O_i} \approx 1$. From figure 4.6 we see that $\frac{\Omega}{O_i} \approx 1$, so that we can at least say that for sufficiently large outbreaks,

$$\Omega^{Poisson} \approx \Omega^M \tag{4.8}$$

. This implies that when T or $\langle k \rangle$ are low in Poisson distributed graphs, the majority of vertices that become infected with both strains recover from the first infection before contracting the second. As the outbreaks become larger, they also overlap more in time, so that vertices will tend to be infected with both for some period of time.

Degree distribution of the outbreaks

The larger outbreaks with smaller overlaps in the Poisson graph are probably due to the differences in degree distribution between the two graphs. In the Zipf graphs, most vertices have very few neighbours, and so have fewer opportunities to become infected, while a small portion of the population have a very large number of neighbours and are likely to be infected. In the Poisson graphs, all vertices have a similar number of neighbours, and so each vertex has a simiar chance of being infected. This means that in the Zipf distributed graphs both outbreaks are likely to spread preferentially in the same high-degree subset of vertices but fail to reach a large fraction of the lowest-degree vertices, leading to larger Υ and Ω but smaller O_1, O_2, than in the well-mixing case. In the Poisson distributed graphs the intersections are only marginally larger than in the well-mixing case.

Figure 4.7 shows $\langle k_O \rangle$ and $\langle k_\Omega \rangle$, for varying T, $\langle k \rangle = 20$. In the Zipf distributed graphs,

$$\langle k_\Omega \rangle > \langle k_O \rangle > \langle k \rangle \tag{4.9}$$

confirming that in Zipf graphs much of the coinfection is taking place in the highest degree vertices, and that the overlap tends to cluster in the higher-degree vertices of the overlap. In Poisson distributed graphs,

$$\langle k_\Omega \rangle \approx \langle k_O \rangle \approx \langle k \rangle \tag{4.10}$$

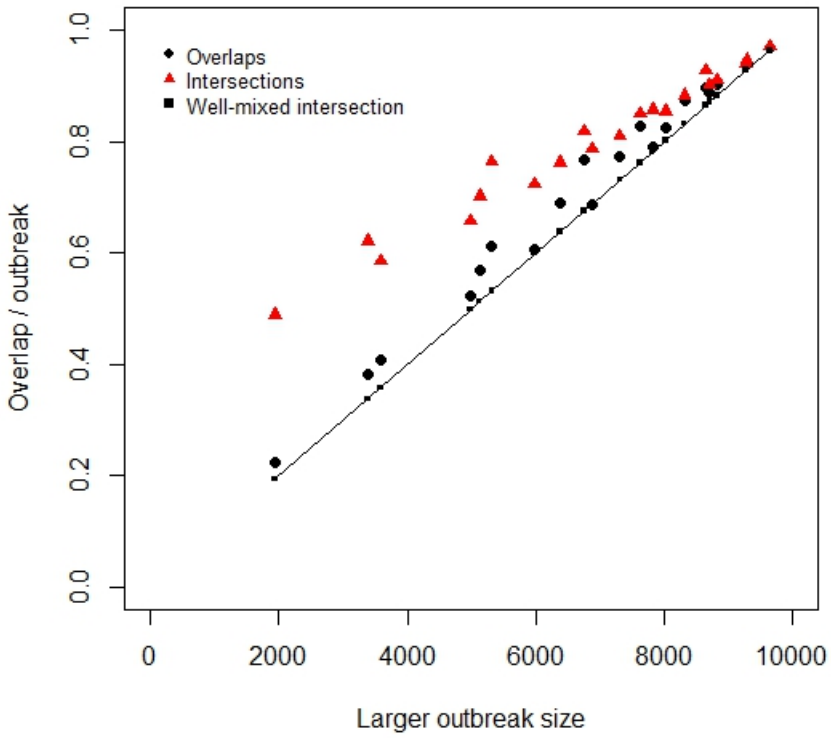


Figure 4.5: $\langle \frac{\Upsilon}{O_l} \rangle$ (red triangles) and $\langle \frac{\Omega}{O_l} \rangle$ (black circles), as a function of O_l, in Zipf distributed graphs. The black squares and line show $\frac{\langle \Upsilon^M \rangle}{O_l^M}$

This must be the case when each infection reaches a significant fraction of the population, given the low variability in the degree of vertices.

If high-degree vertices are infected early in an outbreak, and they are also more likely to be part of the overlap, it is reasonable to expect that they are coinfected early. But it is conceivable that they are not. Depending on the assortativity it is possible that although high degree vertices are always reached they are not reached early. If the graphs are assortative, the overlap could begin in a large cluster of low-degree vertices and only spread to the clustered higher-degree vertices once these are infected. The graph generating algorithm I use does not allow me to specify assortativity, so I

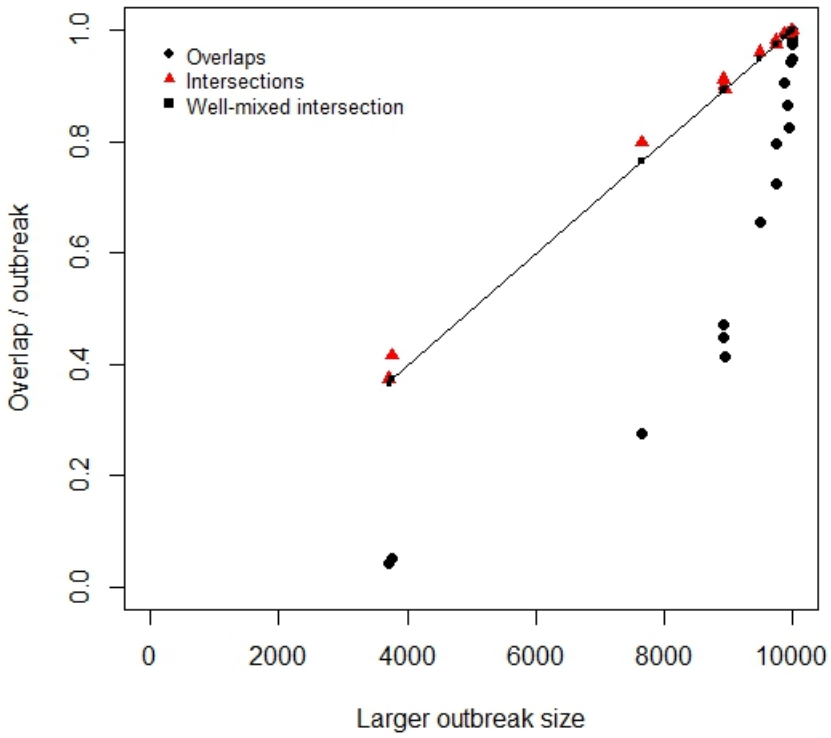


Figure 4.6: $\langle \frac{\Upsilon}{O_l} \rangle$ (red triangles) and $\langle \frac{\Omega}{O_l} \rangle$ (black circles), as a function of O_l, in Poisson distributed graphs. The black squares and line show $\langle \frac{\Upsilon^M}{O_l^M} \rangle$

cannot explore its effect in depth. However, all the graphs used here are quite strongly dissasortative, especially the Zipf distributed grahps, with an assortativity between -0.072 ($\langle k \rangle = 20$) and -0.179 ($\langle k \rangle = 5$), so we should expect the highest-degree vertices to be among the first to be coinfected.

Figure 4.8 shows how $\langle k_I \rangle$ and $\langle k_\omega \rangle$ develop over the course of the outbreak in the Zipf distributed graphs, varying T Each point shows the mean degree of the vertices infected between that point and the previous point, so that for example the point at 2000 vertices shows the mean degree of the 1500th to the 2000th vertex to become infected. Figure 4.9 shows the same for the Poisson distributed graphs. Figure 4.10 shows the corresponding figure for Zipf distributed graphs, varying

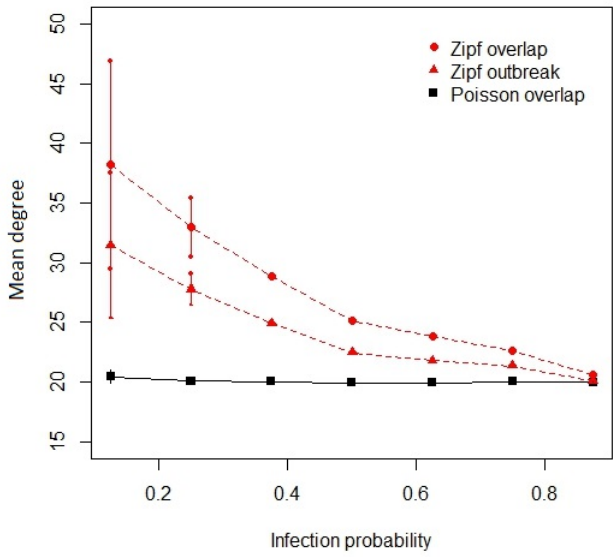


Figure 4.7: $\langle k_O \rangle$ and $\langle k_\Omega \rangle$, varying T. $\langle k \rangle = 20$. The error bars are $\pm 1 s.d.$. Note the overlapping error bars for Zipf $\langle k_\Omega \rangle$ and $\langle k_O \rangle$ when $T = 0.125$. $\langle k_O \rangle$ in Poisson graphs were omitted, as in no case could they be visibly distingushed from $\langle k_\Omega \rangle$, or indeed from $\langle k \rangle$.

$\langle k \rangle$. In the Zipf distributed graphs, $\langle k_I \rangle$ (black solid lines) increases a little from the first 100 vertices to the next 750, suggesting that the highest degree vertices are not always part of the first blush of the outbreak, irrespective of T or $\langle k \rangle$. In contrast, $\langle k_\omega \rangle$ is monotonically decreasing, showing that high-degree vertices *are* among the first to become coinfected.

Although O_i, Ω $\langle k_O \rangle$ and $\langle k_\Omega \rangle$ all depend strongly on T, figure 4.8 shows that T has only a relatively small effect on $\langle k_\omega \rangle$ and $\langle k_I \rangle$. The large effect on the overall mean is due to the cumulated small differences at each stage.

Timing of the overlap

I would also like to know when peak overlap occurs relative to peak outbreak. Given that high-degree vertices are coinfected early in the outbreaks, the overlap might peak before the outbreaks themselves do. On the other hand, after the high-degree vertices have been coinfected, much of

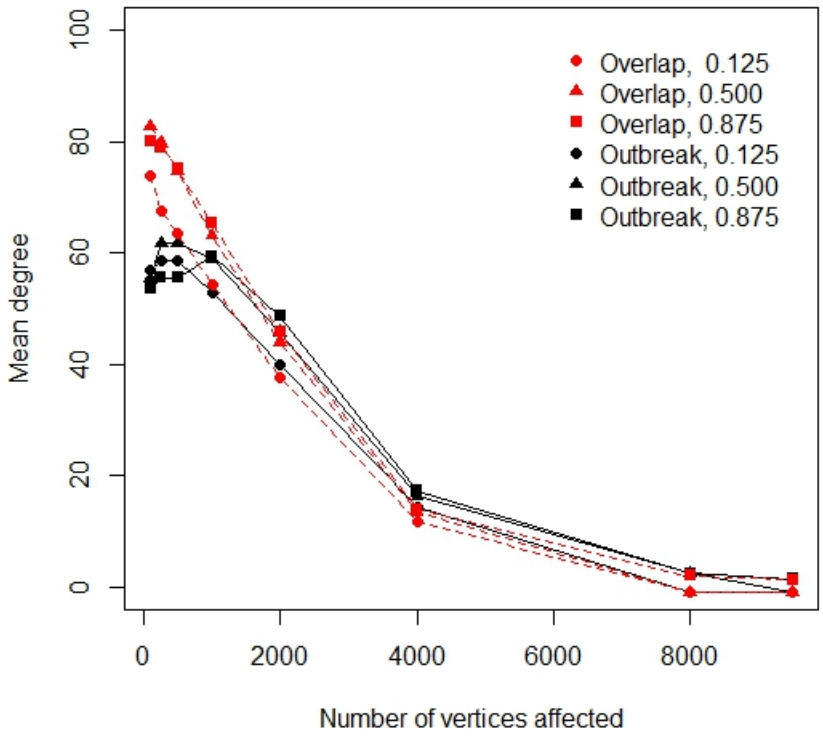


Figure 4.8: $\langle k_\omega \rangle$ (red, dotted lines) and $\langle k_I \rangle$ (solid black lines, i is the strain with larger outbreak) for three different values of T in a Zipf distributed graph with $\langle k \rangle = 20$. Note that the x-axis shows number of vertices affected, not time. Each point represents the mean degree of all vertices infected from the previous point. The points are at 100, 250, 500 750, 1000, 2000, 4000, and 8000 vertices.

the rest of the overlap might happen somewhat randomly and spread out over the duration of the outbreak, so that the overlap might still be growing. Figure 4.11 shows the mean peak times of both outbreaks and their overlap for Zipf distributed graphs, varying T and $\langle k \rangle$; figure 4.12 shows the same for Poisson distributed graphs. Increasing T and $\langle k \rangle$ both cause earlier peak outbreaks and overlaps. This is because they both have the effect of increasing the expected number of succesful transmissions from a vertex, which in turn means that the number of sequential infections required to reach lower-degree vertices decreases, leading to outbreaks growing faster.

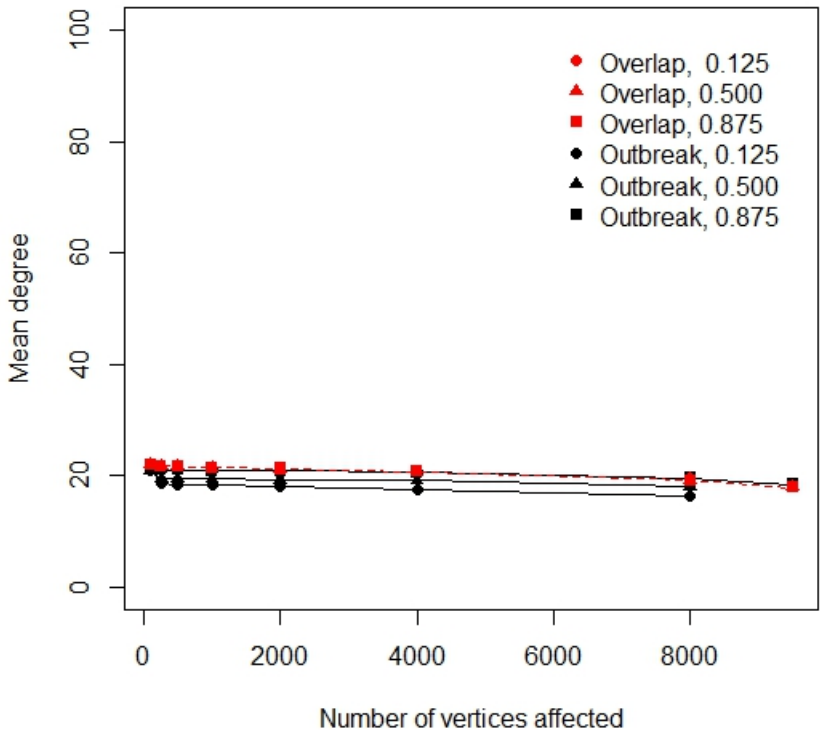


Figure 4.9: $\langle k_\omega \rangle$ (red, dotted lines) and $\langle k_I \rangle$ (solid black lines) for three different values of T in a Poisson distributed graph with $\langle k \rangle = 20$. Note that the x-axis shows number of vertices affected, not time. Each point represents the mean degree of all vertices infected from the previous point. The points are at 100, 250, 500 750, 1000, 2000, 4000, and 8000 vertices.

Outbreaks in Zipf graphs are significantly faster than the corresponding outbreaks in Poisson graphs, except for very high $\langle k \rangle$ or T, where the relationship is reversed. This is because changing $\langle k \rangle$ or T has a much larger effect on the timing of the peak outbreak in Poisson distributed graphs. In both graphs, the mean peak overlap time is very close to the mean peak outbreak time, suggesting that the vertices that are part of the overlap are among those infected early. In Zipf distributed graphs the overlap peaks consistently earlier than the larger outbreak except when $T = 0.125$. However, the difference is less than $1s.d$ of mean peak outbreak time. These results are in good

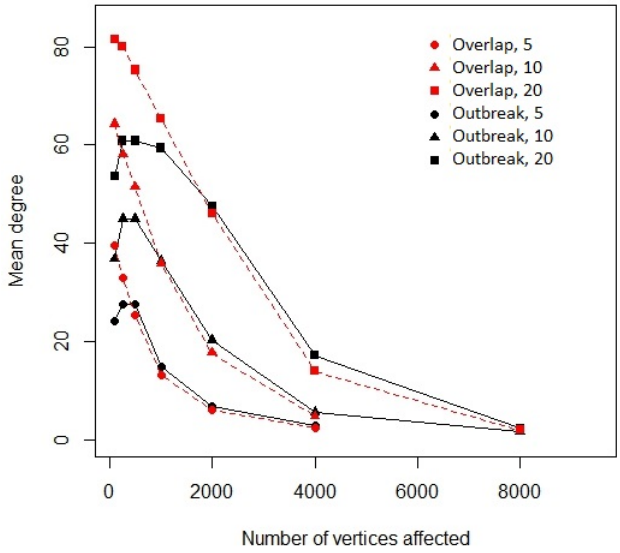


Figure 4.10: $\langle k_\omega \rangle$ (red, dotted lines) and $\langle k_I \rangle$ (solid black lines) for three different values of $\langle k \rangle$ (values shown in figure legend) in Zipf distributed graphs. $T = 0.5$. Note that the x-axis shows number of vertices affected, not time. Each point represents the mean degree of all vertices infected from the previous point. The points are at 100, 250, 500 750, 1000, 2000, 4000, and 8000 vertices.

agreement with the findings that overlaps in Zipf graphs are more clustered in high-degree vertices in Zipf distributed graphs.

Consider the lag between the first and second infection in a vertex that becomes infected with both strains. If the lag is greater than Γ, the vertex is part of the intersection but not the overlap. If the lag is less than Γ, the vertex is part of both. In all simulations $\Gamma = 1$ so this is the threshold lag duration. Earlier I posited that the reason that in Poisson distributed graphs the overlap and intersection are of similar size when T or mean degree is high is that in this case the two infections reach the same vertex close to the same time.

To test this, I calculate the mean lag as a function of T and mean degree. Figure 4.13 shows the difference between mean lag in all coinfections that make up the intersection, and those that make

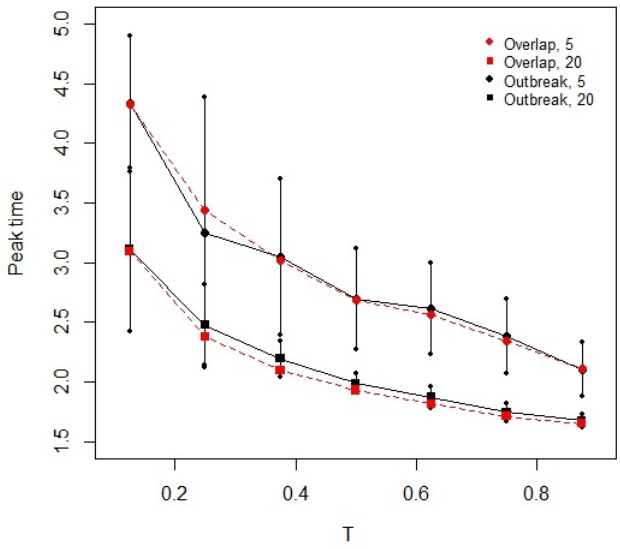


Figure 4.11: Mean peak times of outbreaks (black, solid lines) and overlaps (red, dotted lines) as a function of T. Zipf distributed graphs $\langle k \rangle = 5$ (circles) and 20 (squares).

up the overlap. In the Zipf distributed graphs the difference is in all cases small, indicating that the overlap makes up most of the intersection. In the Poisson distributed graphs, when T or mean degree are low, the difference is very large, indicating that a larger fraction of the intersection is not part of the overlap. When T is high, the difference is much lower, and is similar to the Zipf distributed case.

In Poisson distributed graphs, when T is high, infection spreads so quickly to most of the graph that almost all coinfected vertices will be infected with both strains at the same time. In Zipf distributed graphs, both outbreaks tend to reach high-degree vertices first, and then spread to vertices of progressively lower degree. This ensures that vertices become infected with both strains at similar times. In the absence of this effect, slow-spreading outbreaks on Poisson distributed graphs will often reach a vertex at very different times, so that the vertex has already recovered from infection with the first strain by the time the second arrives.

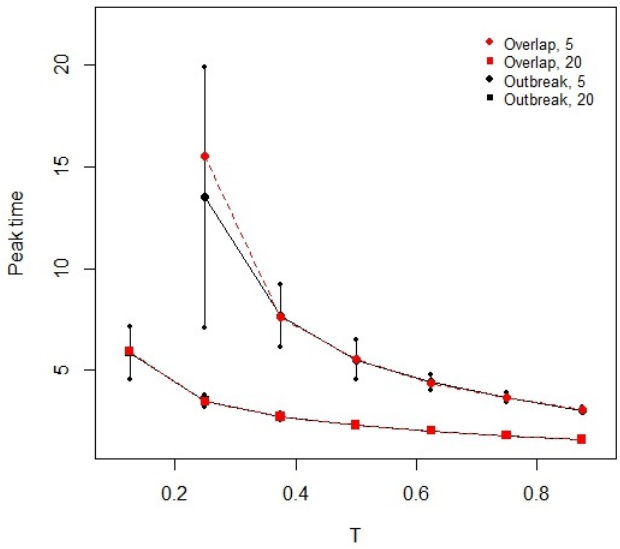


Figure 4.12: Mean peak times of outbreaks (black, solid lines) and overlaps (red, dotted lines) as a function of T. Poisson distributed graphs with $\langle k \rangle = 5$ (circles) and 20 (squares).

When comparing peak times between the two graph types, it is worth looking at timeseries of $\langle I_1 \rangle(t)$, $\langle I_2 \rangle$ and $\langle \omega \rangle$. Figure 4.14 shows timeseries for Zipf (bottom panel) and Poisson (top panel) for three different values of T, with fixed mean degree, and figure 4.15 for three different mean degrees and fixed T. When T or mean degree are low, outbreaks are faster in the Zipf distributed graphs, while for intermediate or high T or mean degree, the opposite is true, so that the speed of the outbreaks varies more with T and $\langle k \rangle$ in the Poisson distributed graph.

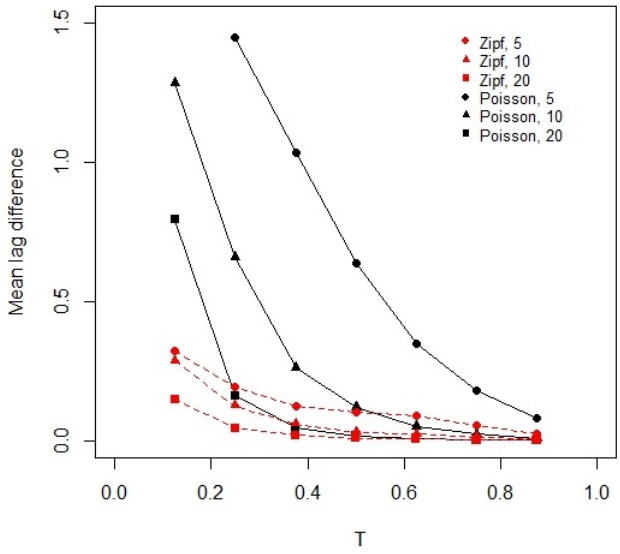


Figure 4.13: The difference between the mean inter-infection lag in the intersection and the overlap, as a function of T.

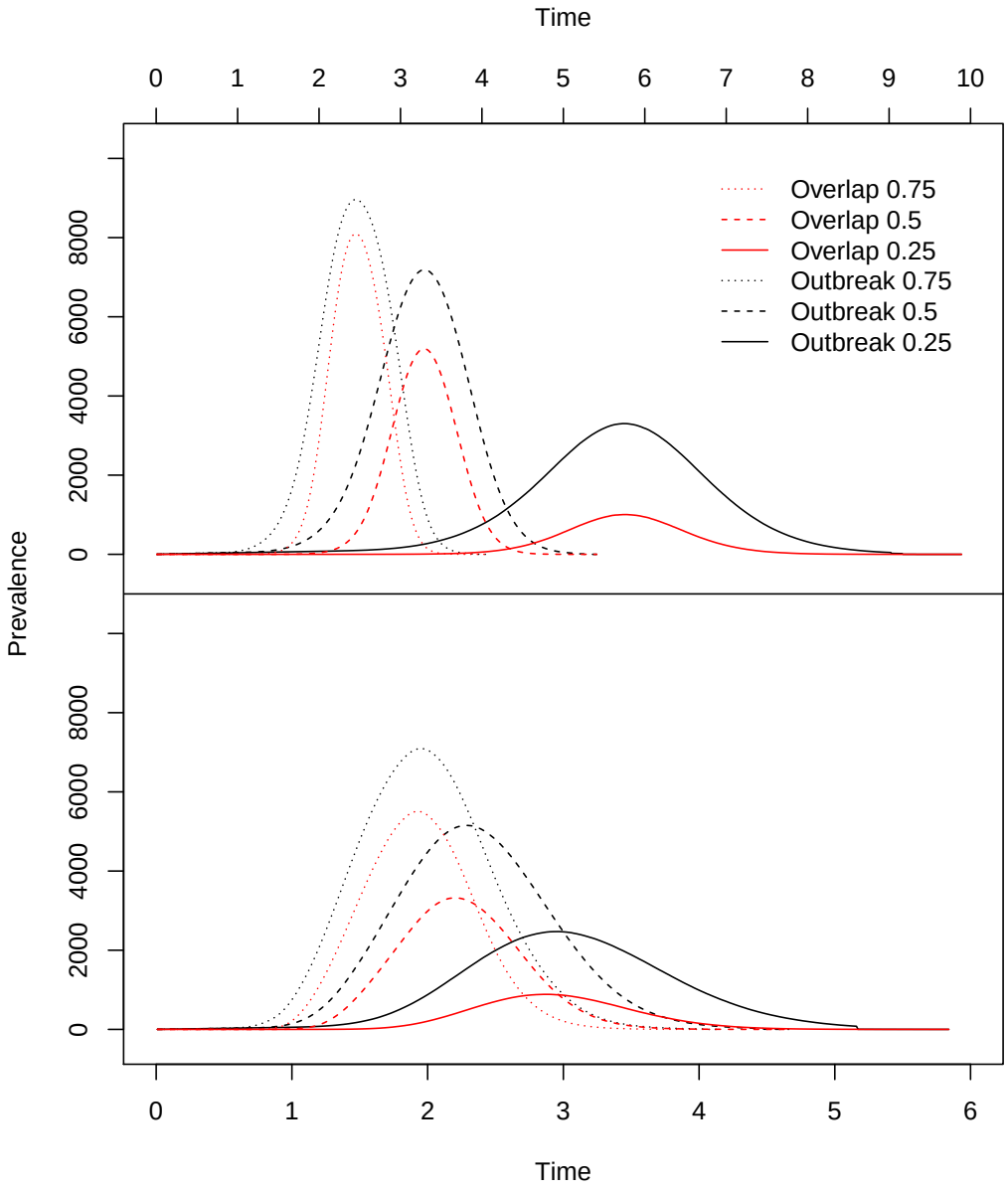


Figure 4.14: $\langle I_i \rangle(t)$ (black lines) and $\langle \omega \rangle(t)$ (red lines) for three different values of T in Zipf (bottom) and Poisson (top) distributed graphs, with $\langle k \rangle = 10$.

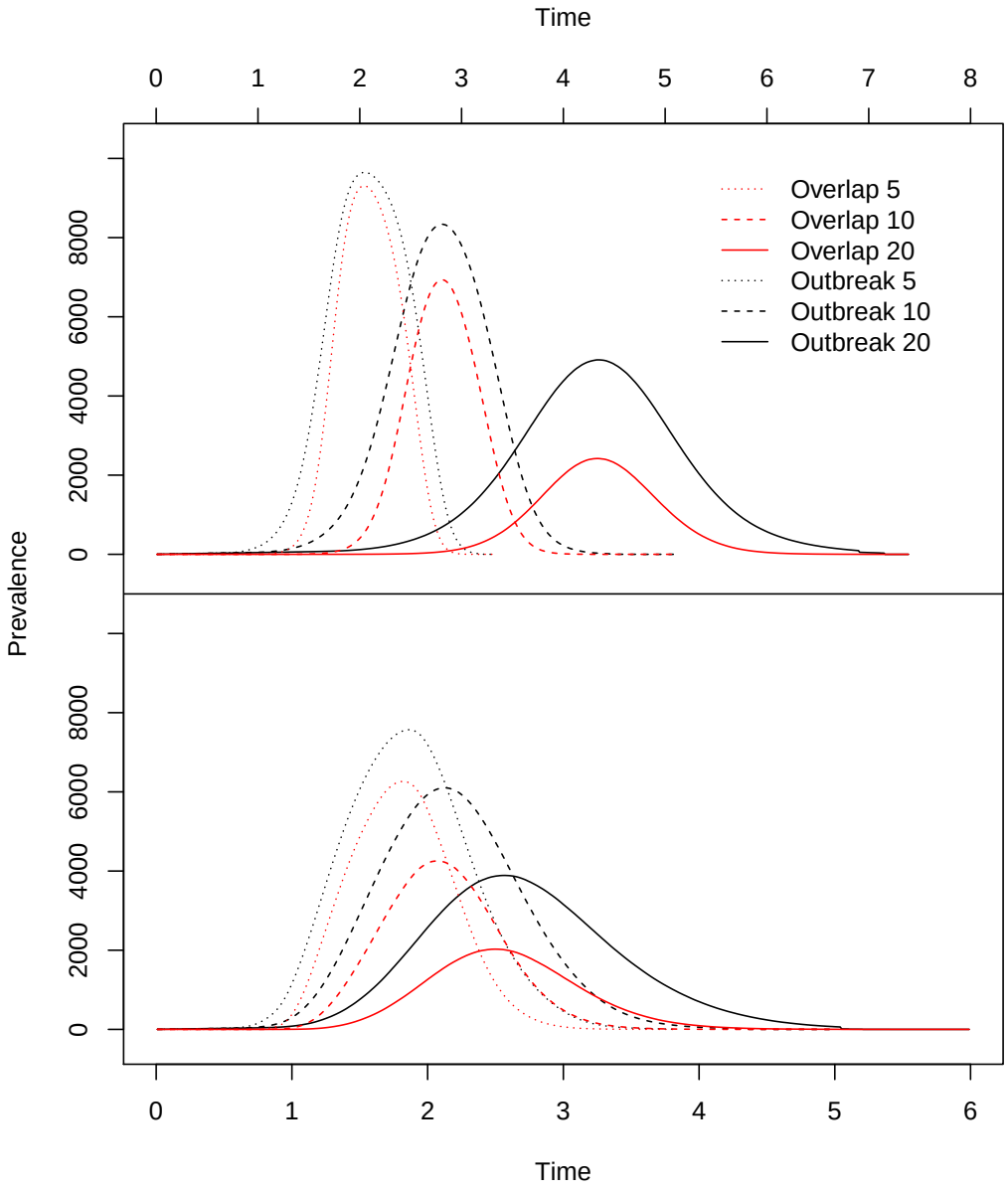


Figure 4.15: $\langle I_i \rangle(t)$ (black lines) and $\langle \omega \rangle(t)$ (red lines) for three different values of $\langle k \rangle$ in Zipf (bottom) and Poisson (top) distributed graphs, with $T = 0.5$.

4.3 Staggered introduction of strains with equal transmissibility

The fact that outbreaks in Zipf graphs with more heterogeneous degree distribution have larger overlaps than in the more homogeneous Poisson graph is interesting, but these results are based on simulations where both strains are introduced to the population simultaneously. This is not a very biologically realistic assumption unless the two strains are introduced by a single coinfected individual, which is not the case in these simulations. Given the very quick single growth-extinction wave of these outbreaks, it might be reasonable to expect that if I introduce even a very small delay between the two outbreaks, Ω will shrink very quickly. To test this, I conducted a set of simulations in which strain 1 is introduced at time $t = 0$, but strain 2 is only introduced after a delay τ, ranging from $0.1t.u$ to $2t.u$. All other parameter values are as in the previous set of simulations. Since the duration of infection is always $1t.u.$, $\tau \geq 1.t.u$ means that at least the first few vertices affected by the first outbreak have already recovered by the time the second outbreak occurs.

Figure 4.16 shows $\langle\Omega\rangle$ in both graph types as τ become progressively greater. In all cases, as τ increases, $\langle\Omega\rangle$ shrinks. When $\langle k\rangle = 20$, the overall reduction in $\langle\Omega\rangle$ is greater in Poisson distributed graphs. In graphs with $\langle k\rangle = 10$ the reduction in $\langle\Omega\rangle$ as τ increases is much smaller, and when $\langle k\rangle = 5$ the relationship between the graph types is reversed, with a smaller reduction in $\langle\Omega\rangle$ for Poisson distributed graphs. Although it is not clear from figure 4.16, the minimum values that $\langle\Omega\rangle$ takes is roughly 100 vertices.

To make the differences between to two graph types clearer, figure 4.17 shows $\langle\Omega\rangle$ for increasing τ as a fraction of $\langle\Omega\rangle$ when $\tau = 0$. It is clear that in graphs with higher $\langle k\rangle$, although the reduction in $\langle\Omega\rangle$ is greater in absolute terms in the Poisson distributed graphs, $\langle\Omega\rangle$ shrinks by nearly 99% in both graph types. However, the effect of increasing τ is much less pronounced in Poisson distributed graphs with lower $\langle k\rangle$. Even when $\tau = 2t.u$, $\langle\Omega\rangle$ is only slightly more than halved, whereas in the Zipf case, the reduction is by more than 80%.

Figure 4.18 shows how $\langle k_\Omega\rangle$ develops over the course of the simulation. The vertices that have been infected at any point during the outbreak form a connected induced subgraph of the whole population graph. Since in a Poisson distributed graph, for almost any induced subgraph $H \subset G$

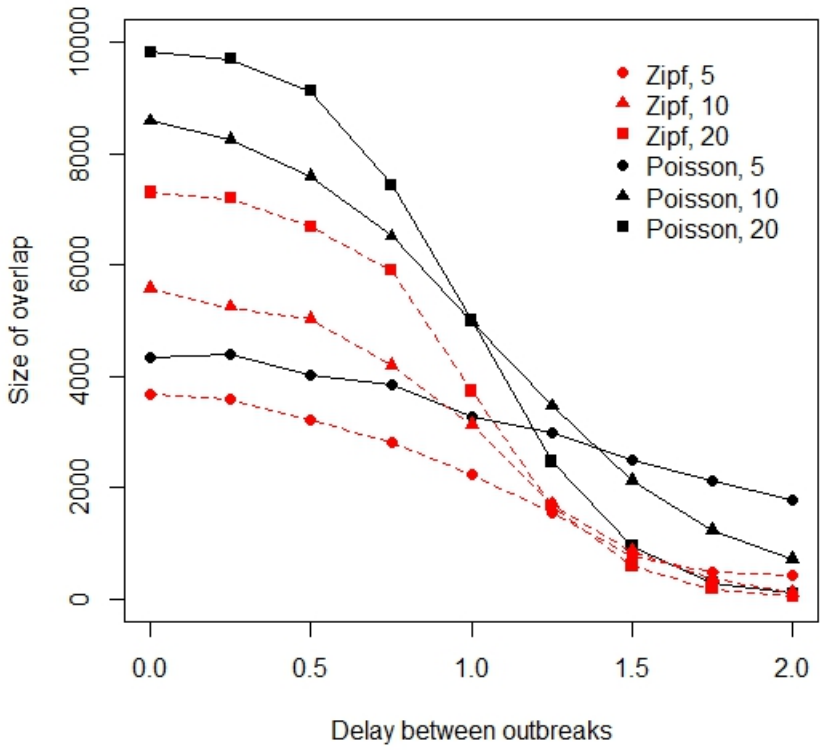


Figure 4.16: $\langle\Omega\rangle$ as a function of τ, for different $\langle k\rangle$ (values in figure legend). $T_1 = T_2 = 0.5$.

$\langle k_H\rangle \approx \langle k_G\rangle$, $\langle k_\Omega\rangle$ will not change much as τ increases. In the Zipf graphs, if $\tau > 1t.u$ the early behaviour of $\langle\Omega\rangle$ changes. When $\tau = 0$, $\langle\Omega\rangle$ decreases monotonically, but when $\tau > 0$, $\langle\Omega\rangle$ initially increases a little before shrinking towards $\langle k\rangle$. This is because the first strain to spread has reached a much larger fraction of the population, so that first vertices to be infected with the second strain are likely already to be infected with the first strain, and hence the early stages of the overlap and the outbreak of the second strain will be much more similar than when $\tau = 0$.

I would also like to know how delaying one outbreak affects the peak time of the overlap. Figure 4.19 shows how the peak overlap time depends on τ for $\langle k\rangle = 20$. Peak overlaps occur earlier in Zipf distributed graphs than in their Poisson distributed counterparts until $\tau = 1.5t.u$, after which

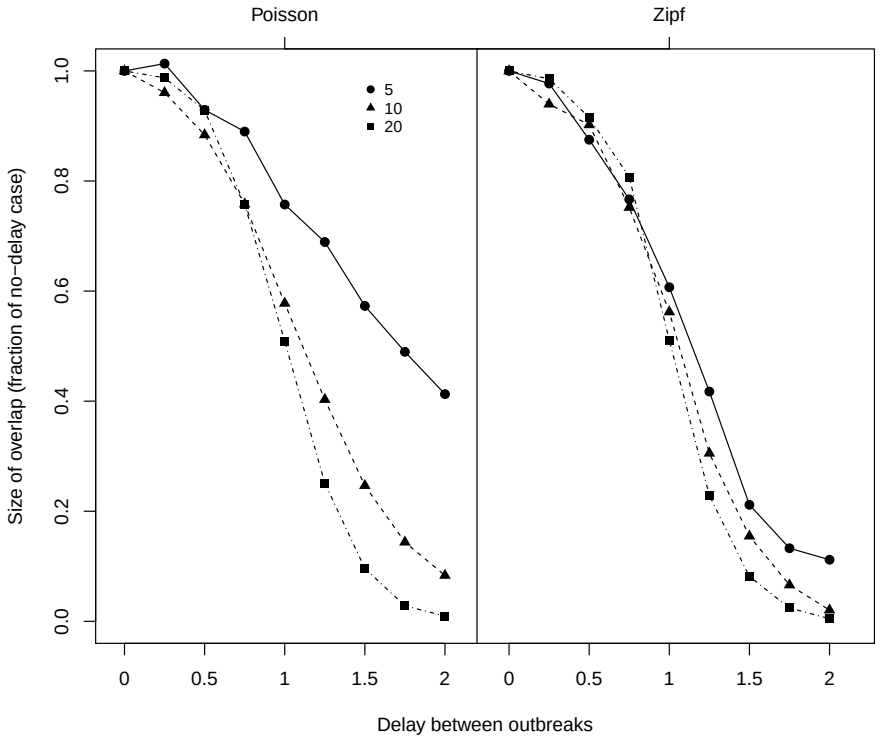


Figure 4.17: $\frac{\langle\Omega\rangle}{\langle\Omega_{\tau=0}\rangle}$ as a function of τ, for different $\langle k\rangle$ (values in figure legend). $T_1 = T_2 = 0.5$.

the overlap peaks earlier in the Poisson distributed graphs. In general, τ has a larger effect on the time of peak overlap in Zipf distriuted graphs ($\tau = 2.0t.u$ delays the peak overlap by $1t.u.$) than in Poisson distributed graphs ($\tau = 2.0t.u.$ delays the peak overlap by less than $0.75t.u.$).

Figure 4.20 shows shows the dependence of peak overlap time on τ when $\langle k\rangle = 5$. Once again peak overlap is postponed relatively more in the Zipf distributed graphs, but in this case outbreaks and overlaps in Poisson distributed graphs are so slow that even outbreaks delayed by $2t.u$ in the Zipf distributed graphs peak before outbreaks without delay in the Poisson distributed graphs.

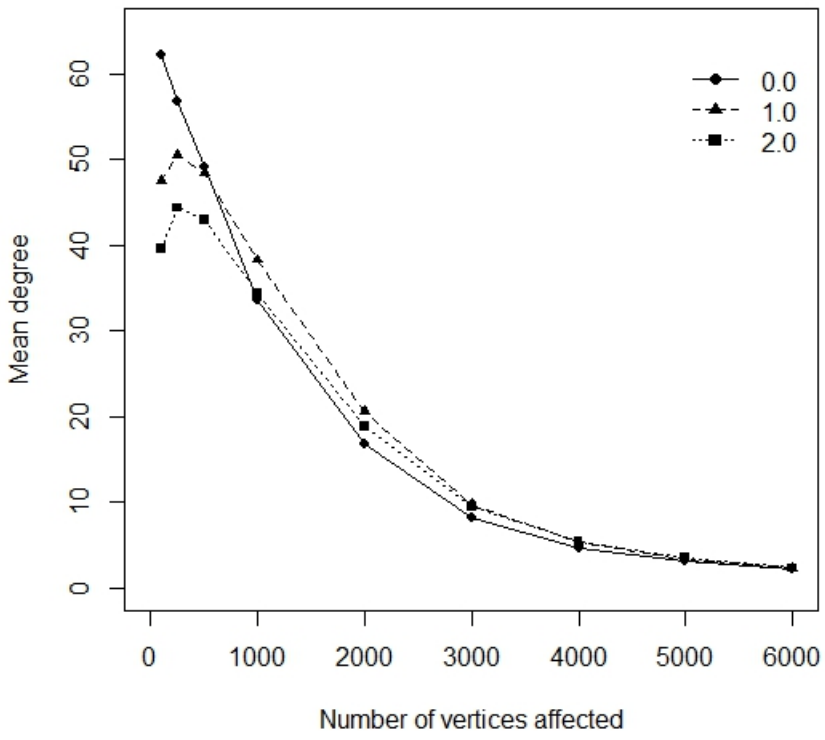


Figure 4.18: $\langle k_\omega \rangle$ for three different values of τ (given in figure legend) in a Zipf distributed graph with $\langle k \rangle = 20$. Note that the x-axis shows number of vertices affected, not time. Each point represents the mean degree of all vertices infected from the previous point. The points are at 100, 250, 500 750, 1000, 2000, 3000, 4000, 5000, and 6000 vertices. $T = 0.5$

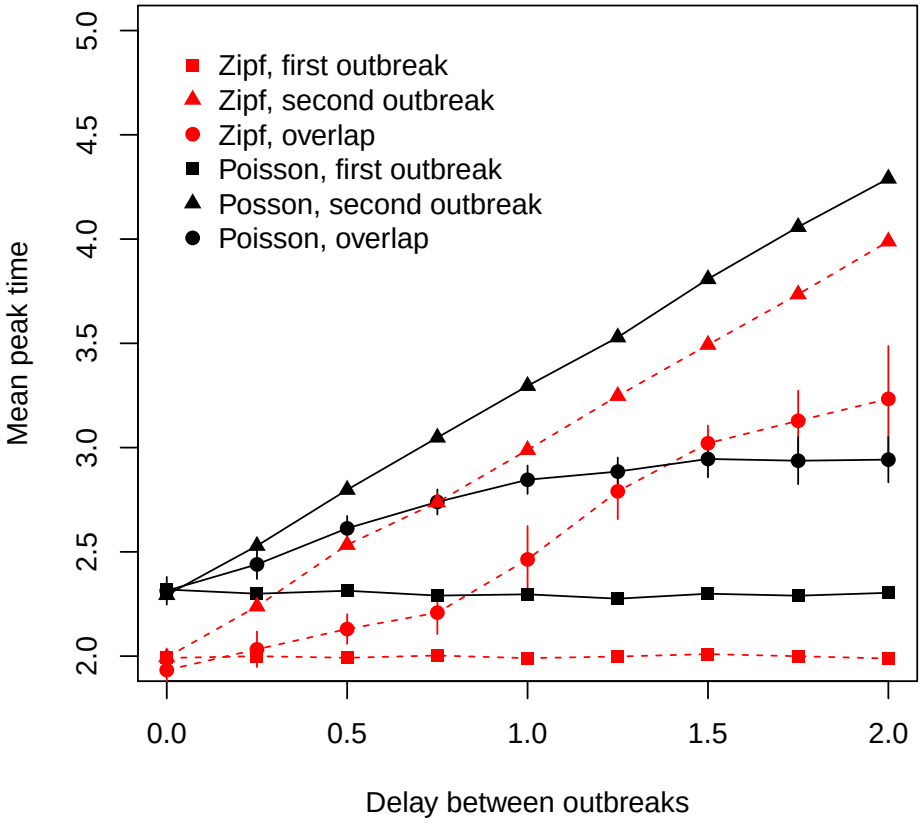


Figure 4.19: Mean peak times for outbreaks (squares, triangles) and overlaps (circles) in Zipf (red dotted lines) and Poisson (black solid lines) distributed graphs wihen $\langle k \rangle = 20$. $T = 0.5$ for both strains, and τ varies.

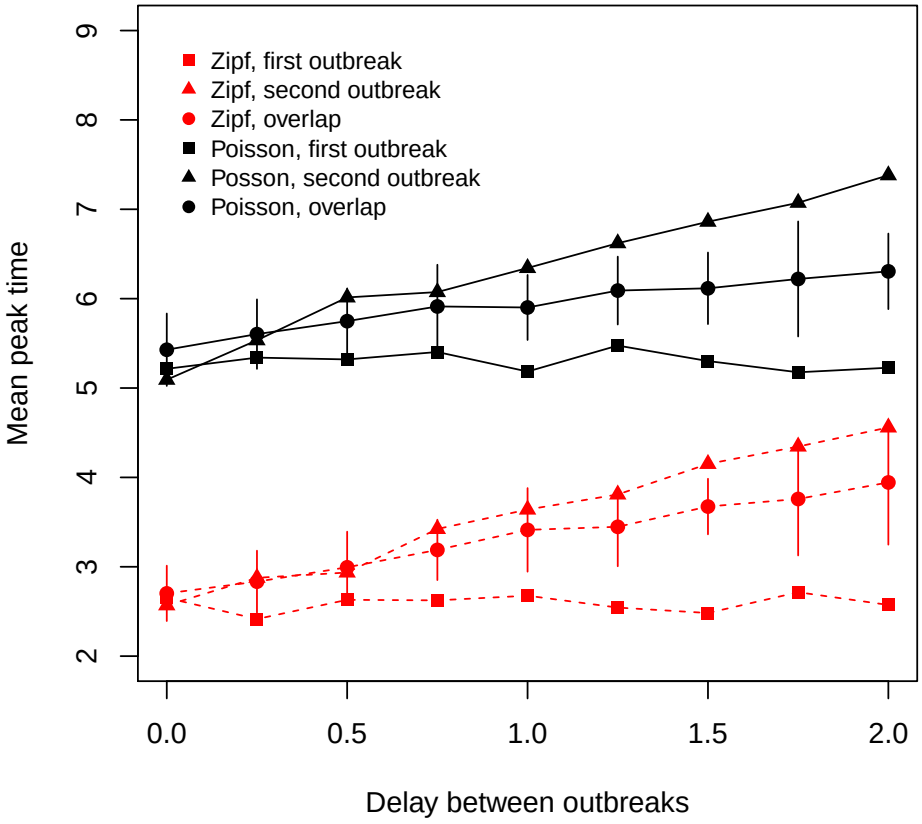


Figure 4.20: Mean peak times for outbreaks (squares, triangles) and overlaps (circles) in Zipf (red dotted lines) and Poisson (black solid lines) distributed graphs with $\langle k \rangle = 5$. $T = 0.5$ for both strains, and τ varies.

4.4 Simultaneous introduction of strains with unequal transmissibility

In all the simulations discussed so far, I assume that the two outbreaks have the same transmissibility. In reality this may often not be the case. In this section I take a brief look at two strains that are introduced to the population simultaneously but have different transmissibility. T_1 is the transmissibility of strain 1, and T_2 correspondingly for strain 2. T_1 takes three values: 0.25, 0.5 and 0.75. For each of these three values I run simulations with $T_2 = T_1 - 0.1$, -0.05, 0, 0.05 and 0.1.

Figure 4.21 shows how varying T_2 while keeping T_1 fixed changes $\langle\Omega\rangle$. In the Zipf distributed graphs the change in Ω is relatively small and roughly linear. In Poisson distributed graphs the effect on $\langle\Omega\rangle$ is much larger when T_1 is low, and smaller when it is very high. When T is low, the speed at which outbreaks reach peak size is more sensitive to small changes in T, and so small differences in T have large effects on $\langle\Omega\rangle$. When T are high the differences in the speed of the outbreak growth are very small and both outbreaks tend to reach almost the entire population, so that the variation in Ω is small.

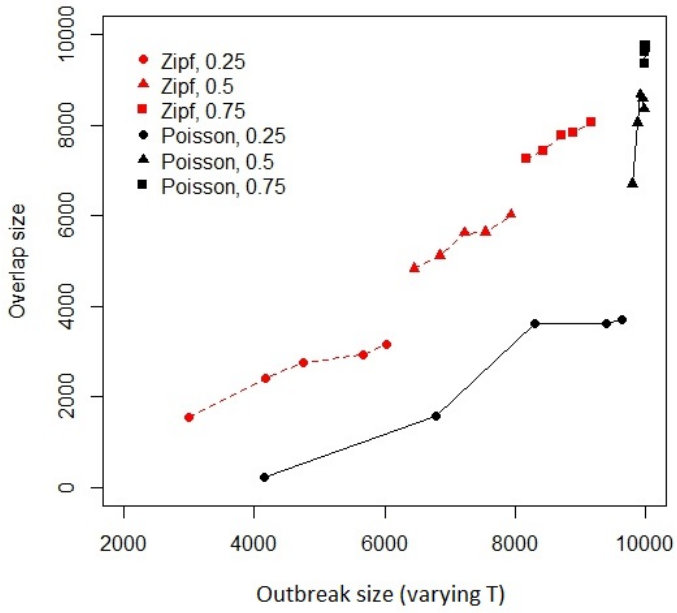


Figure 4.21: $\langle \Omega \rangle$ when $T_1 \neq T_2$, in Zipf (red dotted line) and Poisson (solid black line) distributed graphs with $\langle k \rangle = 10$. The values of T_1 are given in the legend. $T_1 - T_2$ goes from -0.1 to 0.1 in increments of 0.05 going left to right. O_1 for each set of simulations is the same as the x value for the middle point in the line from that set.

4.5 Staggered introduction of strains with unequal transmissibility

In this last section I combine the variations I explored in the two sections before, by allowing the two strains to have different transmissibility ($T_2 \neq T_1$) and to be introduced into the population at different times ($\tau > 0$). I only look at three different values of τ: 0.5, 1.0 and $1.5t.u.$. In one set of simulations $T_1 = 0.6$, $T_2 = 0.6$, 0.7 or 0.8, and $T_2 = 0.6$, 0.7 or 0.8, $T_2 = 0.6$. In the other set $T_1 = 0.3$, $T_2 = 0.3$, 0.4 or 0.5 and $T_1 = 0.3$, 0.4, or 0.5, $T_2 = 0.3$.

Figure 4.22 shows how $\langle\Omega\rangle$ depends on T_1, T_2 and τ. Each panel groups all simulations where $\langle k\rangle$ and the transmissibility of the "fixed" strain are the same. It is clear that T_1, T_2 and τ have a strong effect on $\langle\Omega\rangle$.

When both T_1 and T_2 are high, $\langle\Omega\rangle$ almost always shrinks as τ grows. When $\tau > 0$ and $T_1 < T_2$, the reduction in $\langle\Omega\rangle$ is smaller than when $T_1 > T_2$. This effect is present in graphs of both types, but is much more pronounced in the Poisson distributed graphs. When $T_1 > T_2$, $\langle\Omega^{Poisson}\rangle > \langle\Omega^{Zipf}\rangle$, but when $T_1 < T_2$, $\langle\Omega\rangle$ is very similar in both graph types.

When both transmissibilities are low but $T_1 > T_2$, $\langle\Omega\rangle$ decreases with increased τ. However, when $T_1 < T_2$, $\langle\Omega\rangle$ grows with increased τ in Poisson distributed graphs, but not in Zipf distributed graphs. This effect is more pronounced when $\langle k\rangle = 10$ than when $\langle k\rangle = 20$.

These results are consistent with the results from the previous two sections. Introducing a delay reduces $\langle\Omega\rangle$ because the first outbreak tends to reach peak size before the second, so that many vertices have recovered from the first infection before becoming infected with the second. Reducing T also increases the time to peak outbreak size, while increasing T reduces it. Increasing T_1 or decreasing T_2 therefore increases the time between the peak outbreak sizes of the two strains, and so reduces $\langle\Omega\rangle$. Reducing T_1 or increasing T_2 reduces the time between peak outbreak sizes, increasing $\langle\Omega\rangle$.

In the Poisson distributed graph with $\langle k\rangle = 10$ and low transmissibilities, when $\tau = 0$ the more infectious strain peaks much earlier than the less infectious strain. Increasing T_2 relative to T_1

makes strain 2 peak earlier, and therefore closer to the peak of strain 1, increasing $\langle\Omega\rangle$. When both T_1 and T_2 are higher in Poisson distributed graphs, and in Zipf distributed generally, the time of peak outbreak size is much less sensitive to changes in transmissibility, so that they two strains have the most similar peak time when $\tau = 0$. Introducing a delay therefore always reduces the size of the overlap.

To confirm the reasoning in the last few paragraphs, figures 4.24 and 4.23 show the time series of $\langle I_1\rangle(t)$, $\langle I_2\rangle(t)$ (solid lines), and $\langle\omega\rangle(t)$ (dotted line) from the simulations with tranmissibilities 0.6 and 0.8, and 0.3 and 0.5, i.e the largest differences in transmissibilities. Figure 4.24 shows Zipf distributed graphs, figure 4.23 shows Poisson distributed graphs. Both figures confirm that the difference in the speed of the growth of the outbreak is causing the differences in $\langle\Omega\rangle$. As shown above, in the Poisson graph a small delay increases the overlap when the smaller outbreak starts first.

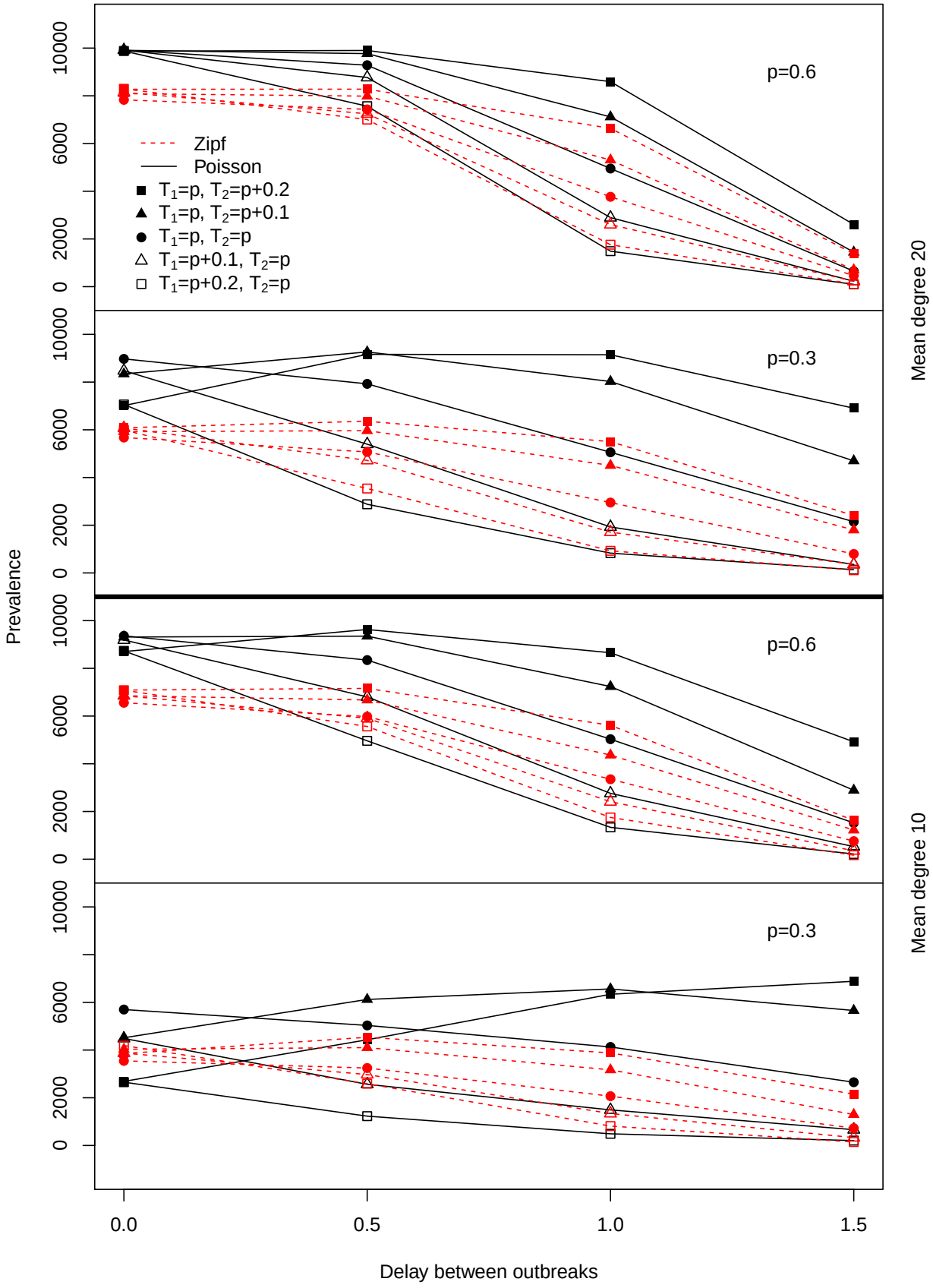


Figure 4.22: $\langle\Omega\rangle$ between two strains with different T and τ.

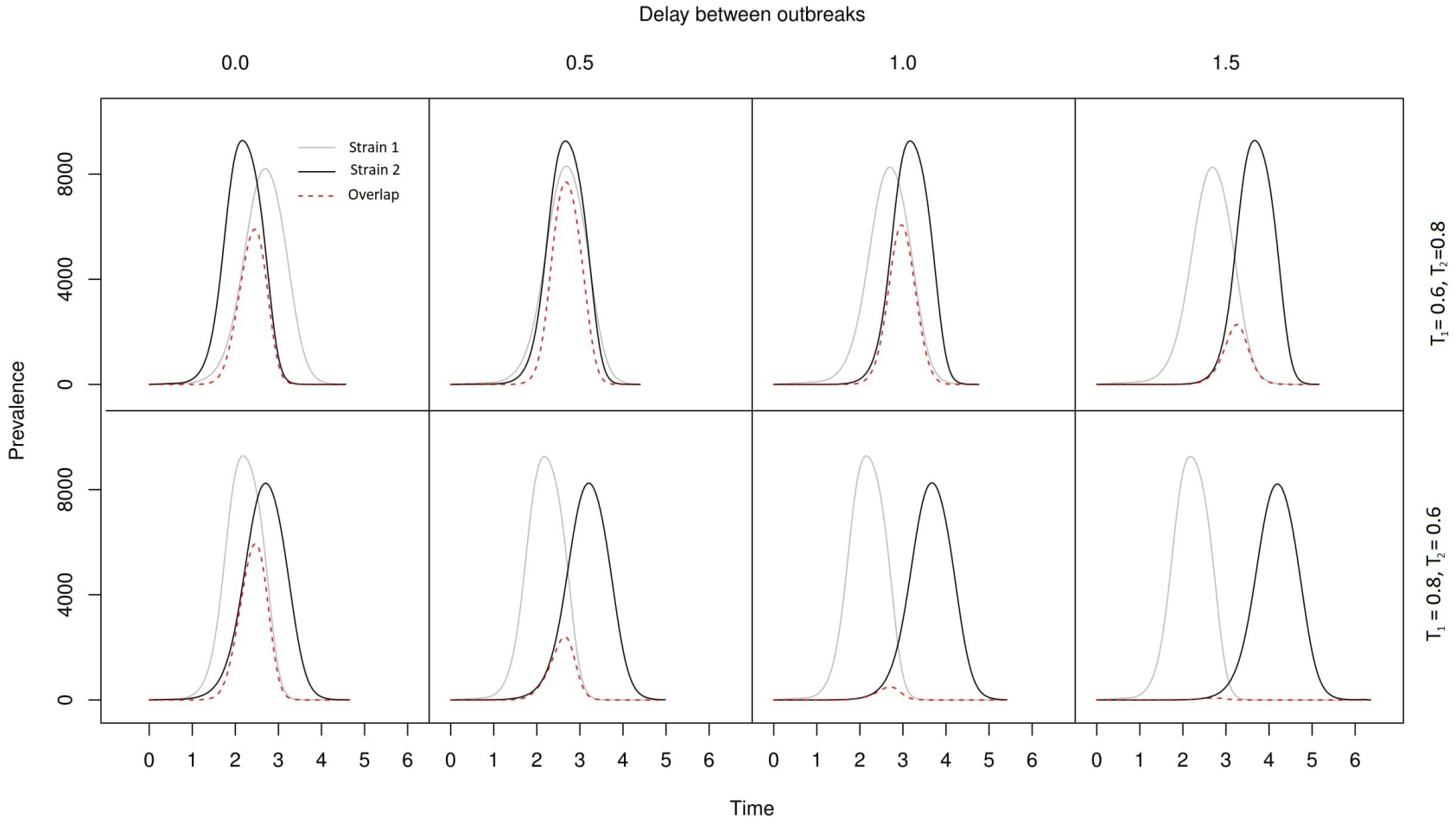


Figure 4.23: $\langle I_1 \rangle(t)$ (black lines), $\langle I_2 \rangle(t)$ (grey lines) and $\langle \omega \rangle(t)$ (dotted red lines) in Poisson distributed graphs with $\langle k \rangle = 20$. In the upper panels, $T_1 = 0.6$ and $T_2 = 0.8$ and vice versa in the lower panels. tau varies going left to right. Time series are calculated in intervals of $0.1t.u$.

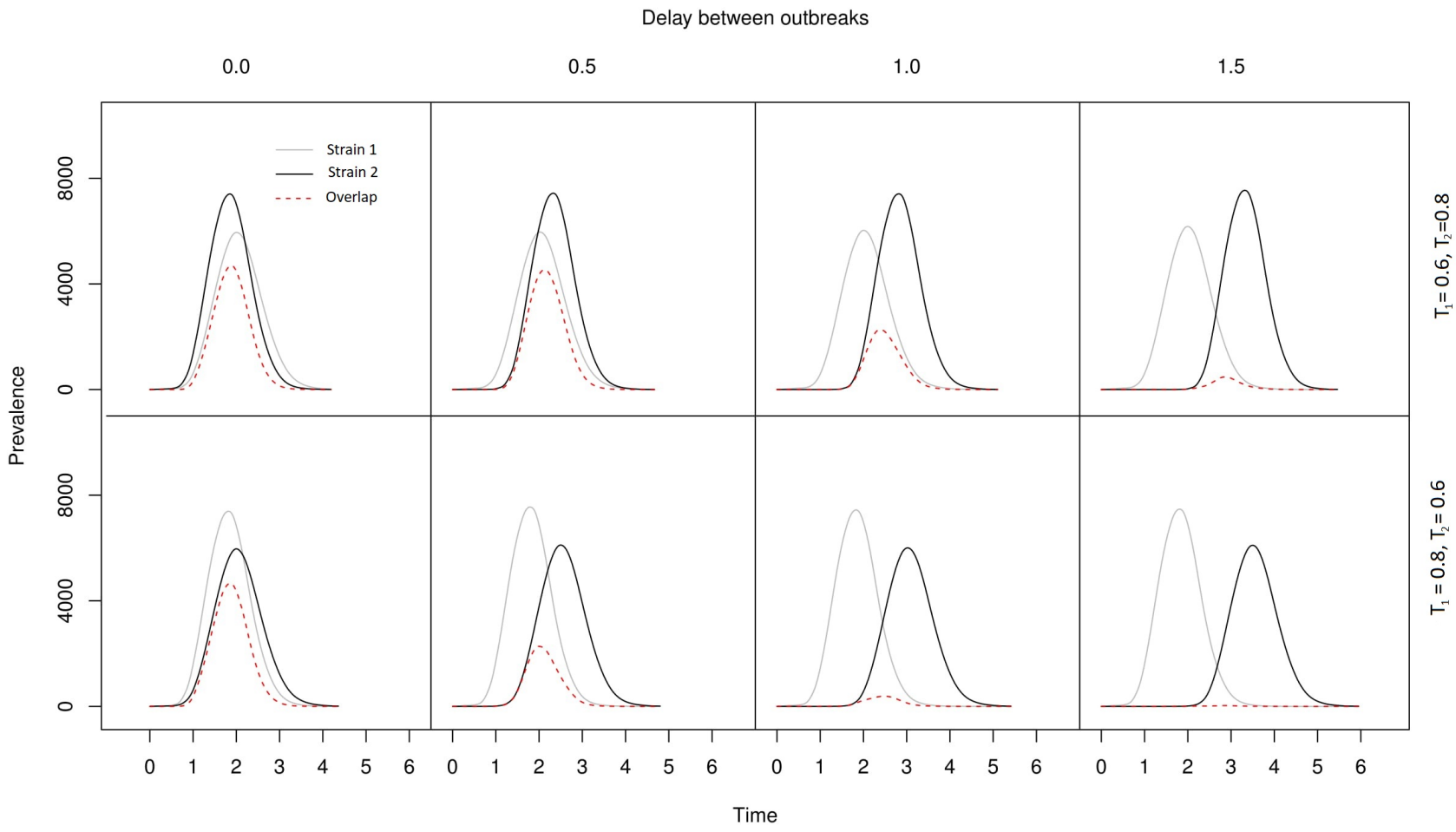


Figure 4.24: $\langle I_1 \rangle(t)$ (black lines), $\langle I_2 \rangle(t)$ (grey lines) and $\langle \omega \rangle(t)$ (dotted red lines) in Zipf distributed graphs with $\langle k \rangle = 20$. In the upper panels, $T_1 = 0.6$ and $T_2 = 0.8$ and vice versa in the lower panels. *tau* varies going left to right. Time series are calculated in intervals of $0.1t.u$.

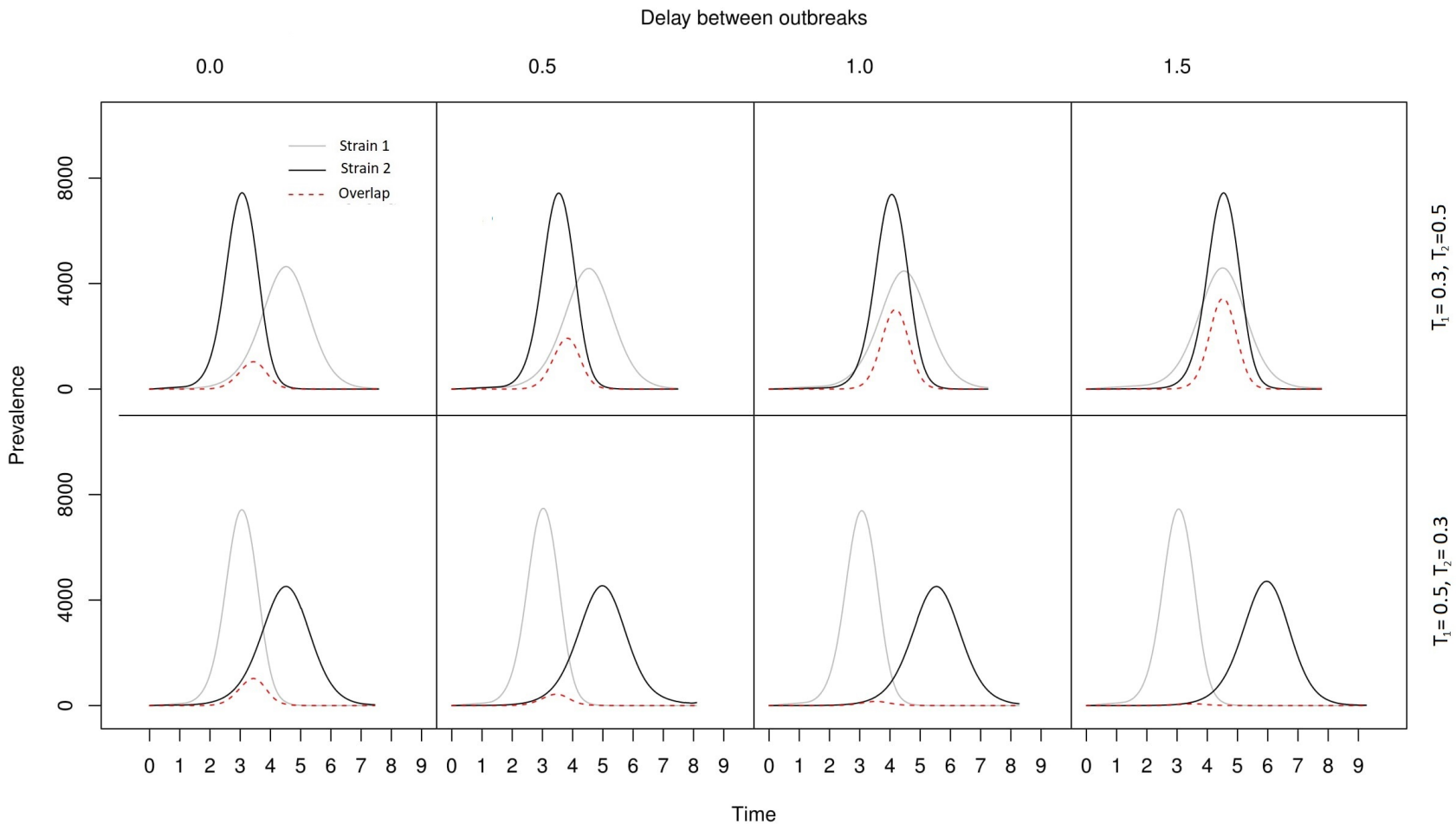


Figure 4.25: $\langle I_1 \rangle(t)$ (black lines), $\langle I_2 \rangle(t)$ (grey lines) and $\langle \omega \rangle(t)$ (dotted red lines) in Poisson distributed graphs with $\langle k \rangle = 20$. In the upper panels, $T_1 = 0.3$ and $T_2 = 0.5$ and vice versa in the lower panels. *tau* varies going left to right. Time series are calculated in intervals of $0.1t.u$.

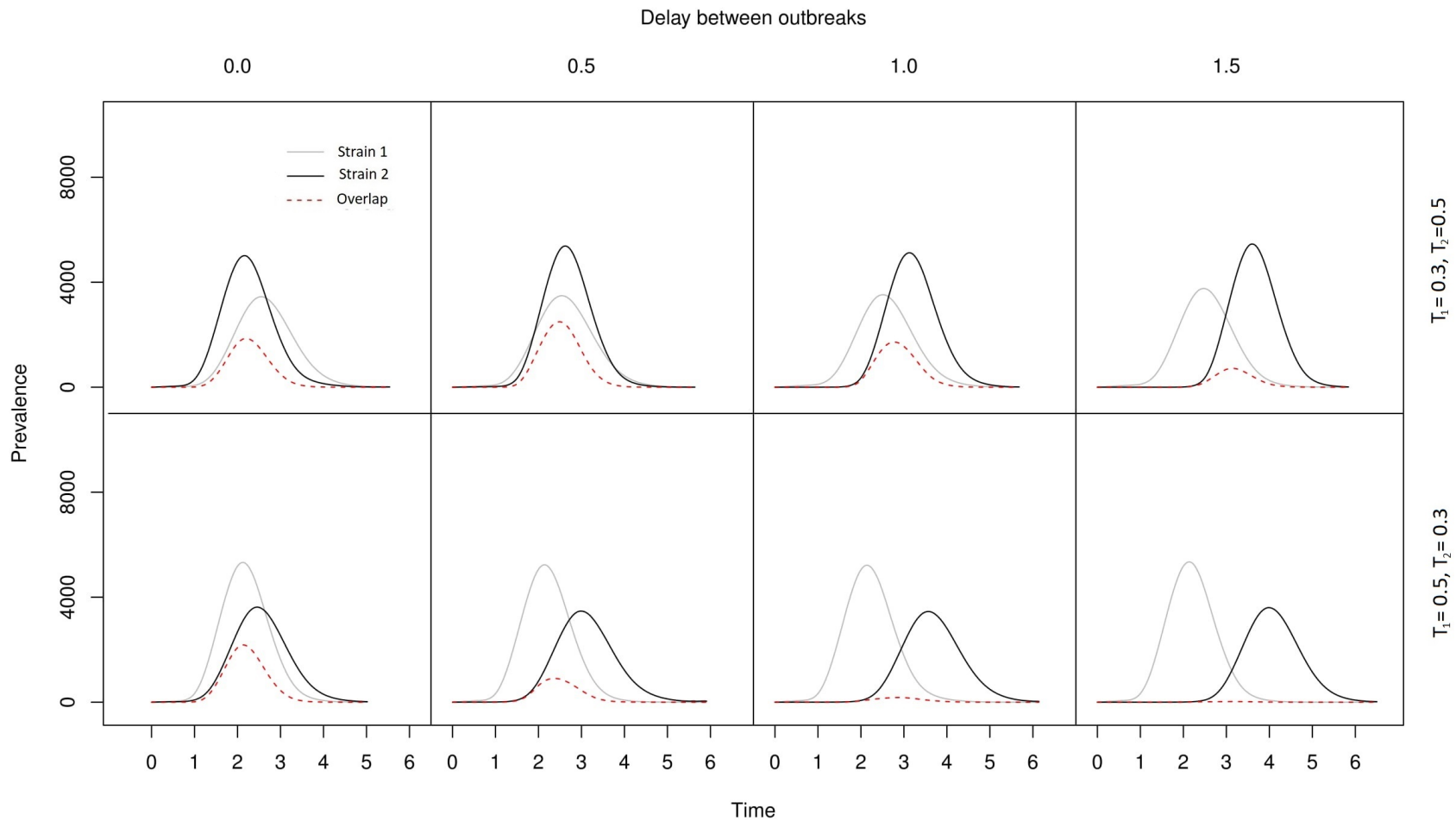


Figure 4.26: $\langle I_1 \rangle(t)$ (black lines), $\langle I_2 \rangle(t)$ (grey lines) and $\langle \omega \rangle(t)$ (dotted red lines) in Zipf distributed graphs with $\langle k \rangle = 20$. In the upper panels, $T_1 = 0.3$ and $T_2 = 0.5$ and vice versa in the lower panels. *tau* varies going left to right. Time series are calculated in intervals of $0.1t.u$.

4.6 Summary

In this chapter I have looked at the overlap between outbreaks in the absence of any interaction between strains within hosts. It is apparent that many factors affect the size of the overlap Ω. When both strains are introduced to the population at the same time and have equal transmissibility, there are three factors that affect overlap size. Changing the mean degree $\langle k \rangle$ of the graph or T both have a very large effect on Ω. Finally, overlaps are larger in Poisson graphs than in Zipf graphs, all other things being equal, but this is due to the individual outbreaks being larger. Overlaps in Zipf distributed graphs are larger relative to the size of the outbreaks than in Poisson distributed graphs. In the Zipf distributed graphs the majority of coinfections overlap in time, and the size of this overlap is larger than in the well-mixing case. In Poisson distributed graphs the majority of coinfections do not overlap in time unless either $\langle k \rangle$ of the graph or transmissibility is very high.

This difference is most likely due to the role that high-degree vertices play in Zipf graphs. High degree vertices are more likely to be infected than other vertices, and are therefore much more likely to be found in the overlap. They make up a disproportionate fraction of the overlap, and are among the first vertices to become coinfected.

When $\langle k \rangle$ or transmissibility or both are high, introducing a delay between outbreaks of strains with the same transmissibility reduces Ω dramatically, and this effect is similar in both graph types, though somewhat larger in the Poisson distributed graphs. When $\langle k \rangle$ is low however, introducing a delay causes a much smaller reduction in overlap size in the Poisson distributed graphs than in the Zipf distributed graphs.

When the two outbreaks start simultaneously but have different transmissibility, overlap size is necessarily capped by the size of the smaller outbreak. This limit notwithstanding, increasing T of one strain increases the overlap size, while decreasing T decreases the overlap size. Increasing T of both strains while keeping the difference in T fixed reduces this effect, while decreasing T of both strains increases the effect. The increase and reduction are much more pronounced in Poisson distributed graphs.
Because the outbreaks and overlaps are clustered in high-degree vertices in the Zipf distributed graphs, the overlap grows less in response to changes in T than in the Poisson distributed graphs.

Finally, when the two outbreaks have different T and one is delayed, we see quite complex behaviour. The most significant result is that the order of the outbreaks matters a great deal - when the larger outbreak is delayed, the reducing effect of the delay on the overlap is dampened, while when the smaller outbreak is delayed the overlap size decreases even faster with greater delay. In Poisson distributed graphs, when $\langle k \rangle$ or T are low, this "dampening" effect is so strong that delaying the more infectious outbreak actually increases the overlap size.

Thus, when two strains are spreading independently in the same population, we should expect coinfections to cluster in the more well-connected individuals, and to cascade from the most well-connected individuals to the less well-connected ones. However, this conclusion should be treated with some care, because it is plausible that the degree of clustering will also affect the overlap, and the graphs generated with my algorithm are more clustered in the Zipf distributed graphs than the Poisson distributed ones. For diseases with short duration, such as influenza, we should not expect that there will be much coinfection unless both strains were introduced into the population at very similar times. This does not hold for diseases that cause permanent infections, such as HIV.

Chapter 5

Two interacting strains

5.1 Introduction

In this chapter I examine the dynamics of two strains that can affect each other's ability to spread. As discussed in chapter 2, there are several ways in which an outbreak of one disease or strain can affect the spread of a second, of which there are broadly two major mechanisms of relevance here: *immune modification*, in which infection with strain 1 can alter the host's ability to resist infection with strain 2, and *transmission modification*, in which infection with strain 1 can alter the ability of the host to transmit strain 2 to other hosts.

The two models are very similar. In the first model, a vertex v that becomes infected with strain 1 experiences a change in its susceptibility to infection with strain 2. Any of its neighbours who are not themselves infected with strain 2 also experience this change indirectly since the chance that they receive infection from v has changed, and hence their overall probability of becoming infected has changed. However, if strain 2 does infect v, its neighbours no longer experience the changed transmissibility by strain 2 from v. In the second model, v experiences no change in susceptibility, but its neighbours do, and this effect persists even if v should become infected with strain 2, because its ability to transmit strain 2 is changed so long as it is infected with strain 1. The difference is illustrated in figure 5.1. As the former model has been studied extensively both in the well-mixing case and on a range of networks, in this chapter I consider only the latter model.

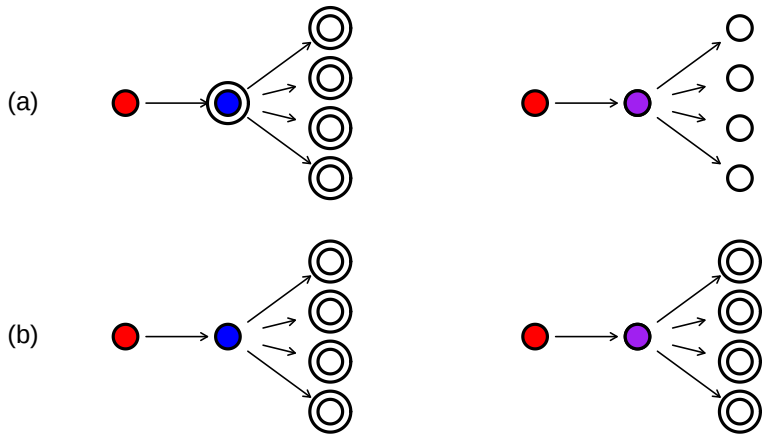


Figure 5.1: Schematic of two different models of the effect of coinfection on the spread of disease: (a) immune modification and (b) transmission modification. Blue indicates infection with the modifying strain, red with the other strain, and purple coinfection. A larger circle surrounding the vertex means that vertex experiences a change in susceptibility to infection.

I consider two scenarios, in which a) infection with strain 1 partially inhibits the transmission of strain 2, or b) enhances the transmission of strain 2.

Beyond limiting interactions to strain transmissibility, I am not trying to model any one specific mechanism of strain interaction seen in nature, and so I have kept the model quite generic. This means that the model is not designed to be an accurate representation of the interactions of the outbreaks of any given set of pathogens or strains, but rather a guide to the behaviour of a range of systems. By varying the parameters of the interaction model, it can provide an approximation

to a range of interactions. For example, as discussed in Chapter 2 reduced transmission due to infection with another strain can be the case because one strain leaves the victim bedridden and so less likely to meet others whom it can infect, or it can model a reduction in the concentration of pathogen in the blood stream due to competition or antagonistic interactions between the two pathogens. Since hosts are not killed or otherwise affected by being infected, the model can also be used to look at scenarios where one strain represents a pathogen and the other represents a host behavioural change that changes the chance of infection, such changes in the likelihood of seeking treatment, adoption of prophylactic measures, or of getting vaccinated.

Extending the basic simulation model to model transmission modification is simply a case of adding two additional sets of parameters: $\lambda_{1|2}$, $\epsilon_{1|2}$, $\Gamma_{1|2}$ and $\lambda_{1|2}$, $\epsilon_{1|2}$, $\Gamma_{1|2}$. As with the non-interaction case, throughout this chapter $\Gamma_{1|2} = \Gamma_{2|1} = \Gamma_1 = \Gamma_2 = 1$, and $\epsilon_{1|2} = \epsilon_{2|1} = \epsilon_1 = \epsilon_2 = 0.2$. Figure 5.2 shows the compartmental diagram in the case of two interacting strains. Only the four infection parameters are indicated. As in previous chapters, the values of the λ parameters are calculated to give a specific transmissibility T, and I will refer to parameter values using T, giving the parameters $T_{1|2}$ and $T_{2|1}$. For $T_{1|2}$ strain 2 is the *modifying strain*, and strain 1 the *modified strain*, and $\delta T = T_{1|2} - T_1$ is the *modifier*.

This chapter covers a several modification schemes. The modifier either increases transmission (positive modifier) or decreases it (negative modifier). Although positive and negative modifications in all probability represent different mechanisms in real systems, I will mostly consider how the dynamics vary over the range of modifier values.

The modification could take place at two stages of the infection cycle: when the host is infected with the modifying strain, or after the modifying infection has been cleared. In this thesis I only consider the former case.

I consider three different schemes of modification:

- strain 2 is modified by strain 1 ($T_{1|2} = T_1 = T_2 \neq T_{2|1}$);
- strain 1 is modified by strain 2 ($T_{1|2} \neq T_1 = T_{2|1} = T_2$);
- both strains modify each other ($T_{1|2} = T_{2|1} \neq T_1 = T_2$).

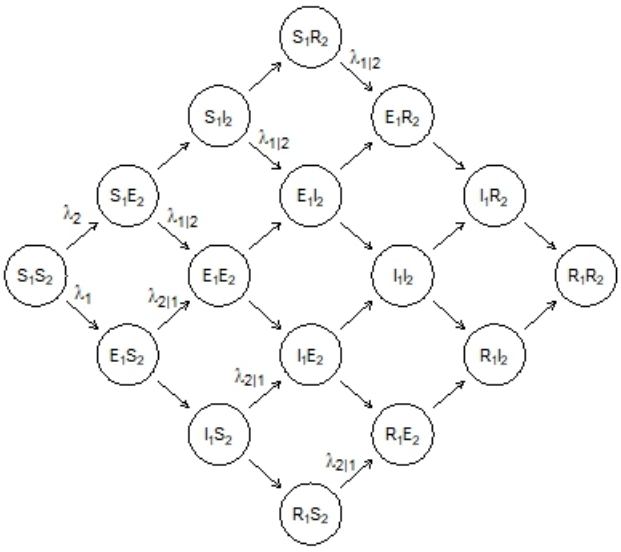


Figure 5.2: SEIR diagram for two interacting strains. Each compartment denotes the state of the vertex with respect to both strains. Only arrows denoting infection are marked with rates.

.

The first two schemes, in which only one strain experiences modification, I call *asymmetric modification*. In both cases I introduce a delay τ between the two strains, where the strain 2 is always the delayed strain. The third scheme I call *symmetric modification*. Here too only strain 2 is delayed. As in the previous chapters, all simulations are done on graphs with 10000 vertices.

5.2 Modifying the delayed strain

In the first modification scheme, infection with strain 1 changes the transmissibility of strain 2. In all simulations $T_1 = T_2 = T_{1|2} = 0.5$. $T_{2|1}$ varies from 0.2 to 0.8 in increments of 0.1, giving $\delta T = \pm 0.1$, 0.2 or 0.3. Strain 2 is either introduced simultaneously with strain 1, or after a delay $\tau 1$ or $2t.u.$. The mean degree $\langle k \rangle$ of the host graphs is 10 or 20.

Figure 5.3 shows $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$ in both Zipf and Poisson distributed graphs with $\langle k \rangle = 10$, and figure 5.4 for $\langle k \rangle = 20$. In Poisson distributed graphs, in the absence of modification both O_1 and $O_2 \geq 0.99N$, and so modification cannot increase the Ω significantly. Although there is a visible difference between $\langle O_1 \rangle$ and $\langle O_2 \rangle$ when δT is strongly negative (-0.2 and -0.3), $O_2 > 0.98N$.

Introducing a delay has no discernible effect on the size of either outbreak, but $\langle \Omega \rangle$ shrinks dramatically. When $\tau = 1t.u.$, $\langle \Omega \rangle$ depends strongly on δT, shrinking when $\delta T < 0$ and growing when $\delta T > 0$.

When $\tau = 2t.u$ in the Poisson distributed graph with $\langle k \rangle = 10$, $\langle \Omega \rangle$ also depends on δT, and when $\delta T = -0.3$, $\langle \Omega \rangle < 0.01N$, rising to approximately $0.2N$ when $\delta T = 0.3$. When $\langle k \rangle = 20$, $\langle \Omega \rangle < 0.01N$ for all but the largest positive modifier, and $\langle \Omega \rangle < 0.05N$ in all cases.

In Zipf distributed graphs, when $\tau = 0$ or $1t.u.$ the presence of a modifying strain changes $\langle O_2 \rangle$. This effect is less pronounced with a delay than without, and is completely absent when $\tau = 2t.u.$

The relationship between $\langle \Omega \rangle$, $\langle O_1 \rangle$ and $\langle O_2 \rangle$ is very similar in Poisson and Zipf distributed graphs. The only noticeable difference is that in Zipf distributed graphs, even for very large positive δT, when $\tau = 2t.u.$, $\langle \Omega \rangle < 0.01N$ both in graphs with $\langle k \rangle = 10$ and with $\langle k \rangle = 20$, whereas in Poisson distributed graphs, $\langle \Omega \rangle > 0.1N$ when $\langle k \rangle = 10$.

Figure 5.5 shows how $\langle k_{I_1} \rangle$, $\langle k_{I_2} \rangle$ and $\langle k_\Omega \rangle$ develop over the course of the simulation in Zipf distributed graphs in the presence of strong negative modifier ($\delta T = -0.3$, lower panel) and for a strong positive modifier ($\delta T = 0.3$, upper panel). $\tau = 0$ Figure 5.6 shows the same for simulations where $\tau = 1t.u$. Comparing these figures to the case when the strains spread independently (figure 4.8), we see that transmission modification does not affect $\langle k_{I_2} \rangle$ or $\langle k_\omega \rangle$ early in the course of the simulation.

However, $\langle k_{O_2} \rangle$ and $\langle k_\Omega \rangle$ *are* affected (figure 5.7, although the effect is small. When strain 1 facilitates the spread of strain 2 ($\delta T > 0$) $\langle k_\Omega \rangle$ declines, and when it inhibits the spread of strain 2 ($\delta T < 0$) it increases. This suggests that the change in $\langle O_2 \rangle$ is largely down to a change in how many vertices of very low degree are infected, since these are often infected later in the course of

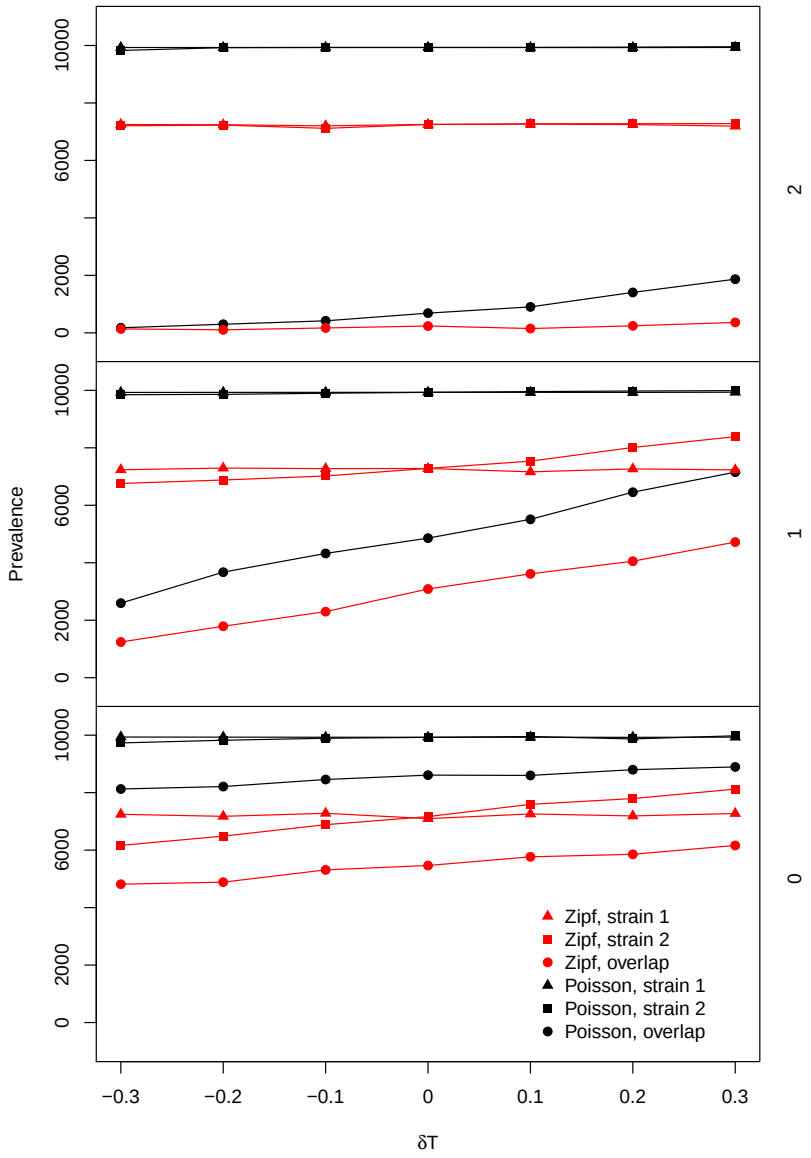


Figure 5.3: $\langle\Omega\rangle$, $\langle O_1\rangle$ and $\langle O_2\rangle$ when $T_{2|1} \neq T_1$, and τ is 0 (bottom), 1 (middle) or $2t.u$ (top panel). $\langle k\rangle = 10$.

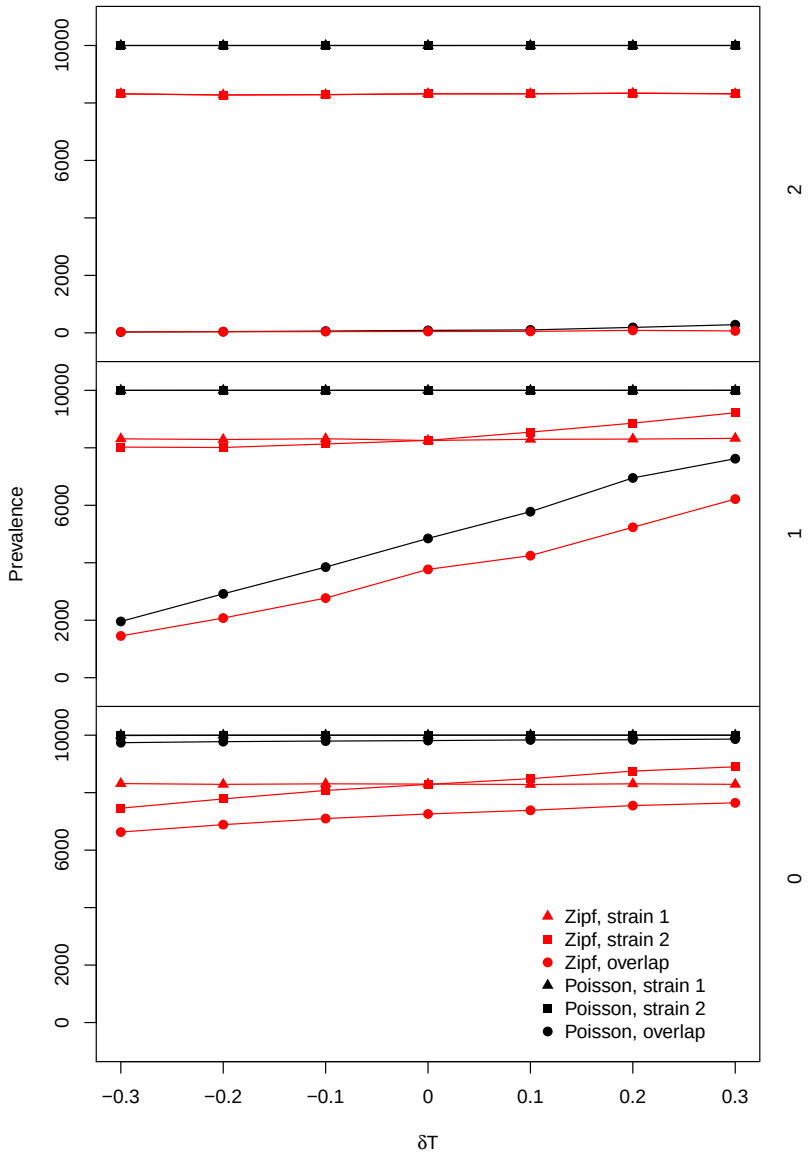


Figure 5.4: $\langle\Omega\rangle$, $\langle O_1\rangle$ and $\langle O_2\rangle$ when $T_{2|1} \neq T_1$, and τ is 0 (bottom), 1 (middle) or $2t.u$ (top panel). $\langle k\rangle = 20$.

the outbreak, and so will affect the overall mean degree but not the mean of the early stagees of the oubreak.

This is what we would expect. Suppose $\delta T = -0.2$, and consider a vertex v with only one neighbour u. If u becomes infected with strain 1 and then subsequently with strain 2, the chance that u transmits strain 2 to v is reduced by 0.2. If on the other hand v has many neighbours, the fact that u is infected with strain 2 only reduces the chance of v becoming infected with strain 2 *via* u. So the more connected a vertex is, the less effective change in susceptibility it experiences per neighbour infected with strain 1.

Note that in the Poisson distributed graphs $\langle\Omega\rangle$ depends on the modifier, despite the fact that $\langle O_2\rangle$ does not. This is because increasing or decreasing T is accomplished by altering the λ parameter used to draw the time to infection, so that a greater or smaller fraction of draws are smaller than Γ. But this also means that the fraction of draws that are smaller than t likewise changes, for all $0 < t < \Gamma$. This is illustrated in figure 5.8. Hence a vertex experiencing a negative modifier will tend to transmit infection later in its infectous period, and one experiencing a positive modifier will transmit earlier. This means that the modifier has the effect of speeding up or slowing down the outbreak of the modified strain, and hence changing the fraction of the intersection that is part of the overlap.

This means that, although the intersection Υ of the two strains will remain unchanged so long as O_1 and O_2 are unchanged, $\frac{\Omega}{\Upsilon}$ wil change. Figures 5.9 and 5.10 show the time of peak outbreak and overlap size for Poisson and Zipf distributed graphs with $\langle k\rangle = 10$ and 20 respectively. Although $\langle O_2\rangle$ is unaffected, introducing a modifier changes the peak outbreak time in both graph types, confirming that changes in $\langle\Omega\rangle$ in the Poisson graph are caused by changes in the timing between outbreaks.

Slowing down an outbreak also has the effect of reducing the peak outbreak size, and so we should expect that outbreaks of strain 2 will have smaller peaks when experiencing a negative modifier, and higher peaks when experiencing a positive modifier. Figures 5.11 and 5.12 show time series of $\langle I_1\rangle(t)$, $\langle I_2\rangle(t)$, and $\langle\omega\rangle(t)$ in graphs of both type with $\langle k\rangle = 10$ with $\tau = 0$ or $1t.u$. The peak value

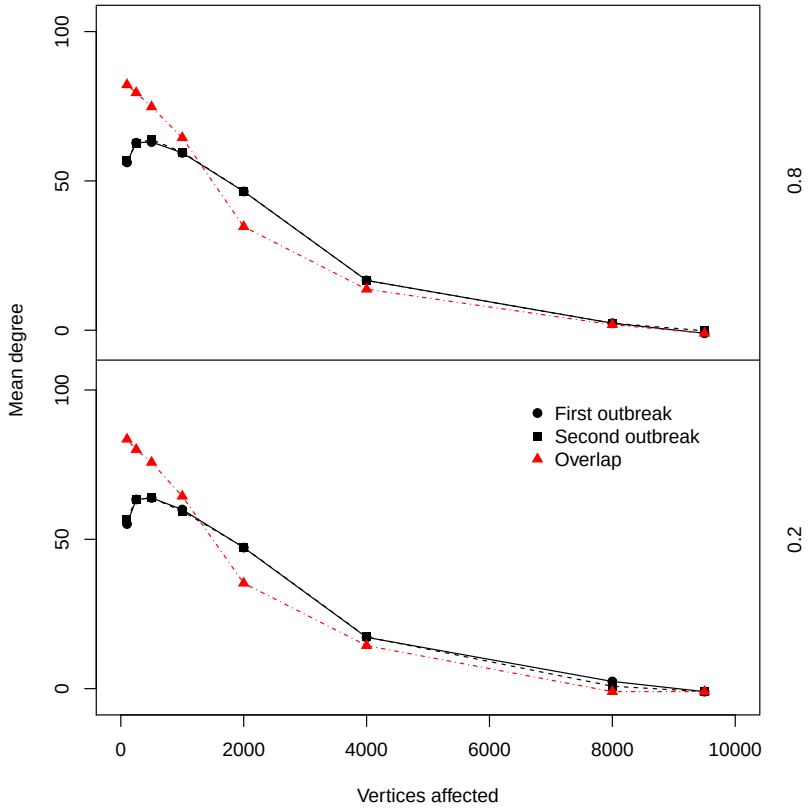


Figure 5.5: Development of $\langle k_{I_1} \rangle$ (black circles), $\langle k_{I_2} \rangle$ (black squares) and $\langle k_{\omega} \rangle$ (red triangles) over the course of the simulations in Zipf distributed graphs when $T_{2|1} \neq T_2$, and $\delta T = -0.3$ (lower panel, $T_{2|1}$ given on right axis), and $\delta T = 0.3$ (upper panel). $\tau = 0t.u.$

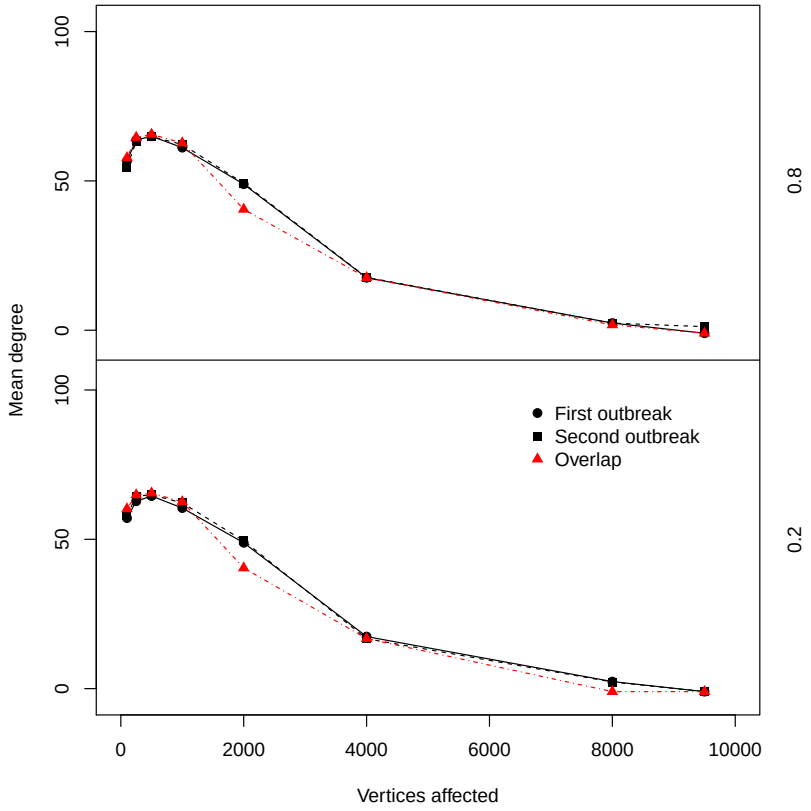


Figure 5.6: Development of $\langle k_{I_1} \rangle$ (black circles), $\langle k_{I_2} \rangle$ (black squares) and $\langle k_{\omega} \rangle$ (red triangles) over the course of the simulations in Zipf distributed graphs when $T_{2|1} \neq T_2$, and $\delta T = -0.3$ (lower panel, $T_{2|1}$ given on right axis), and $\delta T = 0.3$ (upper panel). $\tau = 1.t.u.$

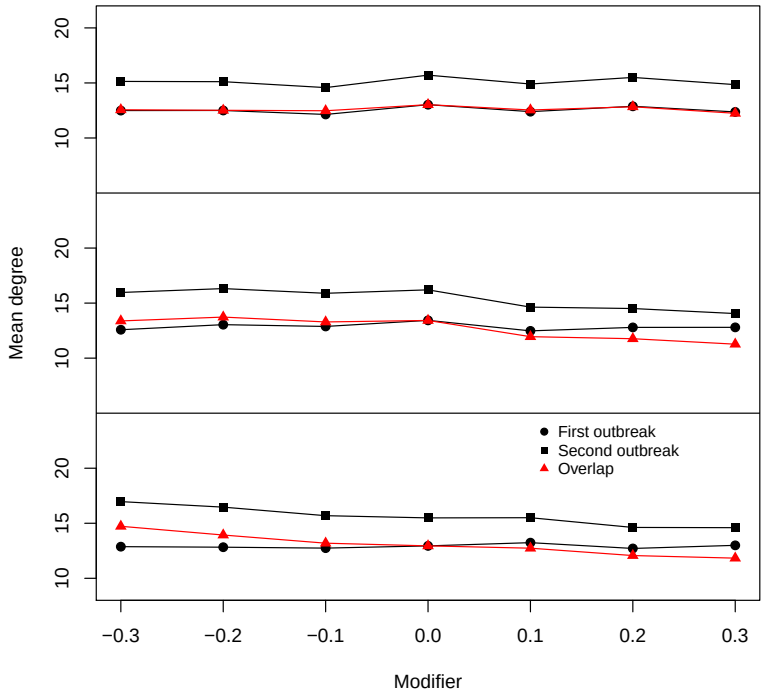


Figure 5.7: $\langle k_{O_1} \rangle$ (black circles), $\langle k_{O_2} \rangle$ (black squares) and $\langle k_{\Omega} \rangle$ (red triangles) in Zipf distributed graphs with $\langle k \rangle = 10$, when $T_{2|1} \neq T_2$ (x-axis). $\tau = 0t.u.$ (bottom panel), $\tau = 1t.u.$ (middle panel) and $\tau = 2t.u.$ (top panel).

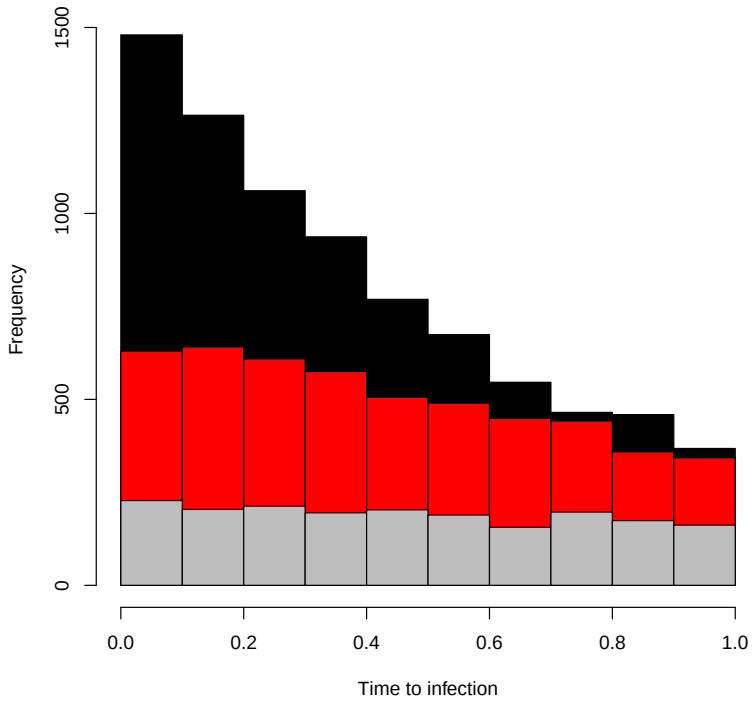


Figure 5.8: Histograms of three exponential distributions with different rate parameters λ The histogram shows the fraction of the distribution that is less than 1.0 (Γ). The λ values are chosen to give $F(x) < 1 = 0.8$ (black), 0.5 (red) and 0.2 (gray). Note the "flattening" of the distribution as $F(x) < 1$ is reduced, implying that sampling from this distribution will tend to draw more values closer to 1. Since $F(x) < 1 = T$, a flattening means that the time between becoming infected and passing the infection on will tend to increase.

of I_1 is lower and occurs later when $T < 0$, and larger and earlier when $T > 0$. This effect is much more pronounced when strain 2 is delayed.

In the conclusion of this chapter I will discuss the qualitative differences between the dynamics of the transmission-modifying coinfection model I investigate here and those of the immune modifying coinfection model, since the latter has been extensively studied previously. Although I do not simulate immune modifying dynamics, it is straightforward to estimate the fraction of the host

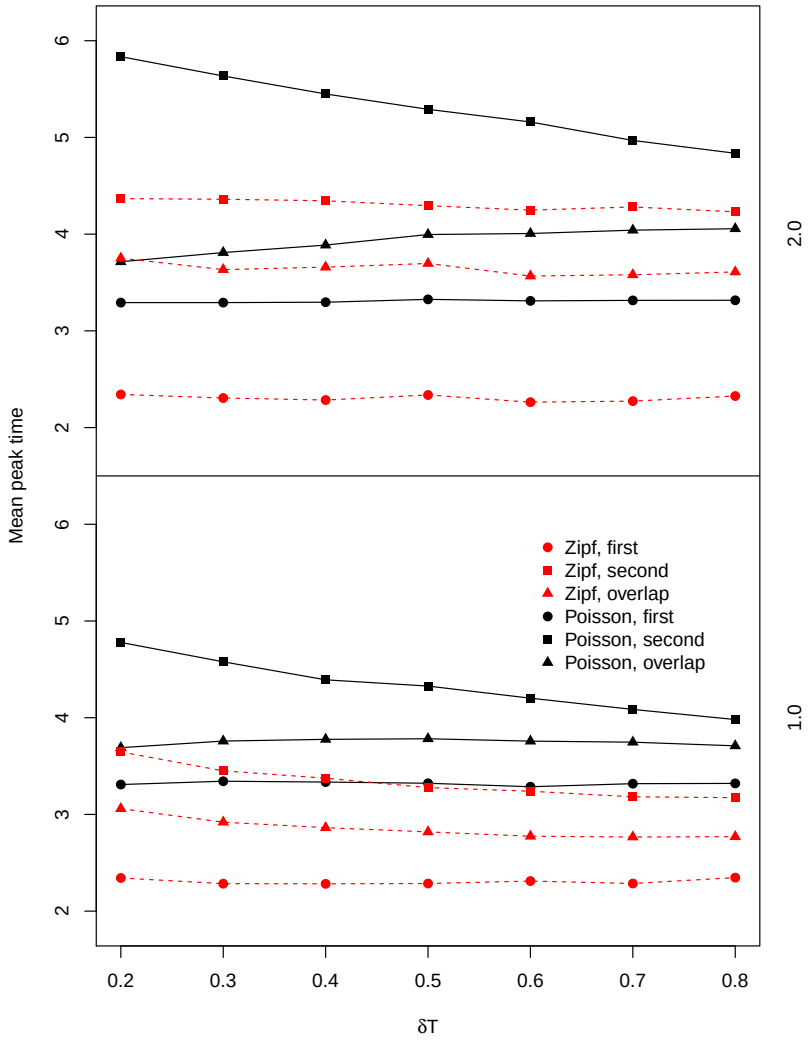


Figure 5.9: Mean peak times of overlaps and outbreaks in Zipf (red dotted lines) and Poisson distributed graphs(solid black lines). $\langle k \rangle = 10$, $\tau = 1.0$ (lower panel) and 2.0 (upper panel), for varying modifier δT.

population that would experience altered susceptibility to strain 2, since it will those vertices infected by strain 1. More precisely it will be those vertices infected by strain 1 *first*, so that the

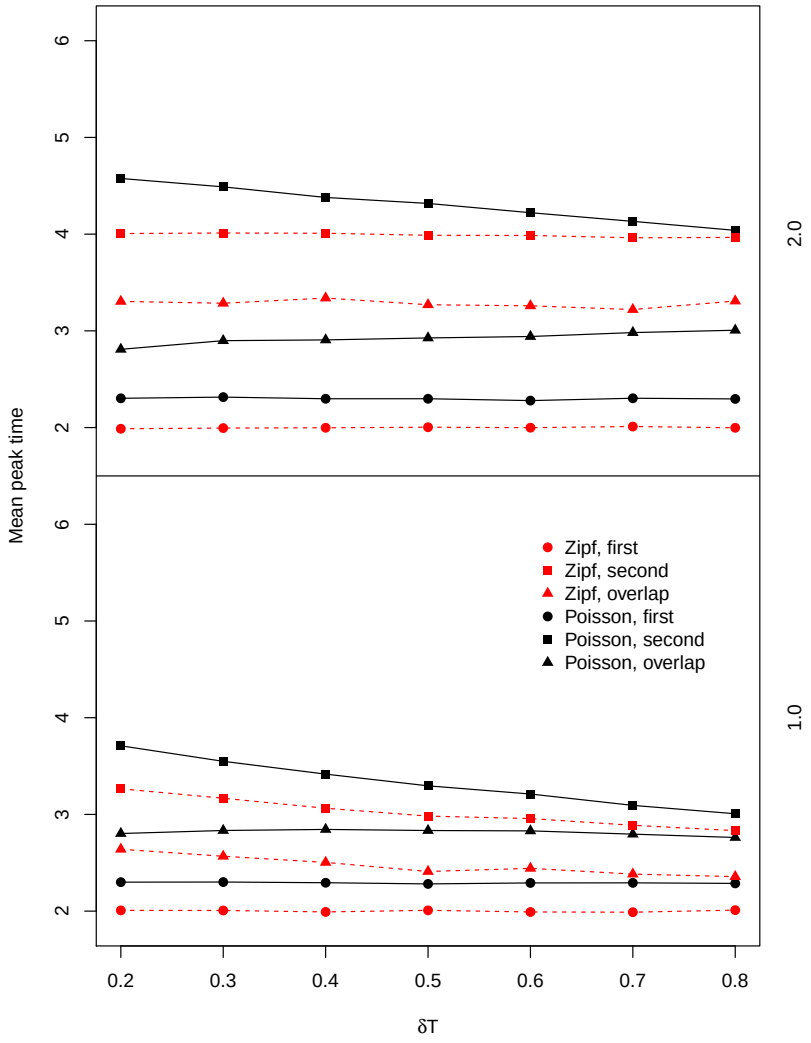


Figure 5.10: Mean peak times of overlaps and outbreaks in Zipf (red dotted lines) and Poisson distributed graphs(solid black lines). $\langle k \rangle = 20$, $\tau = 1.0$ (lower panel) and 2.0 (upper panel), for varying modifier δT.

size of the outbreak of strain 1 is an upper bound on the number of vertices that experience altered susceptibility, but this will do for our comparison.

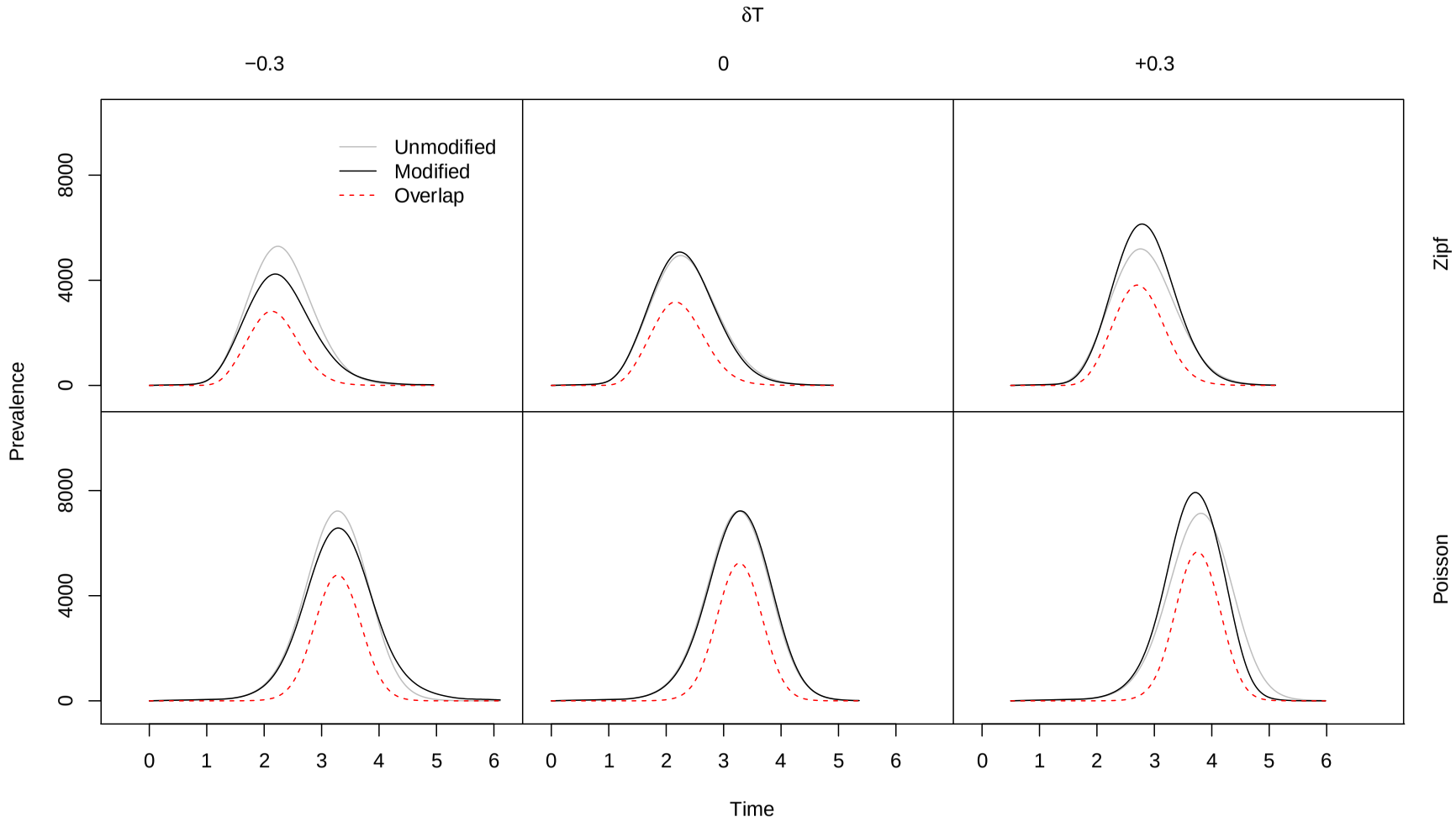


Figure 5.11: $\langle I_1 \rangle(t)$, $\langle I_2 \rangle(t)$ and $\langle \omega \rangle$ over time for different values of δT, with $\tau = 1.0 t.u.$. Strain 1 is shown in grey, strain 2 in black, and the overlap is the dotted red line.

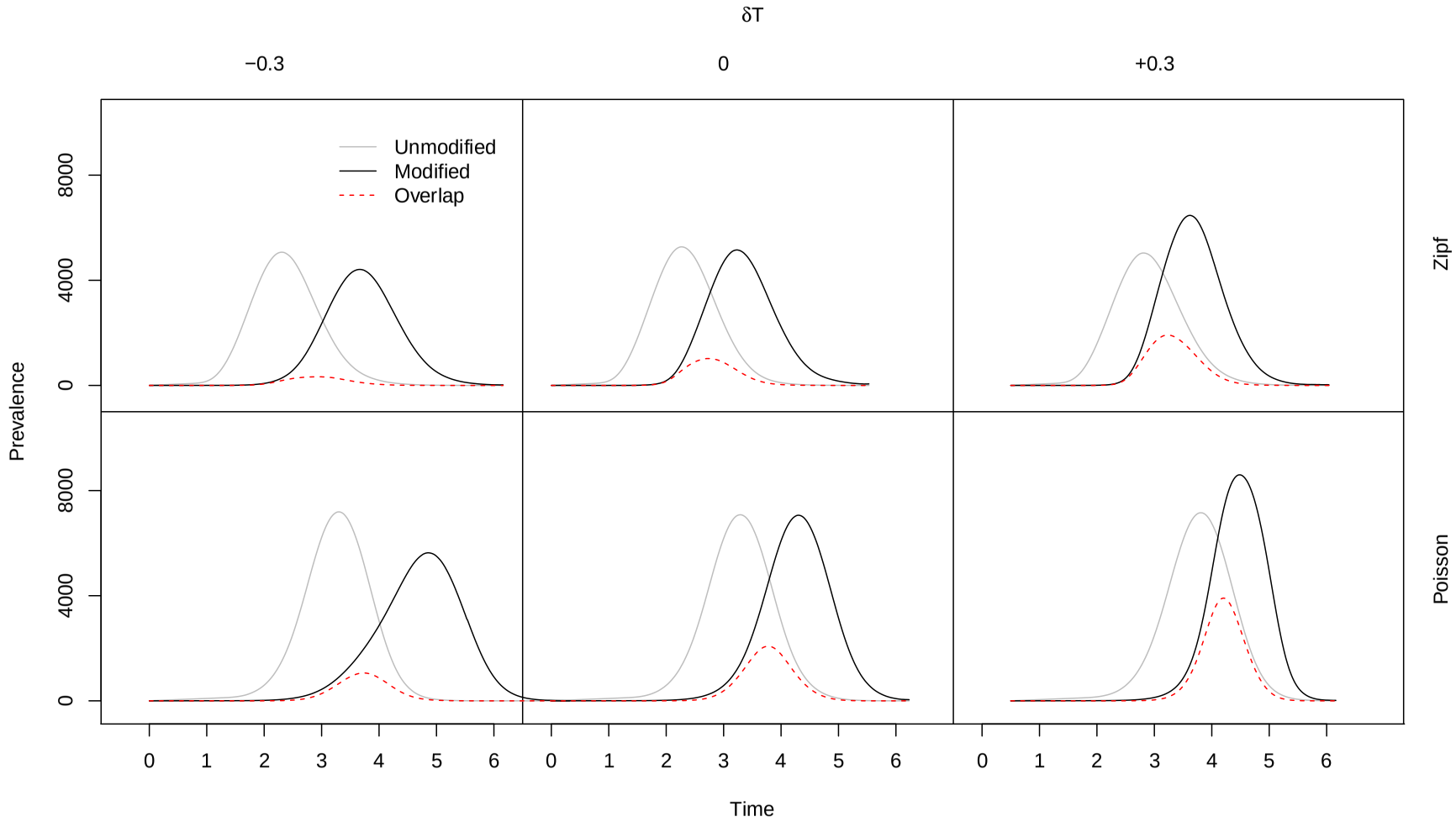


Figure 5.12: $\langle I_1 \rangle(t)$, $\langle I_2 \rangle(t)$ and $\langle \omega \rangle$ over time for different values of δt, $\tau = 2.0 t.u.$. Strain 1 is shown in grey, strain 2 in black, and the overlap is the dotted red line.

We can calculate a similar estimate for the number of vertices that experience altered susceptibility in the transmission-modifying system. In this case we do not count the number of vertices in the outbreak of strain 1, but rather the number of vertices that have a neighbour in the outbreak. Figure 5.13 shows both estimates as time series. Only the curves for graphs with $\langle k \rangle = 10$ and $\tau = 1t.u.$ is shown, but it suffices to demonstrate that the number of vertices that experience some altered T under the transmission-altering model is much greater, although as described above, each of these may be experiencing a change in susceptibility along a single edge.

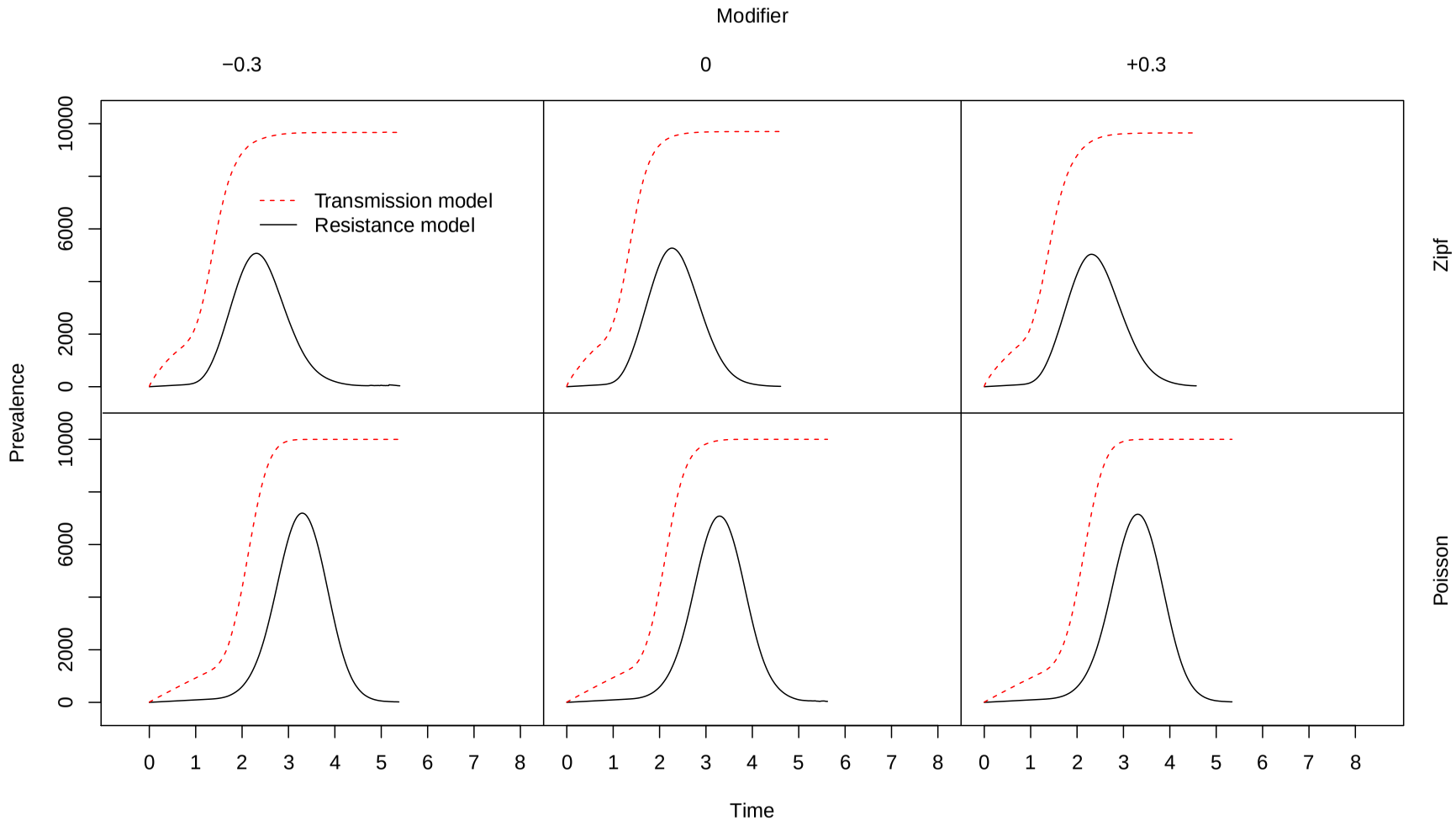


Figure 5.13: The number of vertices at each point in time with at least one neighbour infected with the modifying strain 1, so that the transmissibility from that vertex is $T_{2|1}$ (dotted red line). For comparison, the solid black line shows the number of vertices that would experience a change in effective T under the more widely studied immune modification model. $\langle k \rangle = 10$, $\tau = 1.0 t.u.$

5.3 Modifying the earlier strain

In this section I consider the second modification scheme. I ran several set of simulations. The parameter values used for the first set are identical to those in the simulations discussed in the previous section, except that $T_{2|1} = T_1 = T_2 = 0.5$, and it is $T_{1|2}$ that is allowed to vary. As I will be returning to this set of simulations several times, I refer to it throughout this section as the *main* set.

Figure 5.14 shows $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$, varying δT and τ, in both graph types with $\langle k \rangle = 10$. Figure 5.15 shows the same in graphs with $\langle k \rangle = 20$. When $\tau = 0$ the model is identical to that in the previous section, and we see the same results: $\langle O_1 \rangle$ is not affected in the Poisson distributed graphs, but is in the Zipf distributed graphs. In Poisson distributed graphs with $\langle k \rangle = 10$ and in all Zipf distributed graphs $\langle \Omega \rangle$ depends on δT, and the dependence is very similar in all cases. In Poisson distributed graphs with $\langle k \rangle = 20$ $\langle \Omega \rangle$ does not depend on δT.

Introducing a delay completely negates all effects of the modifier in all cases. When $\tau = 0$ a vertex has an even chance of being infected with strain 1 first or strain 2 first all things being equal, and so we expect a large fraction of the population to experience some change in susceptibility. As we introduce a delay before strain 2 starts spreading, a larger fraction are infected by strain 1 first and so experience no change. This fraction grows as the delay get longer until eventually no vertices experience any effect from strain 2.

To gain a better idea of how long a delay is necessary to completely remove any effect of strain 2 on the spread of strain 1, I ran a set of simulations with $\tau = 0.25$, 0.5 and $0.75t.u.$ All other parameters were as in the main set, except that simulations were only run on graphs with $\langle k \rangle = 10$. Figure 5.16 shows $\frac{O_1}{O_2}$ for increasing τ, for $\delta T \pm 0.3$. In the Poisson distributed graphs O_1 never changes by more than 3%, even when $\tau = 0$, and when the $\tau \geq 0.75t.u.$ there is no noticeable difference. In the Zipf distributed graph O_1 changes by more than 10% relative to $O)2$ when $\tau = 0$. This effect persists until $\tau > 0.75t.u.$ when $\delta T < 0$, and when $\delta T > 0$, it is still noticeable albeit small when $\tau > 1t.u.$

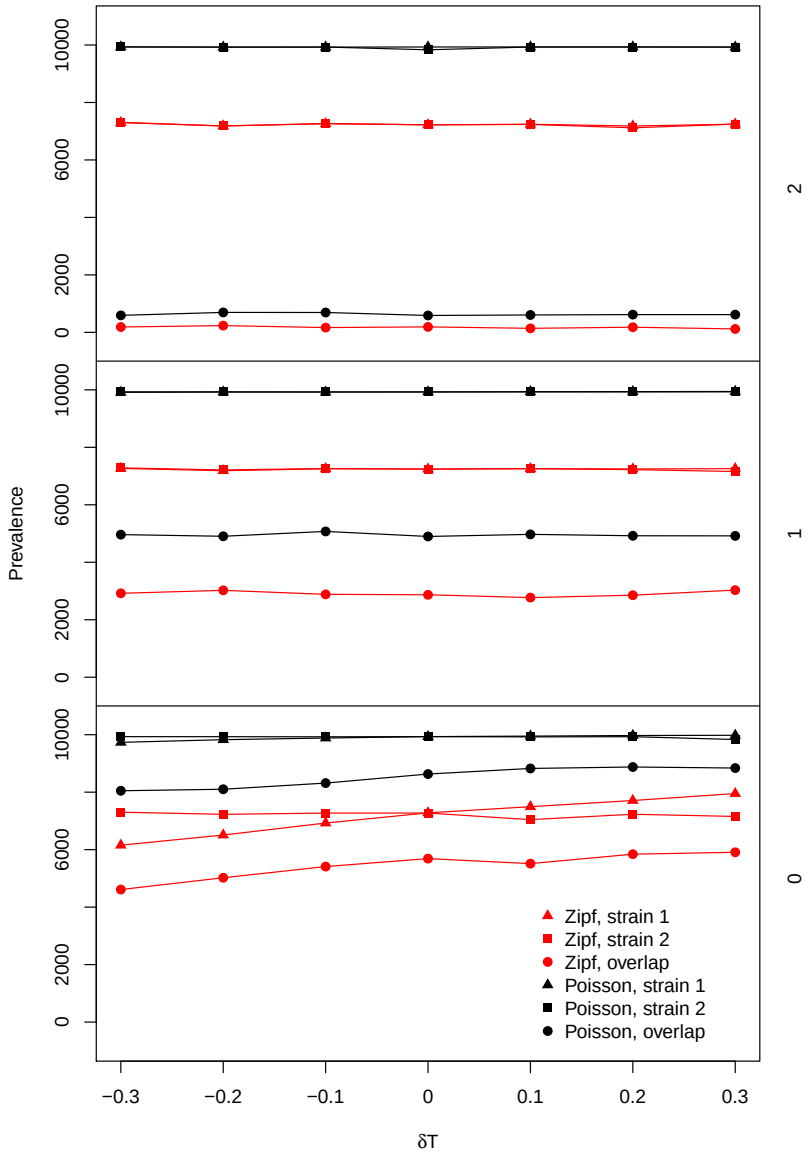


Figure 5.14: $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$ when $T_{1|2} \neq T_1$, and $\tau = 0$ (bottom), 1 (middle) or $2t.u$ (top panel). $\langle k \rangle = 10$.

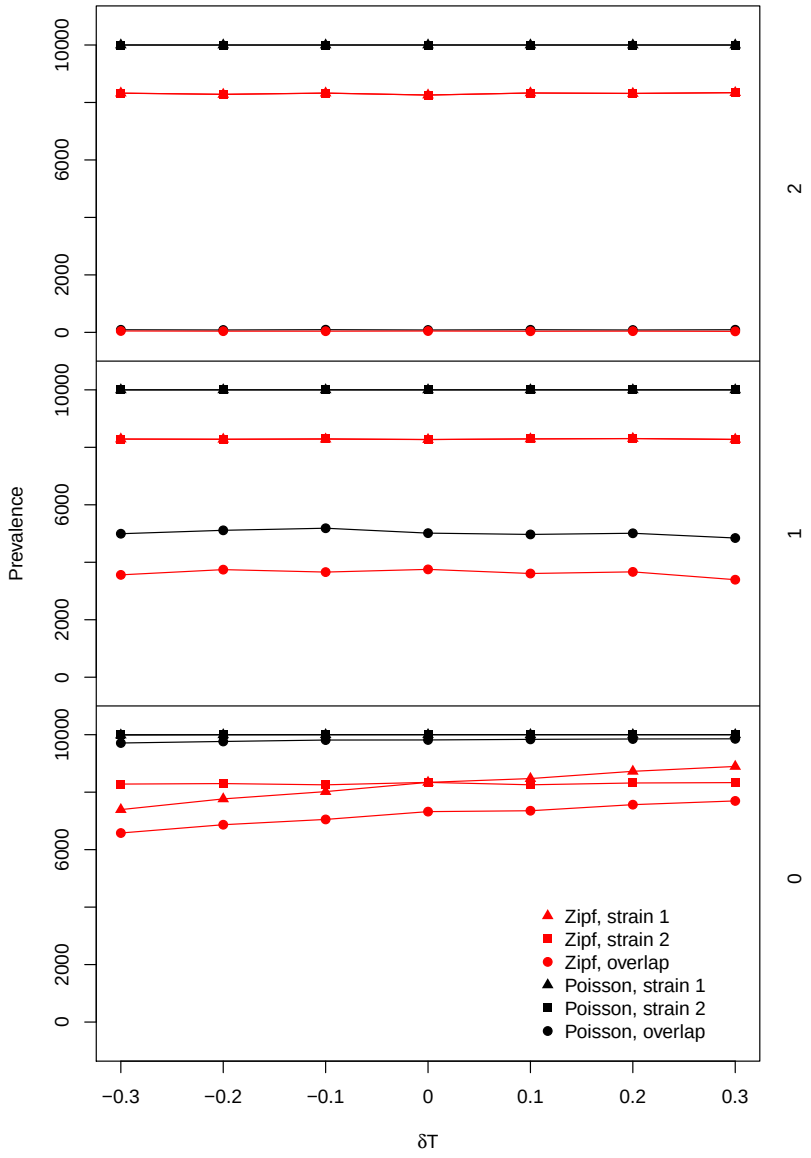


Figure 5.15: $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$ when $T_{1|2} \neq T_1$, and $\tau = 0$ (bottom), 1 (middle) or $2t.u$ (top panel). $\langle k \rangle = 20$.

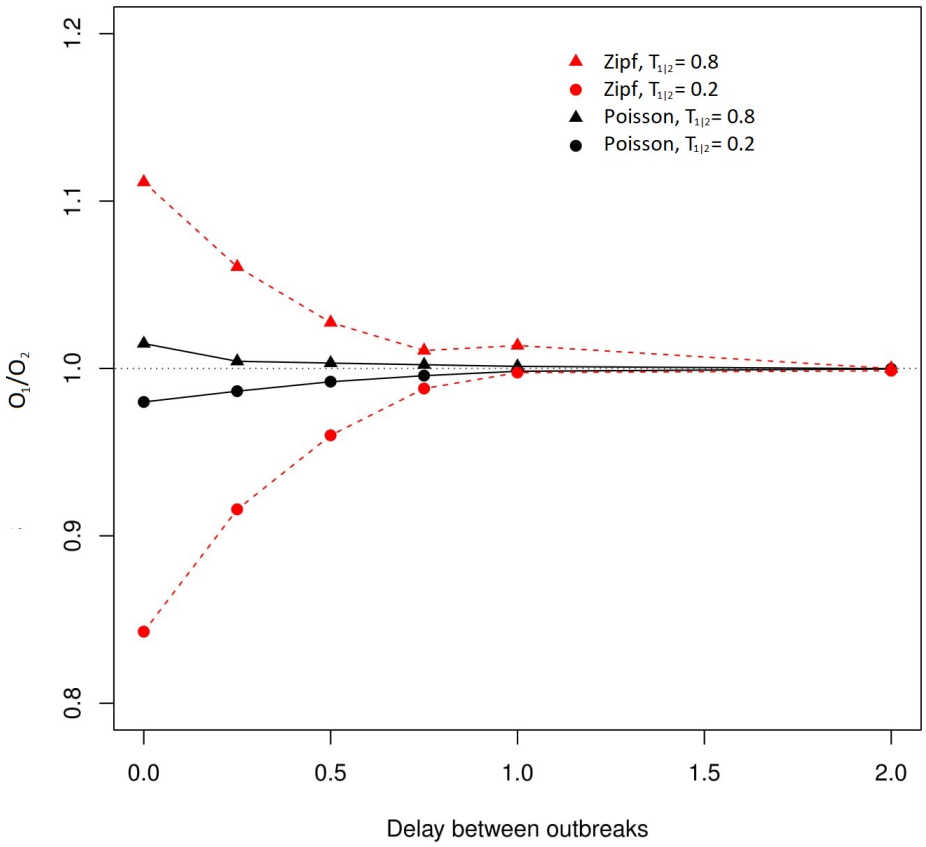


Figure 5.16: $\langle \frac{O_1}{O_2} \rangle$ as a function of τ, $\delta T = \pm 0.3$, in graphs of both types with $\langle k \rangle = 10$.

In both graph types, but in particular in the Poisson distributed graphs, outbreaks with very low transmissibility peak much later than those with higher transmissibility. If strain 2 has low transmissibility and strain 1 experiences a negative modifier, this might sustain the effect of the modifier even when strain 2 is delayed, since strain 1 will be spreading very slowly. To test this I ran a set of simulations in which $T_2 = T_1 = 0.3$ and $T_{1|2}$ was modified by ± 0.1 and 0.2. Simulations were only run in graphs with $\langle k \rangle = 10$, and with $\tau = 0$, 1 or $2t.u$. Figure 5.17 shows $\langle O_i \rangle$ and $\langle \Omega \rangle$ in these simulations.

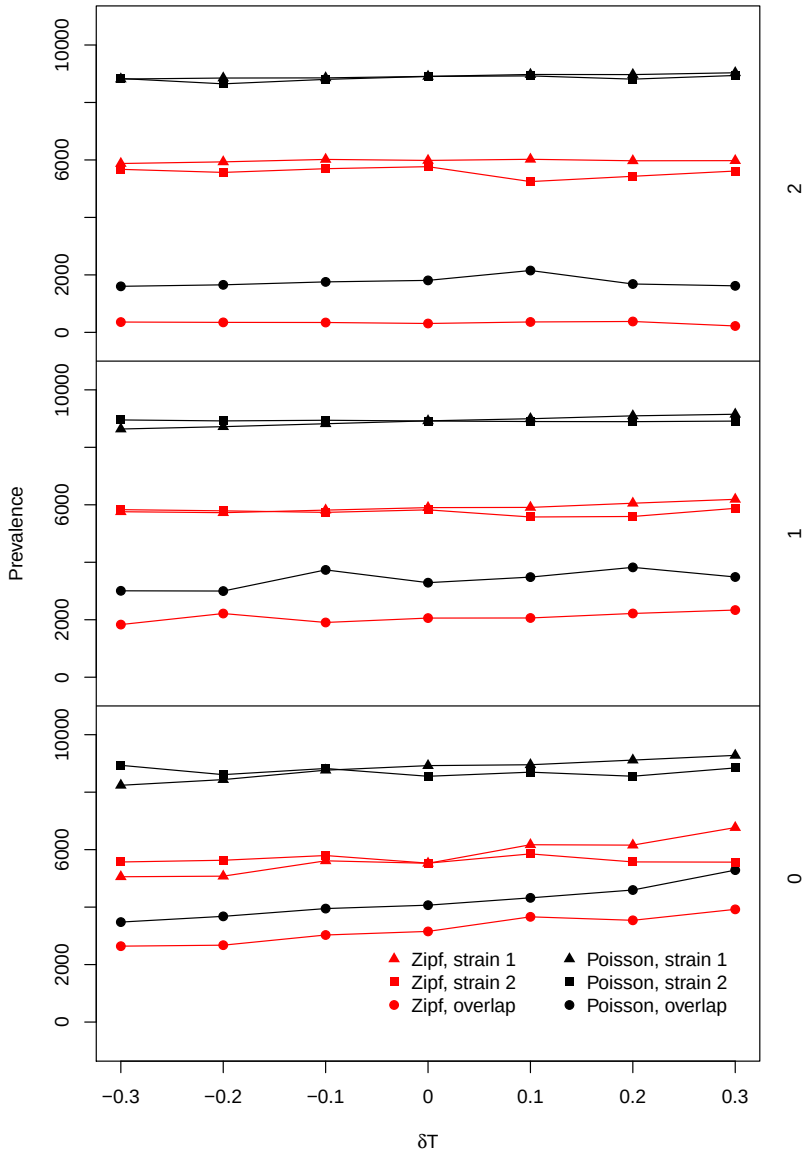


Figure 5.17: $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$ in both graph types when $T_2 = 0.3$ and $T_{1|2} = T_1 \pm 0.1,\ 0.2$. $\langle k \rangle = 10$.

From figure 5.17, it appears slowing down the outbreaks by reducing T does not enhance the modifying effect of the delayed stain on the earlier strain. It is possible that if strain 2 experienced a stronger negative modifier than 0.2 there would be some effect. Likewise, if the transmissibility of strain 1 were kept high, for example 0.7, and $\delta T < -0.2$, we might expect that there will be a more pronounced effect, but I did not simulate these parameter combinations.

Reducing $\langle k \rangle$ also has the effect of slowing outbreaks down, in particular in the Poisson distributed graphs. The results of reducing T suggests that reducing $\langle k \rangle$ will not affect the ability of strain 2 to modify strain 1, but to be sure, I ran a set of simulations identical to the main set, but on graphs with $\langle k \rangle = 5$. $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$ in this case are shown in figure 5.18. These simulations confirm that even when the outbreak of strain 1 is very slow, due to either low mean degree or transmissibility, $\tau = 1t.u.$ is sufficient to negate any modifying effect of the delayed strain on the earlier strain.

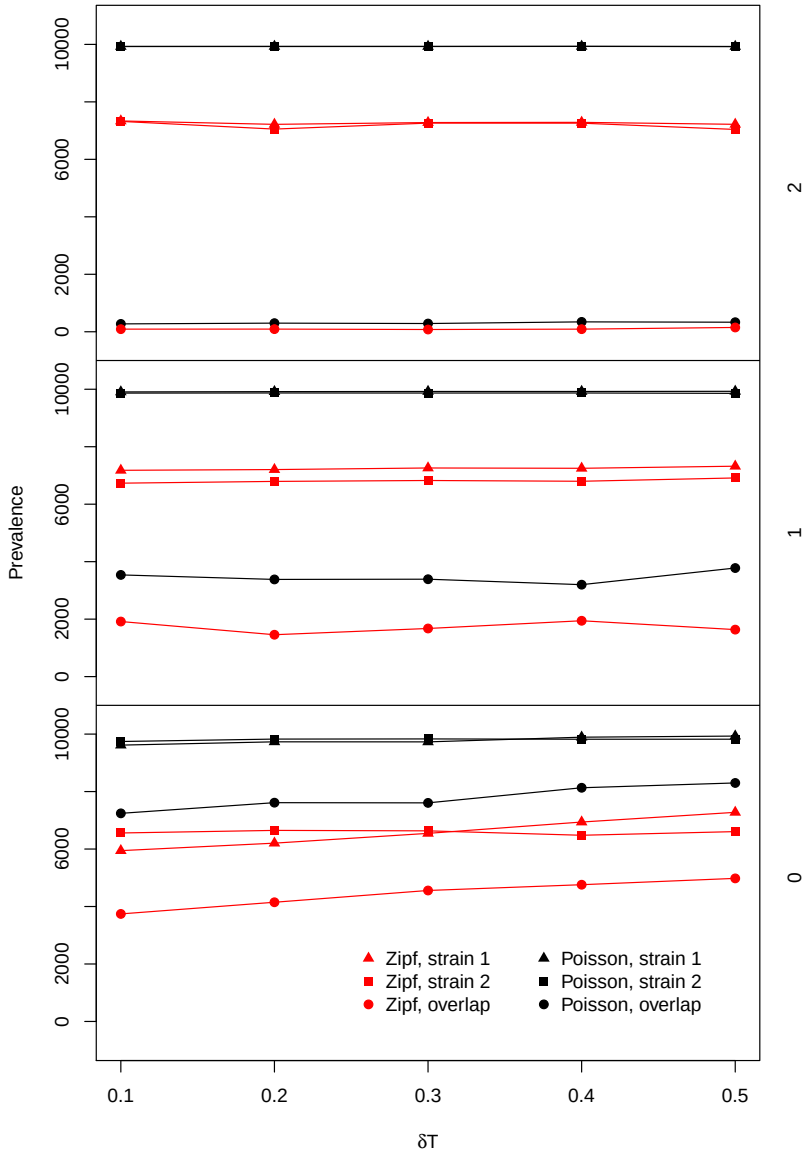


Figure 5.18: $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$ in both graph types when $T_1 = T_2 = 0.5$, varying δT, and $\langle k \rangle = 5$

5.4 Modifying both strains

The final modification scheme I consider in this chapter is symmetric modification, in which both strains modify each other equally ($T_{1|2} = T_{2|1}$). Figure 5.19 shows $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$ in graphs with $\langle k \rangle = 10$, and figure 5.20 in graphs with $\langle k \rangle = 20$.

When $\tau = 0$, the two strains modify each other to the same degree, which is also the same as the effect of strain 1 on strain 2 in the first simulation scheme in this chapter (figures 5.3 and 5.4). Once there is a delay the modification is entirely one-sided: the spread of strain 1 alters O_2, but not vice versa. From the previous two modification schemes, this is what we expect.

When $\tau = 0$, the effect of the modification on $\langle \Omega \rangle$ is larger than in the asymmetric cases, indicating that the modification effect is at least partially additive. If both modifiers tend to affect the same vertices, we might expect that the effect of each modifier on the overlap would be approximately halved, so that the overall effect of the overlap stays the same. Figure 5.21 shows the ratio of $\langle \Omega \rangle$ under symmetric modification to $\langle \Omega \rangle$ under asymmetric modification, when $\tau = 0$. The effect of modification on $\langle \Omega \rangle$ is larger with symmetric modification than asymmetric for all δT and in both graph types The effect is greater in Zipf distributed graphs. This is likely due to the fact that the vertices that experience the effect of modification most in Zipf distributed graphs are low-degree vertices. Because these vertices are very common, there chance of the two outbreaks having a modifying effect on larger separate sections of the graph is greater.

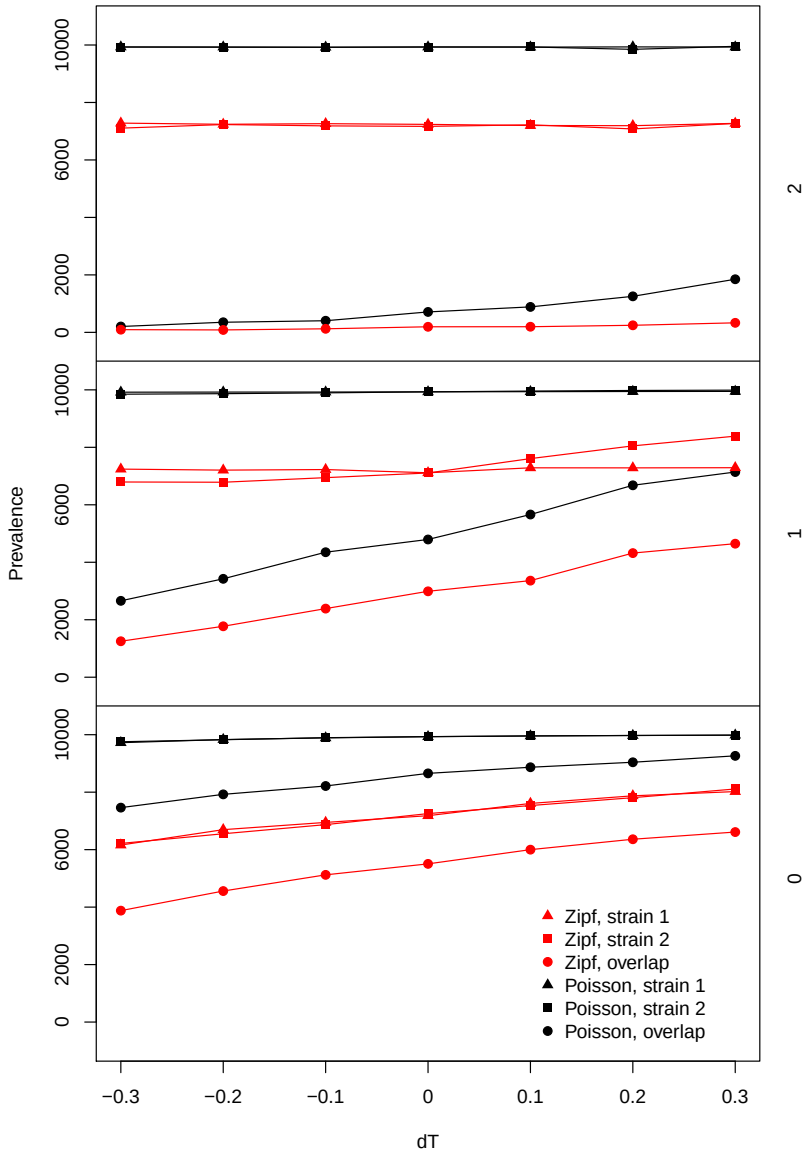


Figure 5.19: $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$ when both strains modify each other's transmissibility ($T_{1|2} = T_{2|1}$), and $\tau = 0$ (bottom), 1 (middle) or $2t.u$ (top panel). $\langle k \rangle = 10$.

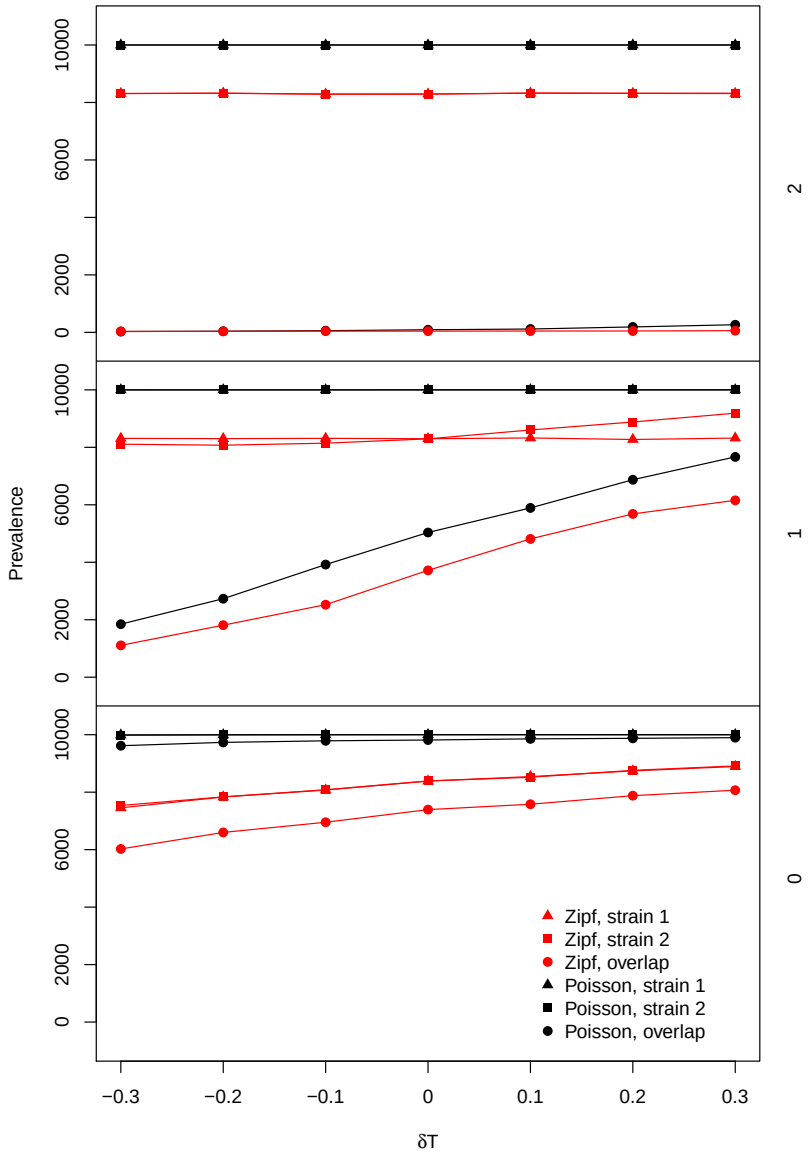


Figure 5.20: $\langle O_1 \rangle$, $\langle O_2 \rangle$ and $\langle \Omega \rangle$ when both strains modify each other's transmissibility ($T_{1|2} = T_{2|1}$), and $\tau = 0$ (bottom), 1 (middle) or $2t.u$ (top panel). $\langle k \rangle = 20$.

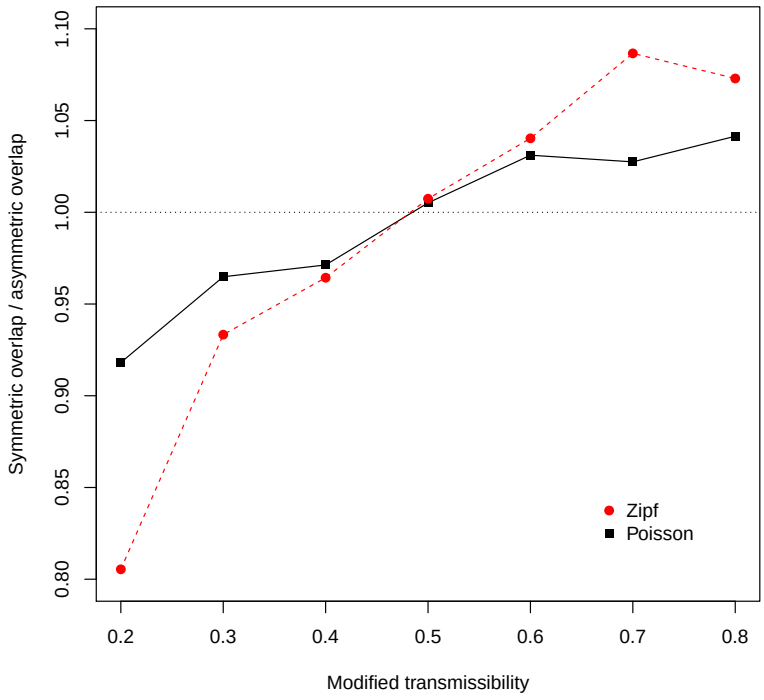


Figure 5.21: The ratio of $\langle\Omega\rangle$ under symmetric modification to $\langle\Omega\rangle$ under asymmetric modification, when $\tau = 0$, varying δT.

5.5 Summary and conclusions

Summary

In this chapter I looked at the overlap dynamics of the outbreaks of two strains that are able to enhance or hinder each other'stransmission through the host contact networks. In general, when each strain negatively modifies the other each outbreak is smaller, and the overlap between the two shrinks in size faster than either outbreak. When the modification is positive both outbreaks and overlap grow similarly.

When one strain has a modifying effect but the other does not we see a qualitatively similar but smaller effect on the size of the overlap and the outbreak of the modified strain.

When the two strains are introduced to the population at different times, the earlier strain exerts a modifying effect on the later so long as it is still present in the population by the time the later strain is introduced. The longer the delay, the smaller the effect. When the first strain modifies the second, the size of the second outbreak is affected in Zipf distributed graphs but not in Poisson distributed graphs. The overlap is affected in both graph types up to a certain delay between the outbreaks. How much delay is needed to completely remove any modifying effect depends on the graph type and mean degree. This also shows that in Zipf distributed graphs changes to the size of the overlap are due to changes in the size of the outbreak of the modified strain, whereas in Poisson distributed graphs changes in the size of the overlap are due to changes in the *speed* of the outbreak of the modified strain.

When the second strain modifies the first the effect is removed by almost any delay, since the first strain in effect spreads ahead of the spread of the modifying strain. This is true even in very sparse graphs and for outbreaks that spread poorly. In Poisson distributed graphs with $\langle k \rangle = 10$, $\tau = 0.75t.u.$ is sufficient to remove any significant modifying effect, whereas in the Zipf distributed graphs a small effect persists at least until $1.0t.u.$

Comparison to immune modification

Although I am not aware of any studies that model transmission modification explicitly, it is worthwhile comparing my results qualitatively to the case of immune modification. The most

directly comparable study is that of Marceau *et al* [115], who conducted similar simulations on both Poisson and power-law distributed graphs, where infection with one strain conveys partial immunity to a second strain. Although the results are broadly similar, there are three major differences.

When one strain conveys complete immunity to a second, the size of the outbreak of the second strain is reduced, and this effect is much more pronounced in power-law distributed graphs than in more homogeneous graphs such as the Poisson distributed graph [70, 95, 127]. This is because high-degree vertices are likely to become immunized early on in the outbreak, and because these vertices play such a significant role in the spread of disease, removing them has a disproportionately large effect.

Marceau *et al* found that when the immunising effect of one strain on the other is not complete, this relationship reverses itself, and we see less pronounced effects in power-law distributed graphs. They found that this is because while high-degree vertices always become partially immunized, each infected neighbour's attempt to infect succeeds or fails independently of the others, and so the reduction in effective susceptibility is inversely related to the degree. Hence high-degree vertices are poorly protected and still function as super-spreaders, albeit to a lesser extent.

Adapted somewhat, the same reasoning applies to transmission-modifying interactions. In this case the degree of protection experienced by a vertex depends not on whether it is itself infected with the protecting strain but rather on how many of its neighbours are infected. A vertex whose only neighbour is infected with the protecting strain, is quite well protected, a vertex with 100 neighbours of which only one is infected with the protecting strain is hardly protected at all. So once again high-degree vertices are unlikely to experience much protection unless the protecting strain has had time to spread to a large fraction of its neighbours.

We should therefore expect to see a more pronounced effect of modification on at least the overlap in Poisson distributed graphs compared to Zipf distributed graphs, but in fact the two are very similar. In the Poisson graph most vertices have a large number of neighbours, and so there are many avenues for infection to reach a given vertex. Infections being transmitted from neighbours that are not coinfected are transmitted with the unmodified probability, so in order to experience

the same degree of protection as in the immune modifying model, a vertex must be completely surrounded by vertices infected with the modifying strain. Hence the typical vertex in the Poisson distributed graph, with its relatively high degree, experiences poor protection. This is compounded by the fact that the Poisson distributed graphs used here are roughly an order of magnitude less clustered than the Zipf distributed graphs, making it less likely that a vertex's entire neighbourhood is infected with the modifying strain at a similar time.

Another significant difference between the model studied in this chapter and that of Marceau *et al* is that they only studied the case where the modifying strain was the delayed strain. They found that the effect on the spread of the other strain was significant, even with a considerable delay. In the present model on the other hand, even a fairly short delay is sufficient to almost entirely remove any effect.

This may be explained by the fact that the my simulation model uses semi-continuous time, with a smallest interval between becoming infected and transmitting infection of $0.2t.u.$, while Marceau et al use a discrete-time model where the interval is $1t.u.$ As I showed in chapter 3 discrete-time models produce slower outbreak dynamics. Indeed, when I considered very short delays, there was a significant protective effect, at least in the Zipf distributed graphs.

The fact that the effect only persists for very short delays has implications for interventions that can be modelled as transmission modification, such as quarantining symptomatic individuals. To the extent that the present work can be used as a model for the real world, it suggests that such measures must be implemented very quickly in order to be effective, and that the time frame for implementing such measures is smaller the more homogeneous and the less clustered a population is.

Unlike the works listed above, I consider the effect of a partially modifying strain introduced shortly before another strain. Similarly to the case when the modifying strain is introduced after a very short delay, the effect of the modifier depends on the degree distribution of the host population, but the effect is much more robust to increased delays.

Chapter 6

Recombination and competition

6.1 Introduction

In this chapter, I consider two kinds of interactions between strains: the generation of novel strains through recombination or reassortment, and intra-host resource competition. There are three sets of simulations. In the first two, only the recombination/reassortment dynamic is introduced. In the third, strains are additionally assumed to be in competition for space within hosts.

6.2 Recombination only - unlimited recombination

The model of recombination that I use here is very simple: whenever a vertex is infected simultaneously with two or more strains, they can give rise to a new strain that then begins to spread independently of its parents. Extending the basic model described in chapter 3 to include this mechanism is straightforward: whenever a strain infects a vertex v that is already infected with at least one other strain, a new strain is generated with probability ρ. If a new *child strain* is generated, it is assigned two *parent strains* chosen randomly from among all strains infecting the vertex. A *recombinant* event R_c is added to the queue with the same time as the infection event that caused it, so that it is the next event to be proscessed.. When the recombinant event is handled, it is treated like an infection events, causing one R event and up to k_v infection events.

The new strain is assigned its own parameter set λ, Γ, ϵ, but in this set of experiments all strains have identical parameter values. All strains spread independently. In particular, note that infection with a parent does not convey any resistance to child strains. Simulations are started with two strains, the *ancestor strains*, that are introduced simultaneously.

The recombination model described above requires that I impose a limitation on the simulations. Since each new recombinant spreads independently of all previous strains and so will add up to 25000 events in the queue, and since each new strain has the potential to generate more strains provided at least one other strain (including the parent strains) is still circulating, simulations could run for a very long time indeed, depending on the resulting dynamics of the system. To make the runtime and memory requirements of the simulation as well as the size of the output manageable, each simulation processes only the first two million events in the queue and then halts, regardless of the state of the system.

In the first set of simulations, two strains with $T = 0.5$ were initiated in Poisson and Zipf distributed graphs with $\langle k \rangle = 20$. The per-infection probability of recombination ρ was either 0.00005, 0.0001, 0.0001, 0.0002, 0.0003, 0.0004, 0.0004 or 0.0008. This range was chosen because preliminary simulations showed that much below this range simulations where a recombinant arose are exceedlingly rare, and much above it the number of recombination events grows so rapidly that the queue becomes completely filled with recombination events and the simulation in effect breaks down.

The first question I want to answer is: how many simulations have at least one recombinant? Let r_{sim} be the number of strains that arise in a simulation. Out of 1600 simulations, in 1136 (71%) $r_{sim} > 0$. In Poisson distributed graphs, 591 simulations (73%) had $r_{sim} > 0$, while in Zipf distributed graphs there were 545 (68%).

Since each coinfection event causes either one recombinant with probability ρ or none with probability $1 - \rho$, the probability of no recombinants is

$$\mathbb{P}(r_{sim} = 0) = (1 - \rho)^{\Omega_{1,2}} \tag{6.1}$$

where $\Omega_{1,2}$ is the size of the overlap between the ancestor outbreaks.

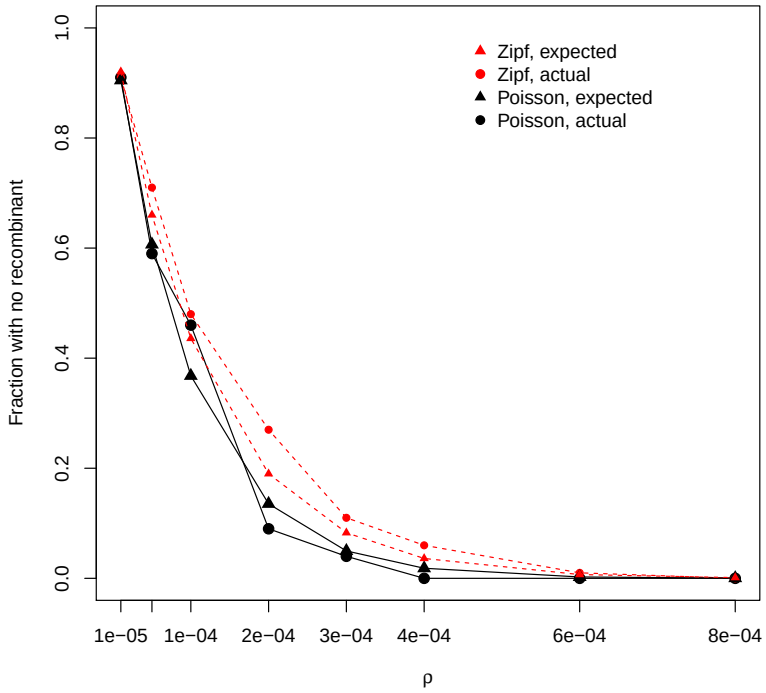


Figure 6.1: The expected rate of simulations where $r_{sim} = 0$ (triangles) and the observed rate from simulations (squares)

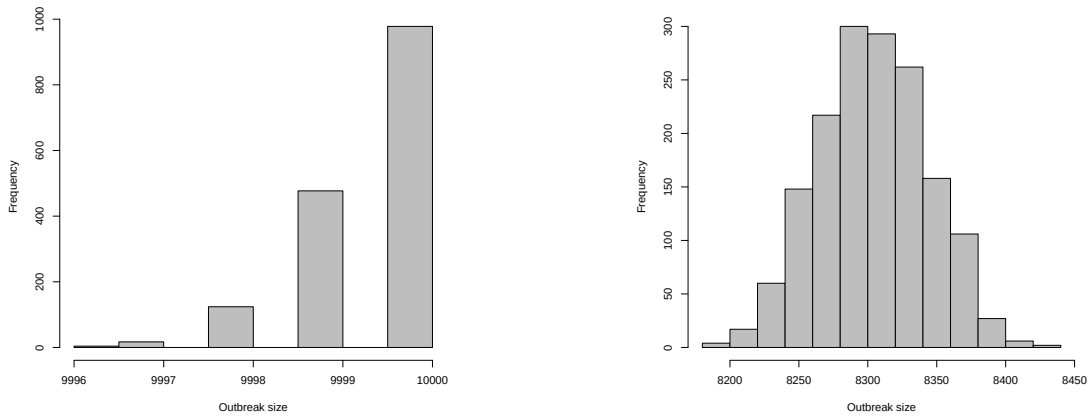
Using equation 6.2 and the mean value of $\Omega_{1,2}$ – 10000 in Poisson and 7654 in Zipf distributed graphs, one can calculate the expected fraction of simulations that generate at least one recombinant. Figure 6.1 compares the expected and observed fractions. In Zipf distributed graphs there are consistently more simulations with no recombinants than expected. In the Poisson distributed graphs the actual number of simulations with no recombinants also differs from the expectation, but not in a consistent way. In both graph types the differences are small, and are probably due to the sample size being 100. The slightly larger differences in the Zipf distributed graph are due to the larger variation in $\Omega_{1,2}$, so that $\langle\Omega_{1,2}\rangle$ is a slightly poorer predictor of actual $\Omega_{1,2}$ in a small sample.

Given that at least one recombinant arises in a simulation, how many recombinants are likely to arise? Calculating this is a little more complicated. I suggest how one might proceed in the discussion, but here I focus on measuring it. Figure 6.3 shows the distribution of r_{sim} in simulations on Poisson distributed graphs, excluding graphs where $r_{sim} = 0$. Figure 6.4 shows the same for Zipf distributed graphs.

The distributions look quite different in the two types of graph for higher values of ρ, but in fact $\langle r_{sim} \rangle$, the mean number of strains per recombination, differs by less than one standard deviation between the two graph types (figure 6.5). The difference in the histograms of the distributions is due to a small number of simulations in the Zipf distributed graphs with a very small nonzero number of recombinants. For larger values of ρ these are completely absent from the Poisson distributed graphs.

The reason that simulations with a small number of recombinants occur in Zipf distributed graphs even when ρ is higher is likely a combination of the fact that overlaps are smaller in Zipf distributed graphs - the sizes of the ancestral outbreaks are given in figure 6.2; that some strains will arise in very low-degree vertices, in which case they may simply fail to infect any of their neighbours; and that some recombination events will occur late in the overlap of their parent strains, which reduces the number of strains available for them to recombine with. Even with a higher value of ρ it is possible that the ancestor strains produce only a few children, and if these happen to occur in low-degree vertices and then fail to spread, or if they appear too late, the system will still go extinct. This is supported by the existence of a few simulations wher $r_{sim} = 15$. This number suggests that, if ρ is sufficiently small, even a system where a fairly large number of strains arose initially is vulnerable to failing due to a series of "late" recombinations.

The very large number of strains that arise when ρ is high suggests that the effect of the cut-off at two million events on $O_i, i \in \mathbb{Z}$ might be quite severe: $O_A \approx 10000$ in the Poisson distributed graphs, and 8300 in the Zipf distributed graphs. With 1000 strains in the simulation, there are 2000 events available per simulation, which implies that $\langle O_i \rangle < 0.25 O_A$, so a large fraction of outbreaks will fail to reach their full size. I would like to know how distributed this problem is. Clearly, strains whose outbreaks are seriously curtailed by the cut-off cannot be considered for most of my analysis. If almost all strains experience some effect of the cut-off, then some adjustment of the simulation



(a) Poisson

(b) Zipf

Figure 6.2: Distribution of ancestral outbreak sizes, $T = 0.5$, $\langle k \rangle = 20$.

method is necessary. If on the other hand the first n strains are almost or entirely unaffected, with the latter strains having extremely small outbreaks, then it might be an acceptable approximation to simply disregard heavily affected strains.

Figure 6.6 shows the $\langle O_i \rangle$ as a function of strain id. The strain id is a unique identifier for each strain in a simulation; they are assigned in the order that the strains arise, so that strain 10 is the 10th strain to arise. In the Poisson distributed graphs the first 50 or so strains are unaffected, whereas in the Zipf graph all but the very first handful of strains are affected. In order to be able to compare the two graph types in subsequent simulations, I therefore need to adjust the simulation procedure.

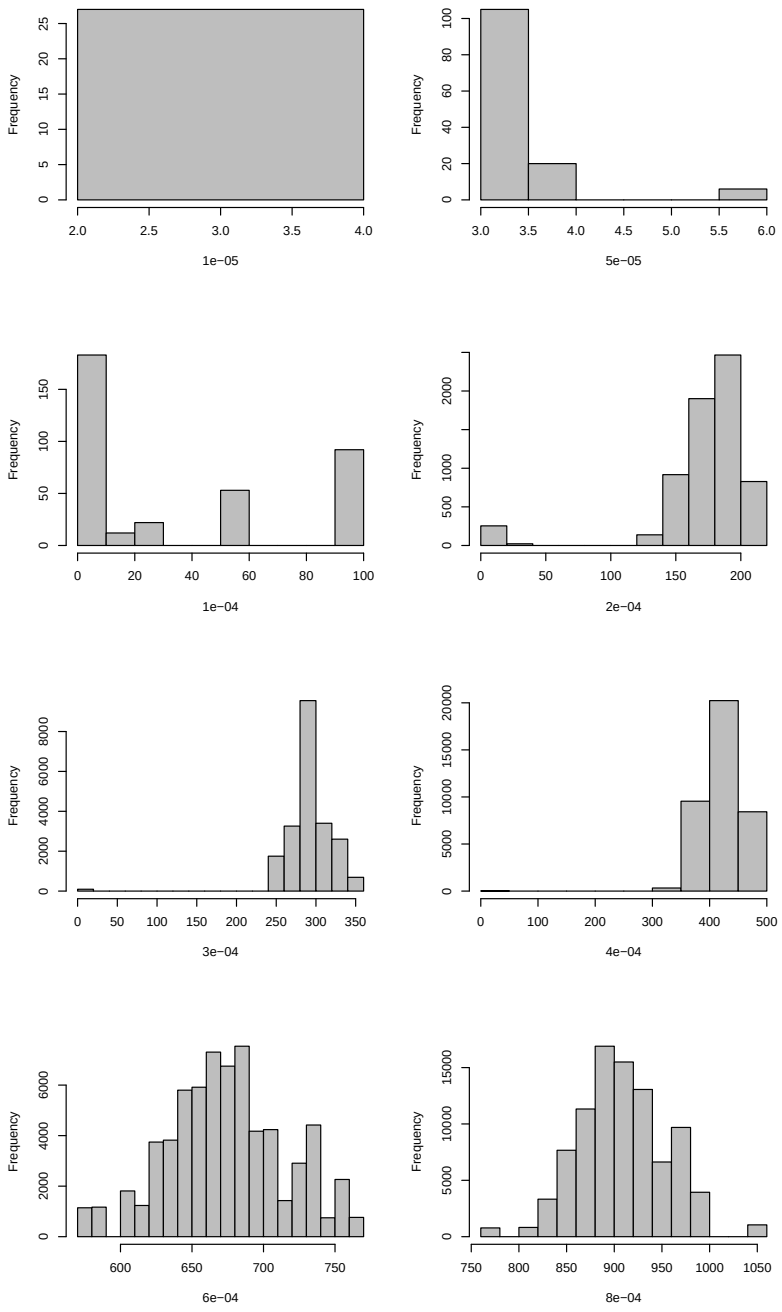


Figure 6.3: The distribution of r_{sim} in simulations where $r_{sim} > 0$ in Poisson distributed graphs, for different values of ρ.

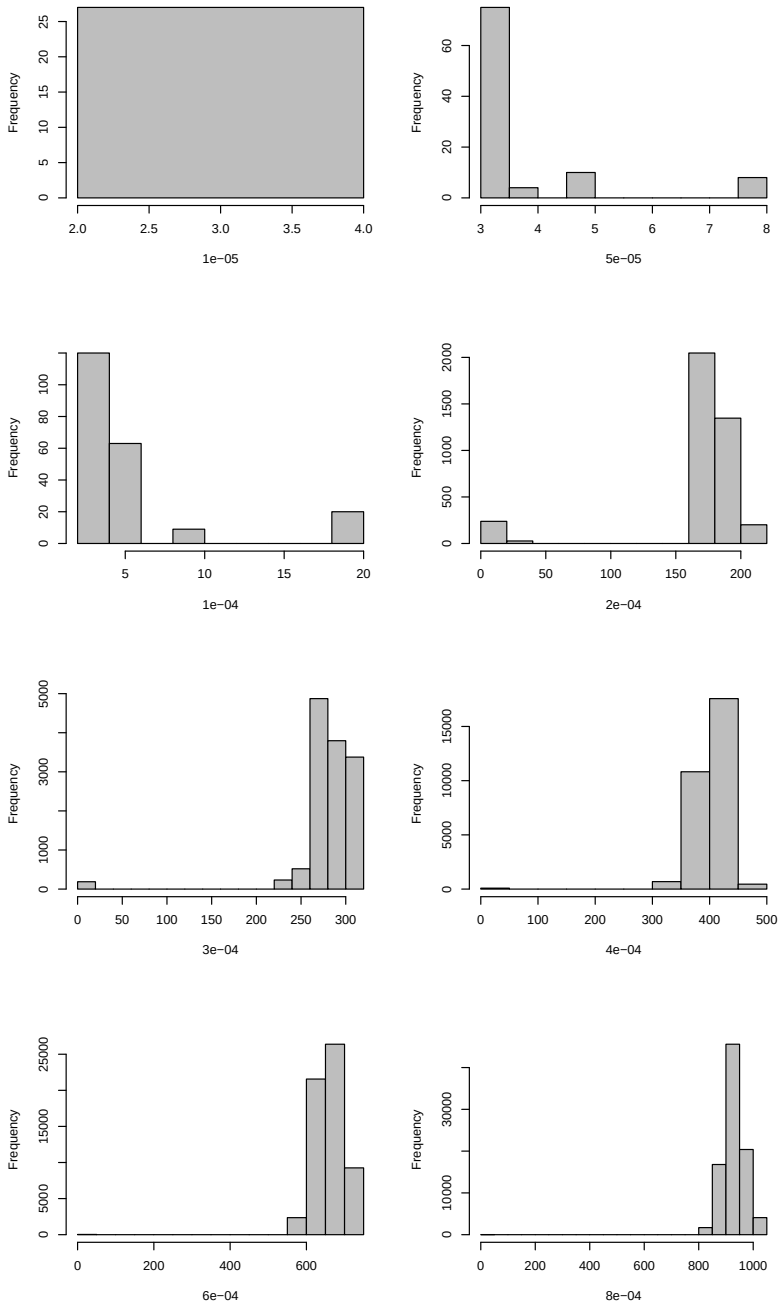


Figure 6.4: The distribution of r_{sim} in simulations where $r_{sim} > 0$ in Zipf distributed graphs, for different values of ρ.

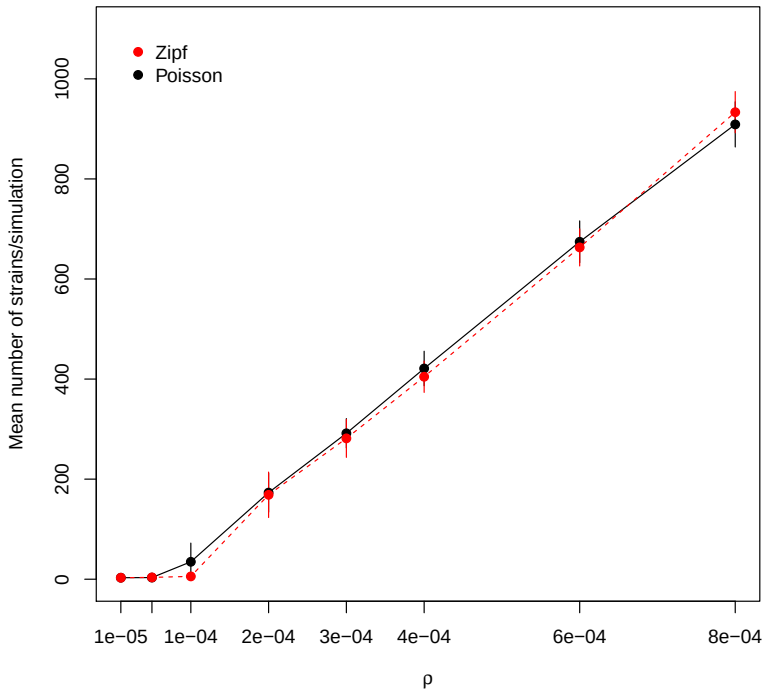


Figure 6.5: $\langle r_{sim} \rangle$ for varying values of ρ in Poisson (solid black line) and Zipf distributed graphs (dotted red line).

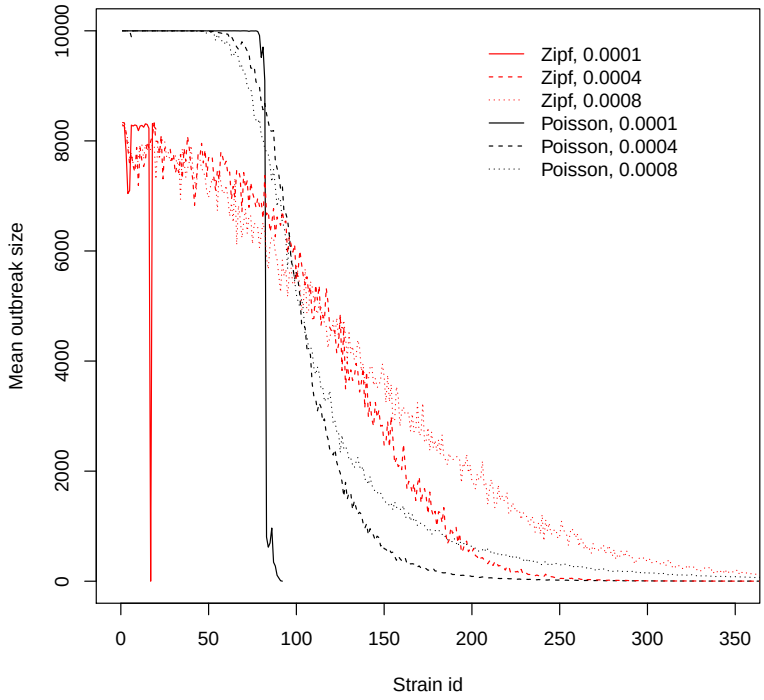


Figure 6.6: $\langle O_i \rangle$ as a function of strain id for for simulations in Poisson (black lines) and Zipf distributed graphs (red lines), with different values of ρ.

6.3 Recombination only - limited recombination

For the values of ρ I consider here it does not appear likely that r_{sim} will be bounded, or if so that the number of strains is too large to accurately simulate the process, even with far greater computational resources. Even simply allowing the strains that have arisen in the first two million events to go extinct would take approximately 25 million events, and so extending the maximum queue length is not an adequate strategy. Instead, I have chosen to impose a limit on the number of strains that can arise. As mentioned in chapter 3, the effective maximum number of steps per strain is about 25000. Allowing for some variation in the number of events required by each strain, and the fact that recombinations are also events, this suggests a maximum number of strains per simulation $r_{max} = 75$. In all subsequent simulations this limit is in effect.

Because I am interested in the dynamics of recombinant strains, for subsequent simulations $\rho = 0.4$, 0.6 or 0.8. This is because for $\rho < 0.4$ a significant fraction of simulations produce no recombinants. All other parameter settings are as in the previous set of simulations.

When we restrict ρ to this higher range, there are very few simulations with $r_{sim} = 0$: 3 in Poisson distributed graphs (1%), and 7 in Zipf distributed graphs (2.5%). In 547 of the remaining 590 (92.7%), $r_{sim} = r_{max}$ (figure 6.7). Of the simulations where $0 < r_{sim} < r_{max}$, in 23 (55%) $r_{sim} = 1$, and there was only one simulation where $r_{sim} > 4$ (in this simulation $r_{sim} = 15$). Figure 6.8 shows when each strain arose in each simulation. Simulations with $r_{sim} < 15$ (red) cluster on the left, indicating that the time between recombinants is quite long in these cases, and that this is a likely driving factor behind the low number of strains in these simulations. We also see that although there is a lot of variation in how fast the first 20 or so strains arise, thereafter the remaining strains all arise very quickly.

Since all strains that arise now are able to reach their full outbreak size O_i and go completely extinct, one might expect that $O_i \approx O_A$ for all strains i. Figure 6.9 shows the distribution of O_i for all strains in Poisson distributed graphs (left panels) and all strains i in Zipf distributed graphs where $O_i > 8000$. The distributions are as expected from the ancestor strains. However, in the Zipf distributed graphs a small fraction of recombinant strains had $O_i \ll O_A$. With only a handful of exceptions, in these strains $O_i = 1$, i.e. only index case was ever infected and there was no

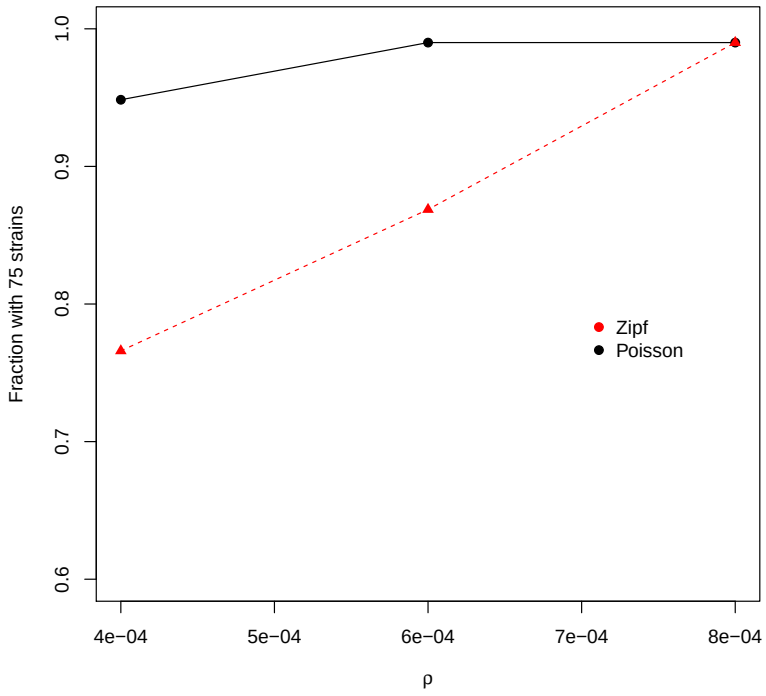


Figure 6.7: The fraction of simulations in which $r_{sim} = r_{max}$ (75) as a function of ρ, in Poisson (solid black line) and Zipf distributed graphs (dotted red line).

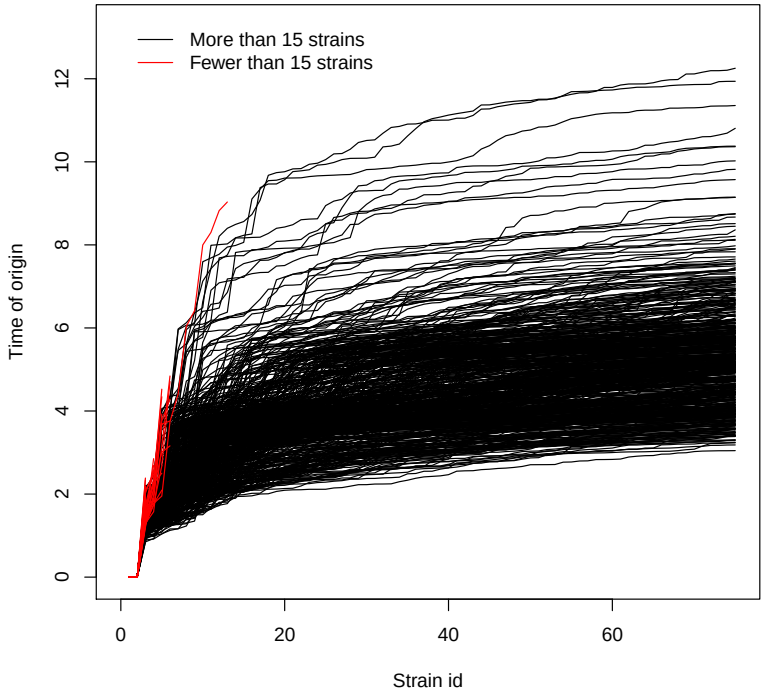


Figure 6.8: Time of origin of each strain. Each line represents a separate simulation.

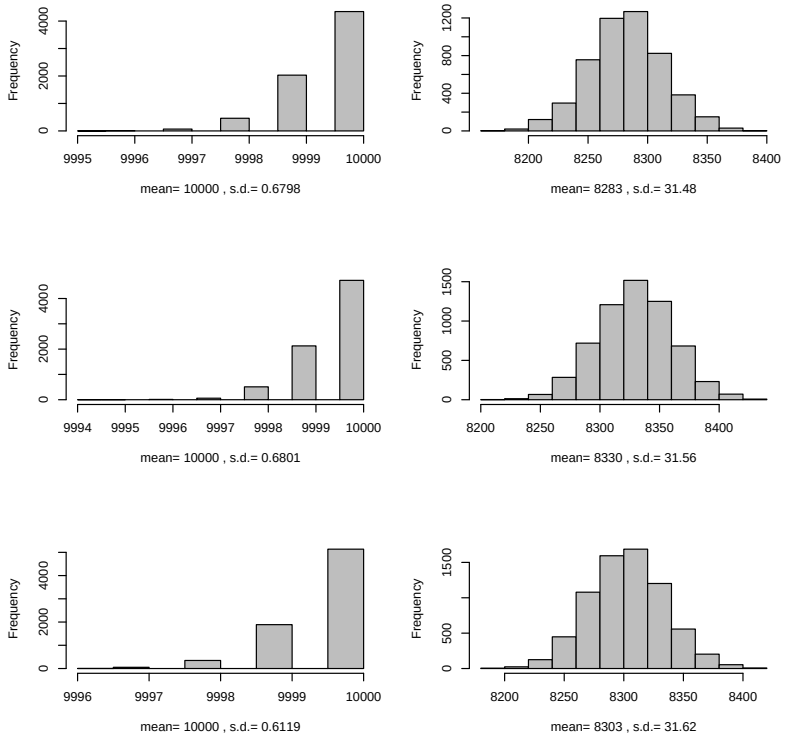


Figure 6.9: The distribution of O_i in Poisson distributed graphs (left panels) and in Zipf distributed graphs (right panels), for ρ =0.0004 (top), 0.0006 (middle) and 0.0008 (bottom panels). In the Zipf distributed graphs, strains with $O_i < 8000$ are not shown. Ih these cases, $O_i \leq 3$, and such cases comprised only 6.8% of strains.

transmission at all. The few that did spread infected at most two other vertices.

These failed recombinants all occur in low-degree vertices. When a strain arises in a vertex with degree k, the probability that it will fail to infect any other vertices is

$$\mathbb{P}(O_i = 1) = (1 - T)^k \tag{6.2}$$

which decays geometrically with increased k. Figure 6.10 shows the fraction of strains where $O_i = 1$ as a function of the degree of the strain's index vertex (solid red circles), along with the expected value (open black circles). The greatest index vertex degee of any of these strains was 8. It is

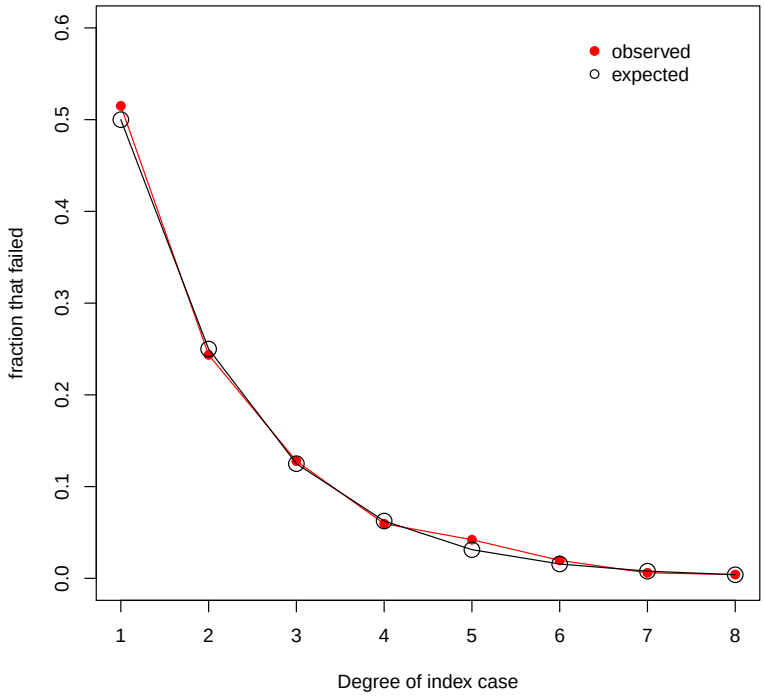


Figure 6.10: The expected (black open circles) and observed fraction (solid red circles) of strains that failed to spread as a function of index degree.

clear that the strains that failed to spread are entirely explained by the expected rate of failure in low-degree vertices.

It should be noted that while low-degree vertices make up the vast majority of the Zipf distributed graphs, the overlap, and hence recombination, occur preferentially in higher-degree vertices. In the Zipf distributed graphs considered here, with $k_{max} = 150$, $\langle k \rangle = 20$, and $N = 10^5$, vertices with degree less thn 9 comprise 58.9% of the graph, but the failed strains make up only 6.8% of all recombinant strains.

Given the disproportionate role that high-degree vertices play in the spread of disease in power-law distributed graphs in general, and their tendency to be over-represented in the overlap between

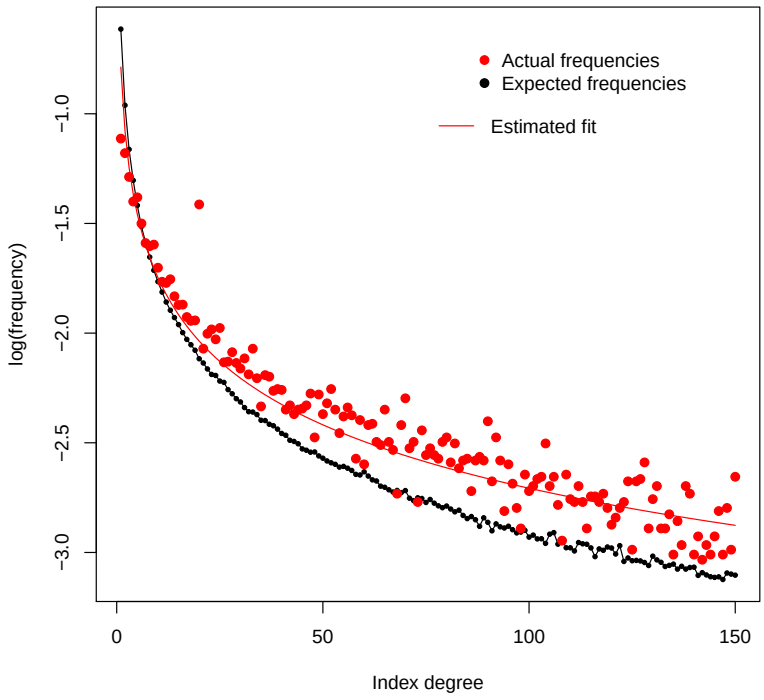


Figure 6.11: The frequency at which vertices of each degree generated recombinants (red points), compared to the frequency of such vertices in the graph (black points, line). The best fit Zipf distribution is shown ($k_{max} = 150, \alpha = 0.96$, but this is only included to indicate the trend in the data, as the fit is not very good (KS two-sample test $D = 0.11333$, $p = 0.2904$). Note the log scale on the y-axis.

outbreaks, it is reasonable to expect that they will also be correspondingly important in the generation of recombinant strains. Figure 6.11 shows the frequency at which vertices with a given degree produced recombinants, compared to the frequency of such vertices in the graph as a whole. The frequencies are given on a log scale. As expected, high-degree vertices produce far more recombinants than their frequency in the graph suggests, up to five or six times as many in some cases. Vertices with degree 20 are also very much over-represented, but this is because all ancestor strains are chosen from among vertices with degree 20, as discussed in chapter 3.

6.4 Recombination and competition

In the range of values of ρ which I have studied so far, this simple model of recombination exhibits one of two behaviours: either sustained growth of the number of circulating recombinant strains, or the eventual extinction of all strains. Several real-world viruses, such as influenza A, exhibit a very different dynamic – annual seasonal outbreaks of different strains, sustained over decades, but without a pronounced increase in the overall number of circulating strains, and with period shifts from one strain being dominant to another [150].

There are many possible reasons why diseases in practice might exhibit more complex dynamics than those shown here, which I will discuss in some more detail in the conclusion of this chapter. One possible reason is that each individual host is a limited resource, so that when there are multiple strains present in the same host, they are in competition for resources, with the losers failing to thrive, and either going extinct in the host or persisting at such low concentration that their ability to be transmitted to other hosts is effectively nil. In this section I introduce a simple model of intra-host resource competition and examine the interaction between recombination and competition.

The competition model I use is very simple: each host is capable of sustaining μ strains simultaneously. If a host is infected with μ strains and another strain either arises through recombination or successfully infects from a neighbouring vertex, then one of the previously present strains is chosen uniformly at random and removed immediately. The removal of a strain is treated identically to the host recovering from infection with that strain, so that the host can never more either become infected or infect others with that strain. μ is the same for all vertices.

This model was run on Poisson and Zipf distributed graphs with the same parameters as the simulations in the previous section, including the overall limit of 75 strains. This is so that any effect of competition on outbreak size is not conflated with the effects of the cut-off at two million events. Two different values for μ, 5 and 10, were considered.

We would not expect the addition of competition to affect the fraction of simulations in which $r_{sim} > 0$, but there is some difference from the simulations in the previous section. In Poisson

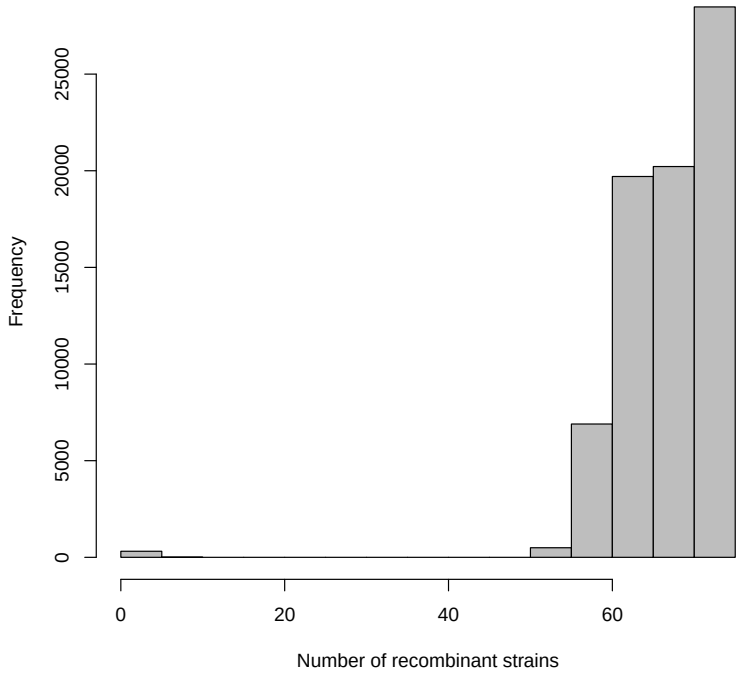


Figure 6.12: The distribution of r_{sim}, excluding simulations where $r_{sim} = 0$.

distributed graphs, two simulations (0.33%) had no recombinants, whereas in Zipf distributed graphs there were 15 (2.5%). Although the difference between graph types is more pronounced than previously, it is so small that it most likely due to chance.

Figure 6.12 shows the distribution of r_{sim} in all simulations for this section, discarding those where $r_{sim} = 0$. As before, simulations fall into two distinct groups: those where $r_{sim} \leq 15$, and those where $r_{sim} \geq 53$. The simulations with few strains comprise 26% of all simulations, which is similar to the case without competition, suggesting that whether an outbreak experiences few or many recombinations is not significantly affected by competition.

In the absence of competition, r_{sim} was either less than 16, or $r_{sim} = r_{max}$. When strains are competing for hosts, we start to see a third group of simulations in which $50 < r_{sim} < r_{max}$. Figure

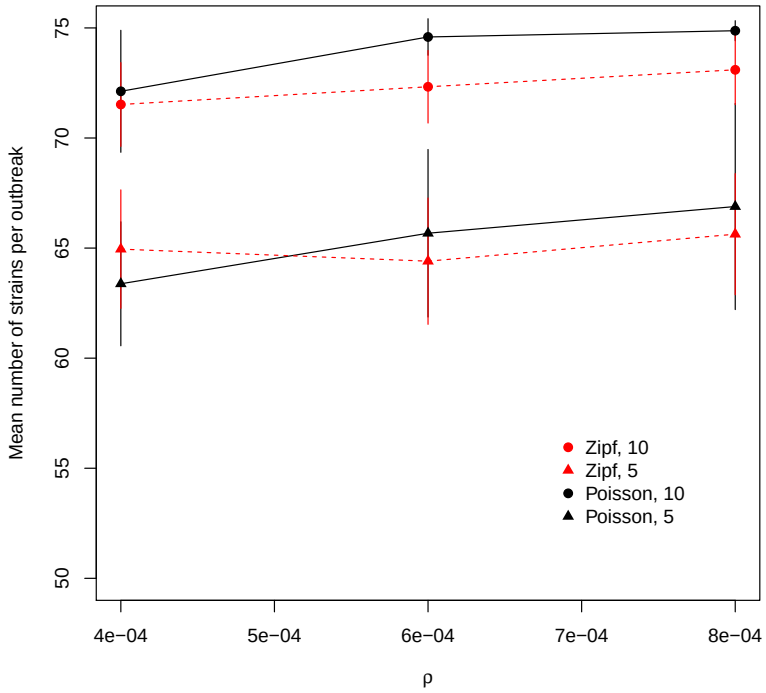


Figure 6.13: $\langle r_{sim} \rangle$, for different ρ, graph type and μ.

6.13 shows how $\langle r_{sim} \rangle$ depends on the graph type, ρ and the intensity of competition, indicated by μ. Neither graph type or ρ have a pronounced effect on $\langle r_{sim} \rangle$, but increasing competition reduces $\langle r_{sim} \rangle$ by 7 to 9 strains per simulation.

Figure 6.14 shows the distribution of O_i for Poisson (left panel) and Zipf distributed graphs (right panel). In the absence of competition the distributions were very different in Poisson and Zipf distributed graphs, but now they are more similar. In both cases the distribution can be divided roughly into three groups: $O_i < 100$, $1000 < O_i < \langle O_A \rangle$, and $O_i \approx \langle O_A \rangle$.

Intensifying the competition by reducing μ from 10 to 5 does not change this relationship, but the fraction of strains where $O_i \approx \langle O_A \rangle$ is much smaller, and the fraction where $O_i < 100$ is much larger. Perhaps the most interesting thing is the large number of strains where $O_i < 100$ in Poisson

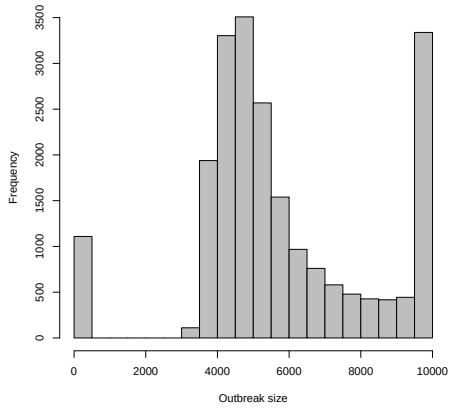
distributed graphs, which did not occur in the absence of competition. Among those strains that did not spread well, in an overwhelming majority (83%) of cases $O_i = 1$. In only 2.5% was $O_i > 2$, and $O_i > 5$ in fewer than 1% of simulations.

Figure 6.15 shows both observed and expected $\mathbb{P}(O_i = 1)$, as a function of the index degree, for simulations on Zipf distributed graphs. For legibility it is split into two panels, showing low index degree (less than 15) and high (greater than 14) respectively. In vertices that have sufficiently low degree we expect some fraction of strains to fail (see figure 6.10), and in the absence of competition, the observed $\mathbb{P}(O_i = 1)$ was exactly as predicted by equation 6.3. Under either level of competition the overall $\mathbb{P}(O_i = 1)$ (solid lines), is much higher than the expected level (dotted line). For strains that arise in vertices with degree 15 or greater, for which the expected fail rate is essentially zero, the degree of the vertex seems to have very little effect on the rate of failure. For strains that arose in vertices with degree 14 or less the probability of failure is as expected, plus a fairly constant amount that depends on μ.

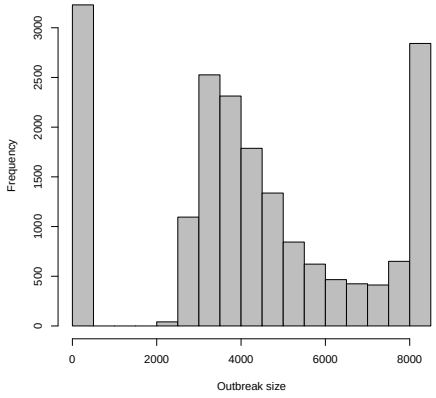
In Poisson distributed graphs $\mathbb{P}(O_i = 1)$ is essentially zero. Introducing competition causes a significant fraction of strains to fail to spread, and intensifying the competition by changing μ from 10 to 5 roughly quadruples this fraction (figure 6.14).

It is clear that competition for resources can prevent recombinant strains that arise from spreading beyond their index vertex. This suggests that the competition is quite severe, even when μ=10, because for this to happen, the recombinant strain has to be removed before the first of its infection events can take place. In a vertex with k neighbours, the time of the first transmission of a recombinant strain is the first order statistic of k exponentially distributed random variables. When k is high, this number will be very small, and so the replacement has a very small window in which to prevent the spread of the new strain to even a single other vertex.

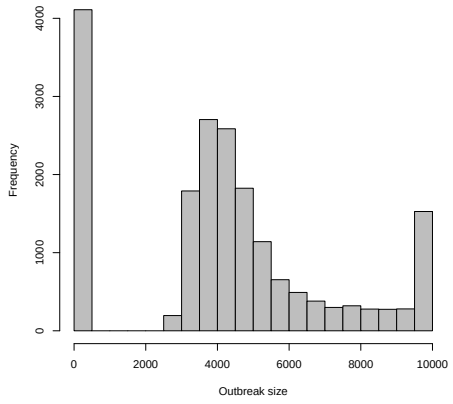
Figure 6.16 shows $\langle \frac{O_i}{O_A} \rangle$ sorted by strain id for different graph types and μ, with ρ=0.0004. The figure shows $\frac{O_i}{O_A}$ instead of O_i in to make comparison between the two graph types easier. As more and more strains arise, competition quickly reduces $\langle O_i \rangle$. At the later stages of the simulation, strains that arose early begin to go extinct, and the pressure of competition eases, so that for the last few strains, $\langle O_i \rangle$ begins to rise again.



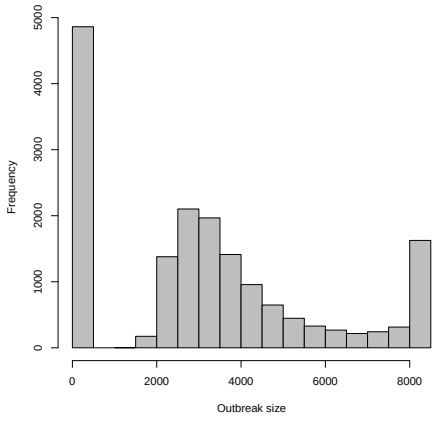
(a) Poisson, $\mu = 10$



(b) Zipf, $\mu = 10$

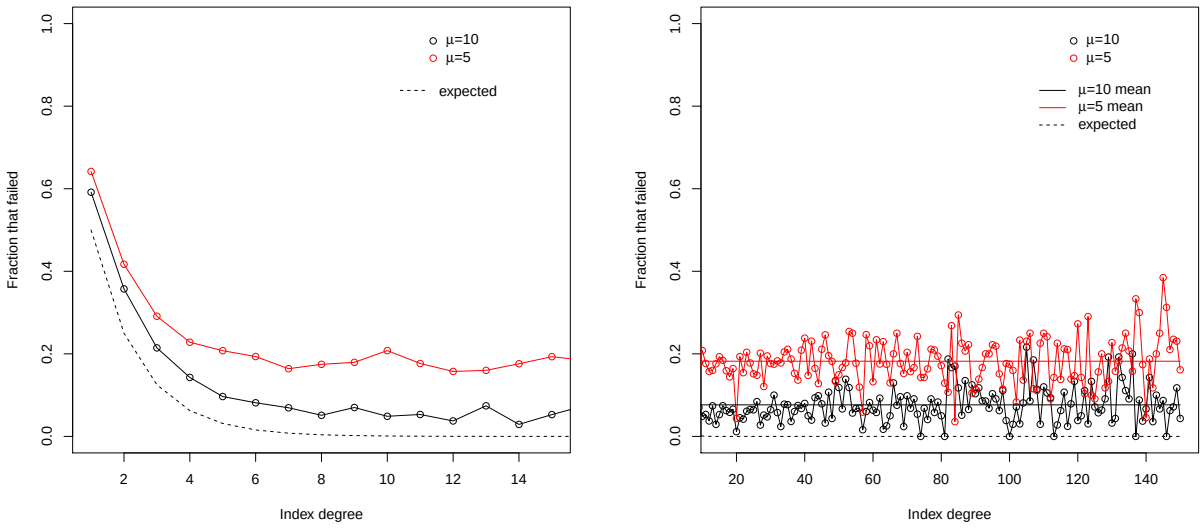


(c) Poisson, $\mu = 5$



(d) Zipf, $\mu = 5$

Figure 6.14: Distribution of O_i under different intensities of resource competition.



(a) Poisson

(b) Zipf

Figure 6.15: $\mathbb{P}(O_i = 1)$ as a function of the degree of their index vertex in Zipf distributed graphs. The rate predicted by the expression $\mathbb{P}(O_i = 1) = (1 - T)^k$ is shown as a dotted black line.

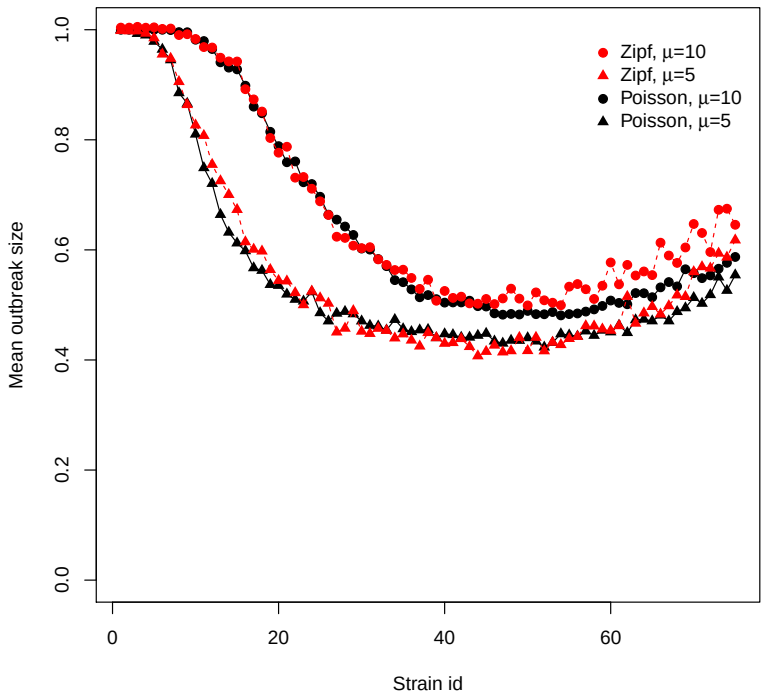


Figure 6.16: $\langle \frac{O_i}{O_A} \rangle$ a function of strain id. ρ =0.0004.

Some of this late-stage increase in outbreak size may be an artefact of the limit $r_{max} = 75$ - if in the absence of the limit more than 75 strains would have arisen, the competition experienced by the strains in the later stages of the simulation is artificially light. This can only be the case in simulations where 75 strains were generated, because if fewer than 75 strains arose, no strains have been prevented from arising due to the cap. Figure 6.19 is a repeat of figure 6.16, excluding simulations where $r_{sim} = r_{max}$. The two figures are extremely similar, so that there is very little reason to expect that the late-stage competition is being severly underrepresented in the data. Figures 6.17 and 6.18 show the same information as figure 6.16 for simulations in which ρ=0.0006 and 0.0008 respectively. These two figures include simulations $r_{sim} = r_{max}$, since removing them also has the effect of reducing sample sizes, and removing these strains did not alter the sizes of late-strain outbreaks for these simulations either (figures not shown).

From figures 6.16-6.18 it is clear that changes in both ρ and μ cause changes to $\langle O_i \rangle$ as well as the number of strains that experience competition. The behaviour is all but indistinguishable between the two graph types. This is somewhat surprising. We might expect that strains are more likely to experience competition in high-degree vertices, and given the crucial role that high-degree vertices often play in the spread of disease, we might expect that a strain being evicted from a high-degree vertex would disproportionately affect the size of its outbreak.

One reason why this is not the case is likely that while strains experience increased competition in high-degree vertices, the opposite is true in low-degree vertices, so that these vertices in effect act as a haven where the strains can persist even when competition is severe. In contrast, there are far fewer low- or high-degree vertices in the Poisson distributed graphs, so that the level of competition a strain is exposed to varies much less.

The final question I would like to consider is the *disease burden* of the population, defined as

$$\frac{1}{N} \sum_i O_i \tag{6.3}$$

In the absence of competition, for sufficiently high ρ r_{sim} grows, apparently without limit, and so each new strains causes more infection in the population. When the strains compete for space in the hosts, this competition tends to drive down the size of the each outbreak, and many strains fail to spread altogether, so that it is reasonable to ask how much this competition lightens the

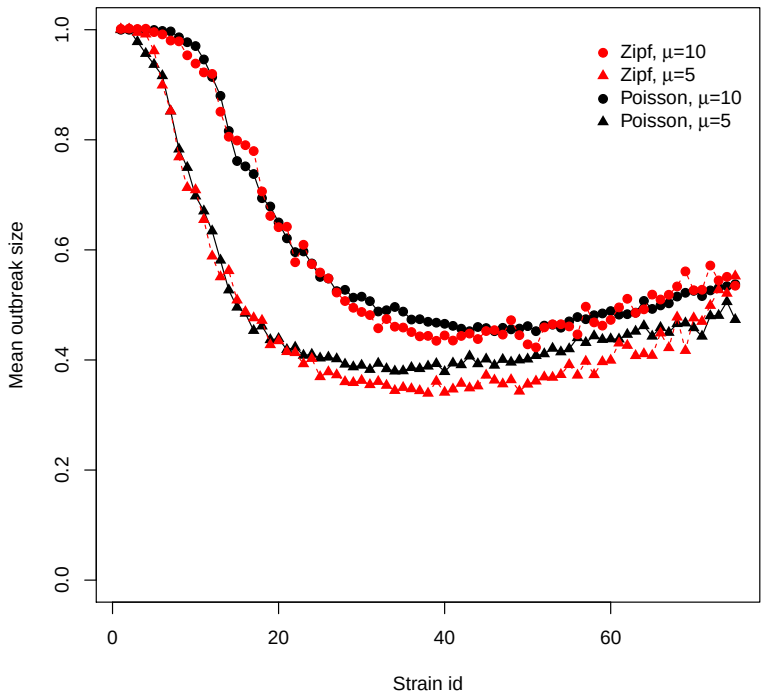


Figure 6.17: $\langle \frac{O_i}{O_A} \rangle$ a function of strain id. $\rho = 0.0006$

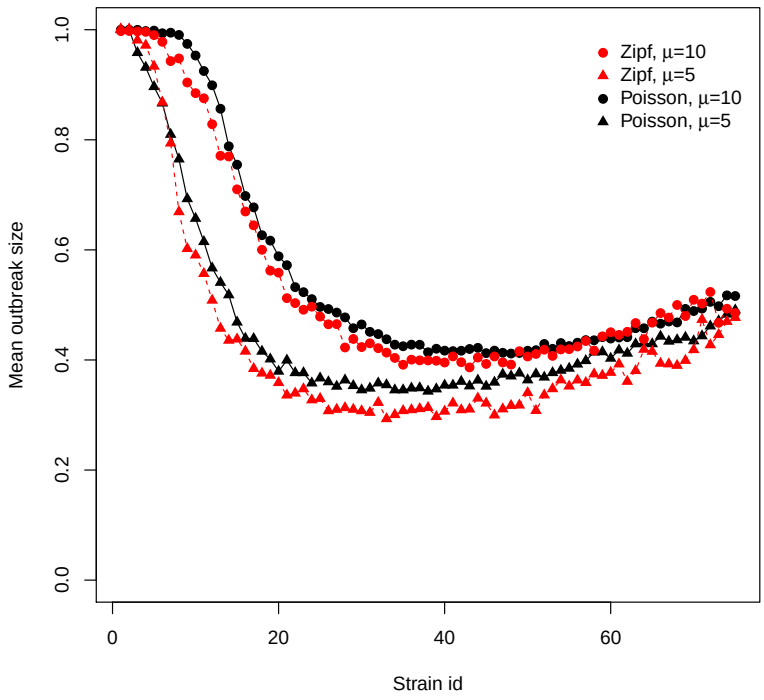


Figure 6.18: $\langle \frac{O_i}{O_A} \rangle$ a function of strain id. $\rho = 0.0008$

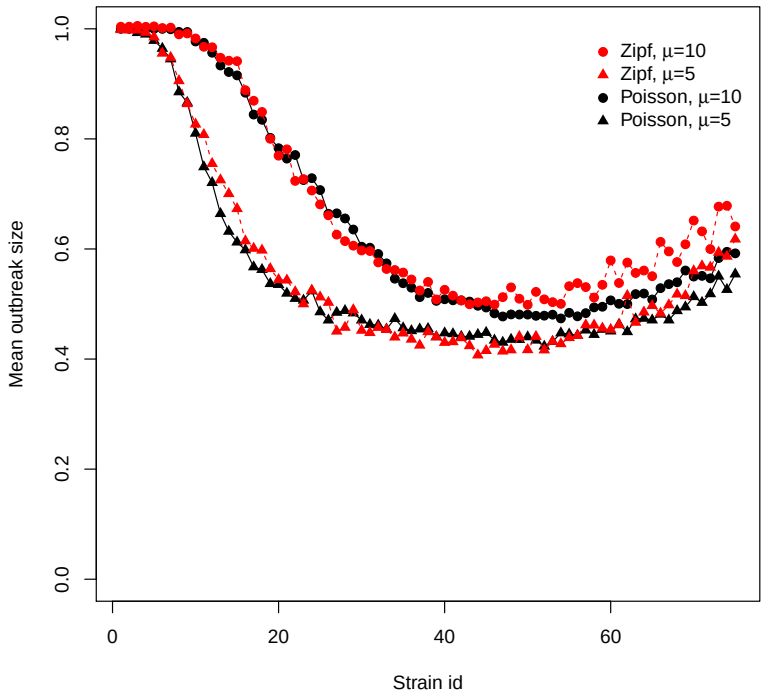


Figure 6.19: $\langle \frac{O_i}{O_A} \rangle$ a function of strain id. $\rho = 0.0004$. All simulations where $r_{sim} = r_{max}$ have been excluded.

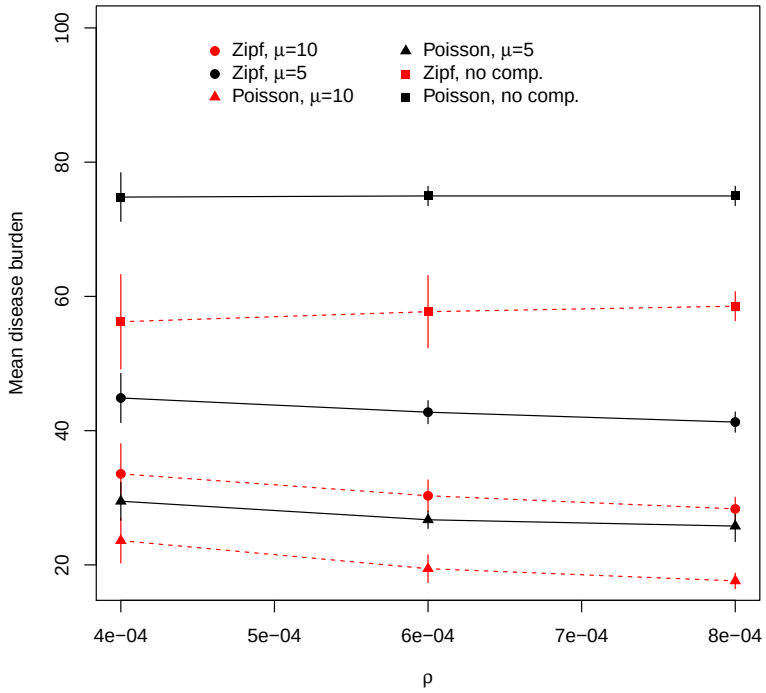


Figure 6.20: Mean disease burden

overall disease burden. Figure 6.20 shows the mean disease burden, varying graph type, ρ and μ. When strains compete for space within the hosts, the overall disease burden in the population is dramatically lowered. In the absence of competition, the disease burden in the Poisson distributed graphs is approximately 74 infections per vertex, almost the theoretical maximum of 75. In Zipf distributed graphs, the larger proportion of simulations with no or few recombinants, as well as the fraction of strains that failed to spread, drives the disease burden down from the maximal 63 infections per vertex to 58. Changing ρ has very little effect on the disease burden, but changing μ has a very large effect. Disease burden is also much greater in Poisson distributed graphs, because each outbreak tends to be significantly larger, but the effect of competition on disease burden is comparable.

6.5 Discussion

Summary

In this chapter, I introduced two additional mechanisms to the basic model. The first mechanism is recombination: two strains coinfecting a vertex give rise to a new strain with probability ρ. The new strain then spreads independently. Due to computational limitations, only the first two million events of the queue were handled in each simulation. In this system, I found that when ρ is sufficiently low almost no recombinants arise, and the system dies out very quickly. When ρ is high, the system almost never dies out within the first two million events, and appears to be still adding recombinants at an increasing rate. The limit on the number of events that could be processed severely affects the accuracy of the results for the strains that arose even quite early on in the simulation. To compensate for this I restricted subsequent simulations to the first 75 strains.

With this cap in place, two major differences between the recombination in Poisson distributed graphs and Zipf distributed graphs became apparent. In Zipf distributed graphs, recombinant strains were significantly more likely to fail to spread. This was found to be entirely due to such strains arising in low-degree vertices. Overall, these failed strains were relatively few, because the overlap between strains, and therefore recombination, preferentially occurs in high-degree vertices, which were found to play a disproportionately large role in the generating recombinant strains. In the Poisson distributed graphs, these effects are entirely absent - recombinant strains never failed to spread.

The second new mechanism I introduced was resource competition among strains - once a vertex was infected with a maximum number μ of strains, a subsequent infection or recombination would cause one of the previously present strains to be replaced. Competition reduced the size of most outbreaks dramatically, and this effect was identical in both graph types when measured relative to the expected outbreak size in the absence of competition. This significantly increased the chances of a recombinant strain failing, and failures were seen in Poisson distributed graphs, at lower but comparable levels to those in Zipf distributed graphs. The reduction in outbreak size and increased chance of failure had the effect of reducing recombination, so that in many simulations the system went extinct without having generated all 75 permitted strains. Finally, competition dramatically

reduces the overall disease burden on the population.

Long-term behaviour of the recombination dynamic

The computational limit of two million events per simulation means that I did not explore the long-term behaviour of the system with recombination. The limit is not absolutely fixed, and could be relaxed. However, the sheer number of strains that arose even in this relatively small early window suggests that simply increasing the number of events handled is unlikely to reveal much more, even if the limit can be extended by many orders of magnitude. An alternative approach might be to attempt to obtain analytical results, for example for the expectation or the distribution of the number of strains generated. The number of recombinants $r_{1,2}$ between the ancestor strains 1 and 2 is a binomially distributed random variable $B(\Omega_{1,2}, \rho)$:

$$\mathbb{P}(r_{1,2} = i) = \binom{\Omega_{1,2}}{i} \rho^i (1-\rho)^{\Omega_{1,2}-i} \tag{6.4}$$

From this one might try to estimate the expected total number of recombinants r_a generated by a given strain a as the sum of the expected number of recombinants between a and every other strain present, giving

$$\mathbb{E}(r_a) = \rho \sum_j \Omega_{a,j} \tag{6.5}$$

In the present model, recombination is tied to infection so that each infection generates at most one recombination, and so the contributions of each overlap $\Omega_{a,j}$ are not additive. Instead, we must use the union of all overlaps Ω_a:

$$\Omega_a = \bigcup_j \Omega_{a,j} \tag{6.6}$$

summed over all other strains j present, so that equation 6.5 becomes

$$\mathbb{E}(r_a) = \rho \Omega_a \tag{6.7}$$

We could then calculate the number of recombinant strains over time as a continuous-time branching process, where the number of offspring of a given strain a is binomially distributed. To obtain the distribution of the times at which the offspring strains arise, we normalise the "prevalence" curve of Ω_a. The size of Ω_a is not straightforward to calculate, and obtaining a general expression for its evolution in time is even more complex. However, were these complications to be surmounted, there still remains the problem that generations are not distinct - a strain can recombine with its

own offspring. I do not attempt to derive an expression for this, but we can still speculate about the long-term behaviour of the system.

Calculating the exact Ω_a that each strain a generates is a non-trivial problem, but it must at least be bounded above by N, the size of the population. If there are very few strains arising, either because ρ is very low, or the system is in its earliest phase, the probability that strains arise sufficiently far apart in time that they have almost no overlap becomes high - a strain that arises from the very last coinfection between its parent strains is unlikely to be able to recombine with either of them. If this is the case, the system is likely to die out.

However, if we assume that the system survives until some point where there strains are arising quite close to each other in time, then the size of O_A is (approximately) bounded below by the smallest outbreak size of all other strains present. Since outbreaks sizes for independently spreading identical strains do not vary very much, we can argue that there exists some typical value of O_{typ} that is representative of most strains at this stage in the process. Then the expected number of recombinant strains generated by a strain a is approximately

$$\mathbb{E}(r_a) = \rho\Omega^{typical} \tag{6.8}$$

. We see that if $\mathbb{E}(r_a) < 1$, new strains are arriving in the population too slowly to replace the extant strains, and eventually the system reaches a diseases-free situation. If $\mathbb{E}(r_a) \gg 1$, then new strains are arising much faster than their parents are dying, and the number of strains will simply continue to grow indefinitely. When $\mathbb{E}(r_a) \approx 1$, the behaviour is less clear cut. It seems likely that in this scenario the system is susceptible to stochastic fluctuations, so that it will eventually die out due to several strains in a row for which r happens to be 0. Of course, we have that

$$\mathbb{P}(r_a = 0) = (1-\rho)^{\Omega_a} \tag{6.9}$$

which shrinks very fast as the population size, and hence Ω_a, grows. So for larger populations, the behaviour when $\mathbb{E}(r_a) \approx 1$ may be more steady, but it is not clear to me whether there exists a steady state in which the number of strains remains fairly constant and nonzero indefinitely, or whether, analogously with R_0, $r_a = 1$ denotes a critical threshold between sustained growth and certain extinction. The latter seems likely, given that the system is similar to a branching process. In particular, it is striking that this system is essentially the Reed-Frost model, but with

the complication of calculating Ω. Reed-Frost models and branching processes in general do exhibit critical thresholds, suggesting that this is what we should expect. Determining this threshold, or the behaviour of the system close to the threhsold, is outwith the scope of this thesis.

Long-term behaviour with competition

When competition is introduced, a significant fraction of simulations do not generate 75 strains. In these cases, the effect of competition is to drive the system extinct. Before μ strains have arisen there is no competition, and even after this point some time will elapse before most vertices are infected with μ strains. The system is therefore usually able to sustain itself beyond the first 50 strains. However, as competition intensifies, outbreak sizes shrink. This necessarily reduces Ω_a, so that the expected number of strains generated by a "typical" strain, $\mathbb{E}(r_a)$, also shrinks. In the cases where the system goes extinct, $\mathbb{E}(r_a)$ has clearly dropped below 1.

What is not so clear is what happens in the cases where all 75 strains arise. Most likely these are simply cases where the number of strains, by chance, is slightly larger. If we extended the limit to 100 strains, say, we would then expect that these systems to also go extinct. This is not certain, however. It is possible that in these cases r_a has remained above 1, and that therefore in some cases strains that die out will be replaced continuously, so that the system sustains itself for a long time, although the more likely case is that once competition kicks in it reduces $\mathbb{E}(r_a)$ below 1, and this drives the system extinct.

It is also worth noting that since the system in this case always begins with only two strains, and the population size is only 10000, the process is very vulnerable to random extinctions due to "bad luck". Running similar simulations in larger populations, or with a larger initial strain pool, might also reveal more of the long-term behaviour of the system.

Realism of the recombination model

The model of recombination I used here is deliberately nonspecific with regard to the mechanism of recombination. It simply assumes that whenever two or more strains are present in the same host, they can give rise to a new strain, with no specification of how precisely the genetic material of the ancestors are combined in the offspring. In this sense, it encapsulates recombination as it occurs in

different pathogens, including recombination in HIV and reassortment in influenza viruses. In this section, I briefly compare the model to these.

HIV. Although the present model can encapsulate different types of recombination, recombination is tied to infection, with each infection event generating at most one recombinant with a constant probability ρ. This means that within the host recombination is density independent. This is not the case for HIV. Recombination depends on the viral concentrations of each strain [108], and so is density dependent. Cells can be simultaneously infected with more than two strains [93], and so a newly arrived strain can recombine with several of the strains that are present [170].

Recombination rates are usually expressed in terms of recombinations per sequence (or per site) per replication. In contrast, in this model all recombination that occurs within a single host is rolled into one parameter ρ. This complicates comparing the model to observed recombination in pathogens, but given the high rates of recombination in coinfected hosts [135, 163], we might say that the effective ρ for HIV is very high, probably close to 1. That is, given coinfection, the emergence of at least one recombinant over the long infection duration is almost certain

In the model presented here, with such a high ρ we would expect uncontrolled growth in the number of strains. This might in fact be the case for HIV. The observed genetic diversity of the global HIV pandemic appears to be increasing [44, 73, 154, 184]. While several strains are often transmitted simultaneously [77] usually only one strain will successfully infect the new host [96], which is captured in this model by the fact that infection events are tied to individual strains.

On the other hand, HIV is not a very easily transmitted pathogen, with most estimates of per-contact transmissibility being less than 5% for even the most high-risk sexual contacts [11], so that in the framework of my model, T is low. This will have the dual effects of reducing Ω_a and increasing the probability $(1-T)^k$ that a newly arisen strain fails to spread. I did not examine different T, and so it is unclear whether the effect of high ρ or low T would win out, and so whether one would observe HIV-like diversity patterns with lower T

The underlying compartmental model that I use is a SEIR model. This is not a very good fit for HIV for several reasons. The incubation period of HIV is on the order of 2-4 weeks [2],

but the disease persists in the host for years. Hence ϵ makes up a vanishingly small fraction of Γ. Because HIV persists over such a long time frame, a population model should also take into account demographic dynamics (births, deaths and migration), which my model does not. However, while SEIR models can exhibit very complex dynamics that SIR models do not [83], in the range of parameters I use they are very similar. A SIR model is a good fit for HIV, assuming that the R compartment is taken to represent death, rather than permanent immunity, and so the use of a SEIR model probably does not invalidate the conclusions that can be drawn.

Finally, very structured population models, such as random graphs, are a good fit for sexually transmitted diseases like HIV in which transmission occurs via very clearly defined contact types, although for long-term diseases like HIV demographic dynamics and the transitory nature of contacts should ideally be included by using a dynamic graph, in which edges are rewired during the process of the epidemic [180].

Influenza A The ratio of ϵ to Γ I use is comparable to estimates for influenza A. In influenza two sequential infections overlapping in time have a very narrow window in which to generate reassortants. This is emulated here by the effect that differences between the origin times of strains have on their overlap, although in Influenza the reassortment window is even smaller relative the Γ [169]. This can to some extent be reflected in the model by using a lower ρ.

Influenza A exhibits a fairly constant and limited level of genetic diversity over time, with repeated cycles every few years of one or a few strains surviving while almost all others go extinct [191]. This is generally due to mutation, rather than extensive reassortment, although the exact mechanisms are still uncertain [19, 152].

Whether the kind of competition that I model here occurs in influenza A in humans is uncertain. There is evidence that when viral concentrations in cells are low, some strains go extinct earlier than others, which is similar to the replacement mechanism modelled here. Unlike my model, this generally happens after most of the viral shedding - and hence transmission - has happened [172].

The role of reassortment in maintaining genetic diversity in influenza is also not clear. Some evidence suggests that reassortment in humans is fairly infrequenct [168], but is frequent in other

species in which influenza A circulates [113]. Perhaps the major role of reassortment in influenza A as a human disease is the occasional introduction of novel gene segments by reassortment of strains circulating in humans with strains circulating in pigs or birds [92]. Such dynamics are obviously not captured by my model.

In my simulations I did not observe any simulations in which a constant low number of recombinant strains was maintained. This might be due to the relatively short time frame of the simulations. It is possible that letting the simulations run for longer would revel a stabilised level of diversity maintained by recombination and competition balancing each other out. However, it seems more likely that competition wins out and the systems generally go extinct. To the extent that my model can be considered a good fit for influenza, this might help explain why the levels of reassortment seen experimentally in humans is low [168], and that mutation is the more important driver of standing diversity in influenza in humans, with reassortment contributing largely via regular injections of segments circulating in other hosts organisms.

Comparison with other models

To my knowledge, to date the only other model to explicitly incorporate recombination on a graph is that of Buckee et al [35]. Their model differs quite substantially in several ways. They considered only regular and small-world graphs, and their model included both selection, mutation, and infection with a parent strain conveying resistance to a recombinant strain. In this scenario, they found that recombination was not a major driver of genetic diversity, because recombinant strains always arose in an environment in which almost all available neighbours were at least partially resistant. In contrast, in my model recombinant strains are considered sufficiently different from their parent strains that there is no cross-immunity, and I find that recombinant strains arise and spread, frequently as well as their parent strains. These two models present opposite extremes in terms of resistance to recombinant strains, and clearly there is room to explore the range in between, in which infection with a parent conveys only partial resistance to recombinant offspring.

Future extensions

There are several issues with the model I used here, and these can be addressed in future work. Firstly, the present model does not illuminate the long-term behaviour of the system. To explore

this, an approximate agent-based simulation, in which each strain rather than each host is an agent, could be used. Such a model requires knowing the expected size of the overlap, which as discussed above is not straightforward to calculate. However, the approximate model could be parameterised from overlap sizes calculated from the current model.

Secondly, recombination is often a density dependent process. Density-dependence was not included in the current model because the number of strains generated grows much faster with density-dependence. However, the approximate agent-based model might be able to handle this scenario. Moreover, when recombination is density-dependent calculating O_A is much more straightforward, not only making the approximate simulation more accurate, but making deriving a tractable mathematical description more feasible.

Many models have been developed which encapsulate the complexities of recombination within a single host (e.g. [170]), which this model does not. Incorporating such a model might lead to more biologically accurate results.

Finally, the primary reason that pathogen recombination is of practical interest is that it can facilitate the rise of novel strains that can cause pandemics, as was the case with the pandemic 2009 influenza A(H1N1) strain [124], or else recombination between two strains each resistant to different drugs can give rise to multi-resistant strains [170]. The present model does not include a specific model of how the genetic material of the parent strains is combined in the offspring, and no differences in fitness between strains are included. Extending the model to include an explicit mechanism of genetic admixture and consequent differences in fitness would allow one to assess whether recombination as a driver of drug resistance is affected by host population structure. Including mutation as a source of diversity, and also perhaps modelling a second reservoir population in which recombination/reassortment is higher and from which novel strains can be transmitted to the main population might also be an interesting scenario. Such models are planned in the future.

Chapter 7

Final conclusions, future directions

In this thesis I have investigated the dynamics of multiple strains, circulating simultaneously or with some delay between them, on two types of random graphs with radically different degree distribution. The disease was modelled using multiple simultaneous SEIR models. The graphs considered have either Poisson or Zipf distributed degrees, and were generated using a *vertex-focused* variation on the configuration model.

In chapter 4 I examined the dynamics of the *overlap* – the number of vertices simultaneously infected with both strains – when the two strains have no effect on each others' transmission. The most important determinants of the size of the overlap were the transmissibility of the pathogens, as the overlap depends strongly on the size of each strain's outbreak; and the delay between the introduction of the two strains, which had the effect of reducing the size of the overlap. When the two strains were introduced simultaneously but had different transmissibility, the differences in the speed at which the outbreaks reached their peak size at high and low transmissibility were found to have a similar effect to introducing a delay. Interestingly, I found that when the two strains have different transmissibility, a more infectious strain being introduced shortly after a less infectious strain had the effect of *increasing* the size of the overlap relative to the case when the two strains are introduced separately, because the faster-spreading second strain will reach peak outbreak size at roughly the same time as the slower-spreading first strain.

Comparing the behaviour of the overlap in the two graph types, the most striking difference is that overlaps in Poisson distributed graphs depend more strongly on transmissibility and delay, and that comparing outbreaks of equal size, overlaps are much larger in Zipf distributed graphs. Just as outbreaks cluster in the high-degree vertices in Zipf distributed graphs, overlaps on Zipf distributed graphs were found to cluster even more strongly in the vertices with highest degree.

This exploration revealed that the overlap of outbreaks depends on the dynamics of the underlying host population structure. Understanding the overlap of diseases in the absence of interactions can be important. Several members of the virus family *Herpesviridae*, for example Cytomegalovirus (CMV) are extremely prevalent in human populations but in the absence of immunosuppresion conditions very rarely cause serious symptoms [84]. However, when HIV infection causes immunosuppression, these viruses can cause serious illness [20,109]. Understanding how common coinfection with CMV and HIV or HSV and HIV is likely to be is therefore useful for estimating and preparing for the likely burden of disease in the HIV positive population.

The model studied in this chapter assumes that there are no interactions between different strains within the host that can lead to changes in the extent of the spread of either disease. Many disease do affect each other's ability to be transmitted. However, if the contact graphs underlying the spread of two diseases are sufficiently well known, this model allows one to predict the expected overlap in the absence of any interactions, and can therefore serve as a null model, against which one can compare observed rates of coinfection in order to infer interactions between the two diseases.

Planned developments of this work include a more extensive exploration of different random graph types, as well as examining the effect of different pruning schemes in which vertices are selectively removed from the graphs.

In chapter 5 I extended the model from chapter 4 to allow strains to interact by modifying each others' transmissibility: infection with one strain made the vertex either more or less able to pass on infection with a subsequent strain. Both symmetric modification, in which each strain changes susceptibility to the other; and one-sided modification, in which one strain alters susceptibility but the other does not were examined. These schemes were examined with a variety of delays between the strains being introduced.

If one strain enhances the transmissibility of a second strain, then the second strain will infect a larger fraction of the population. Likewise, if one strain hinders transmission of the other, then the hinderd strain will infect a fewer hosts. Introducting a delay between the two strains arriving in the host population reduces their overlap and so reduces the impact of the modifying strain on the other.

The modified strain experiences a larger change in the number of hosts infected in Zipf distributed graphs, but the size of the overlap between the two graphs is changed more in the Poisson distributed graphs. These differences are very small, despite the differences in the way in which the diseases spread in the two graphs

The order in which the strains are introduced matters. When the modifying strain is introduced first, it has time to affect a significant portion of the population. When the modifying strain spreads after the other, it has a limited effect on the spread of the other. In this latter case, the effect is larger and is present with larger delays in the Zipf distributed graphs, where both diseases tend to reach well-connected hosts first and then cascade out to less well-connected vertices.

Although to my knowledge no previous work has studied the effect of tranmission modification in isolation on graphs, several authors have studied the similar mechanism of immune modification, in which a host infected with one strain experiences a changed susceptibility to infection with another. Comparing my results to theirs, it appears firstly that the effect of transmission modification depends much less on the degree heterogeneity of the graphs. In my simulations, although there were quantitative differences in the effect of modification in the two graph types, these were slight.

Marceau *et al*, who studied immune modification in the same two graph types, found a much more pronounced effect in Poisson distributed graphs than in Zipf distributed graphs. It appears that this difference is due to the fact that under immune modification the change in susceptibility to one strain in a host depends on that host's infection status relative to the second strain, while under transmission modification the change in suscueptibility depends on the status of all its neighbours.

I therefore speculate that the degree of clustering in the graph might have a significant effect, and that the differences between graph types under immune modification, which one should expect

to obtain at least partially under transmission modification, are offset by the higher clustering in Zipf distributed graphs. This bears further study, and future work might focus on comparing the effects transmission modfication on graphs with equal degree distributions but differing clustering coefficients.

In chapter 6 I extended the model from hapter 4 in two ways - first, by allowing the two strains to recombine to produce novel strains, that could themselves reproduce, and secondly by imposing an upper limit on the number of strains that could simultaneously infect a vertex. Once a sufficient number of recombinant strains arises, this introduces competition among strains in a different form than that studied in chapter 5.

In the absence of competition, each simulation generated either very few recombinant strains or a great many, and nothing in between. Ther did not appear to be a limit to the number of strains generated in those simulations where more than a few arose, although due to computational limitations the long-term behaviour of the system could not be simulated. The largest effect of graph structure in the models considered in this chapter was that of the recombinant strains that arose, far more failed to spread beyond their index vertex in the Zipf distributed graph than in the Poisson distributed graph. This was found to be due to the fact that strains that arise in very poorly connected vertices have a larger chance of failing to infect any of their neighbours.

Future planned work from these experiments include studying the long-term behaviour of the system, determining the threshold between extinction and uncontrolled growth, and the behaviour of the system around this threshold. A promising direction for this work is the framing of the problem on the scale of strain rather than individual hosts, which suggests several simulation approaches and the use of branching processes for analytical results. The main obstacle is a sufficiently thorough understanding of the overlap of more than two strains, and the dependence of that overlap on the time of origin of each strain. Another potential area of investigation is the role of other topological features of graphs on the number of strains generated. In particular, altering the degree assortativity may strengthen or mitigate the chances of recombinant strains going extinct in heterogeneous graphs.

The main effect of imposing competition was to reduce the size of each outbreak. This in turn causes the overlaps between outbreaks to becomes smaller, leading to fewer recombinants arising later in the simulations. This lead to a large number of simulations in which a relatively large number of strains arose, but where the system went extinct within the time frame studied. This extinction is heralded by an increase in the mean size of outbreaks. Even under competition some systems still generated the maximum number of strains allowed due to computational limitations. It remains to be seen whether these simulations represent a case in which the competing pressures of competition and recombination are balanced and a constant number of strains is maintained in the long term; or whether there is simply more variation in the number of recombinants generated before extinction than the computational limitations imposed allowed me to determine.

In the presence of competition many strains failed to spread beyond their index host, even in the Poisson distributed graphs. This effect was not found to depend on graph structure, and despite the well-documented differences in outbreak and overlap sizes on different graph types, there were no other observed differences between the two graph types studied here.

Concern about the rise of immunologically novel pathogen strains and multi-resistant strains are major reasons to be interested in recombination dynamics in structured host populations. Two model extensions I am currently examining introduce either a vaccination scheme or a drug treatment scheme, with recombination and mutation allowing strains to develop resistance.

Bibliography

[1] H. Abbey. An examination of the Reed-Frost theory of epidemics. *Human Biology*, 24, 1952.

[2] P. Alcabes, A. Muñoz, D. Vlahov, and G. H. Friedland. Incubation period of human immunodeficiency virus. *Epidem. Rev.*, 15, 1993.

[3] S. Alizon. Decreased overall virulence in coinfected hosts leads to persisitence of virulent parasites. *Am. Naturalist*, 172, 2008.

[4] R. M Anderson and R. M May. Regulation and stability of host-parasite interactions. I. regulatory processes. *J. Anim. Eco.*, 47, 1978.

[5] R. M. Anderson and R. M. May. Regulation and stability of host-parasite interactions. II. destabilizing processes. *J. Anim. Eco.*, 47, 1978.

[6] R. M. Anderson and R. M. May. Population biology of infectious diseases. part I. *Nature*, 280, 1979.

[7] R. M. Anderson and R. M. May. Population biology of infectious diseases. part II. *Nature*, 280, 1979.

[8] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[9] J. Archer, J. W. Pinney, J. Fan, E. Simon-Loriere, and E. J. Arts *et al.* Identifying the important HIV-1 recombination breakpoints. *PLoS Comp. Bio.*, 4, 2008.

[10] N. Azimi-Tafreshi. Cooperative epidemics on multiplex networks. *Phys. Rev. E*, 93, 2016.

[11] R. F. Baggaley, R. G. White, and M.-C. Boily. HIV transmission risk through anal intercourse: systematic review, meta-analysis and implications for HIV prevention. *Int. J. Epidemiol.*, 39, 2010.

[12] F. Ball and T House. Heterogeneous network epidemics: real-time growth, variance and extinction of infection. *J. Math. Bio.*, 75, 2017.

[13] F. Ball and P. Neal. Network epidemic models with two levels of mixing. *Math. Biosci.*, 212, 2008.

[14] F. G. Ball, D. J. Sirl, and P. Trapman. Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Adv. Appl. Probab.*, 41, 2009.

[15] F. G. Ball, D. J. Sirl, and P. Trapman. Analysis of a stochastic SIR epidemic on a random network incorporating household structure. *Math. Biosci.*, 224, 2010.

[16] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[17] C. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. V. Marathé. EpiSimdemics: an efficient and scalable framework for simulating the spread of infectious disease on large social networks. *Proceedings of the 2008 ACM/IEEE conference on Supercomputing.*, 2008.

[18] C. Bauch and D. A. Rand. A moment closure model for sexually transmitted disease transmission through a concurrent partnership network. *Proc. R. Soc. Lond. B*, 267, 2000.

[19] T. Bedford, A. Rambaut, and M. Pascual. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Bio.*, 10, 2012.

[20] S. Bibert, A. Wojtowicz, P. Taffé, O. Mauel, and E. Bernasconi *et al.* The IFNL3/4 ΔG variant increases susceptibility to cytomegalovirus retinitis among HIV-infected patients. *AIDS*, 28, 2014.

[21] K. R. Bisset, J. Chen, and X. Feng. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. *Proceedings of the 23rd ACM International Conference on Supercomputing*, 2009.

[22] O. N. Bjørnstad and C. Viboud. Timing and periodicity of influenza epidemics. *Proc. Natl. Acad. Sci. USA*, 113, 2016.

[23] M. Boguñá, C. Castellan, and R. Pastor-Satorras. Nature of the epidemic threshold for the susceptible-infected-susceptible dynamics in networks. *Phys. Rev. Lett.*, 111, 2013.

[24] M. Boguña, L. F. Lafuerza, R. Toral, and M. A. Serrano. Simulating non-Markovian stochastic processes. *Phys. Rev. E*, 90, 2014.

[25] M. Boguñá and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Phys. Rev. E*, 66, 2002.

[26] B. Bollobás. *Random Graphs (2nd edition)*. Cambridge University Press, 2001.

[27] M. F. Boni, J. R. Gog, V. Andreasen, and F. B. Christiansen. Influenza drift and epidemic size: the race between generating and escaping immunity. *Theor. Pop. Bio.*, 65, 2004.

[28] H. Bradley, L. E. Markowitz, T. Gibson, and G. M. McQuillan. Seroprevalence of Herpes simplex types 1 and 2, United States, 1999-2010. *J. Inf. Dis.*, 209, 2014.

[29] F. Brauer. Compartmental models in epidemiology. In F. Brauer, P. van den Driessche, and J. Wu, editors, *Mathematical epidemiology*, chapter 2. Springer, Berlin, 2008.

[30] F. Brauer, P. van den Driessche, and J. Wu. *Mathematical epidemiology*. Springer, 2008.

[31] T. Britton. Stochastic epidemic models: a survey. *Math. Biosci.*, 225, 2010.

[32] S. Broadbent and J. Hammersley. Percolation processes I. Crystals and mazes. *Proc. Cam. Phil. Soc.*, 53, 1957.

[33] J. Bruchfield, M. Correia-Never, and G. Källenius. Tuberculosis and HIV coinfection. *Cold Spring Harb. Perspect. Med.*, 5, 2015.

[34] J. F. Brundage and G. D. Shanks. Deaths from bacterial pneumonia during 1918-19 influenza pandemic. *Emerg. Inf. Dis.*, 14, 2008.

[35] C. O'F. Buckee, K. Koelle, M. J. Mustard, and S. Gupta. The effects of host contact network structure on pathogen diversity and strain structure. *Proc. Natl. Acad. Sci. USA*, 101, 2004.

[36] K. M. Carley, D. B. Fridsma, E. Casman, A. Yahja, and N. Altman *et al.* BioWar: scalable agent-based model of bioattacks. *IEEE Transactions on Systems. Man. and Cybernetics-Part A: Systems and Humans.*, 36, 2006.

[37] C. Castellano and R. Pastor-Satorras. Thresholds for epidemic spreading in networks. *Phys. Rev. Lett.*, 105, 2010.

[38] E. Cator and P. Van Mieghem. Second-order mean-field susceptible-infected-susceptible epidemic threshold. *Phys. Rev. E*, 85, 2012.

[39] E. Cator and P. Van Mieghem. Susceptible-infected-susceptible epidemics on the complete graph and the star graph: exact analysis. *Phys. Rev. E*, 87, 2013.

[40] D. Chakrabarti, Y. Wang, C. Wang, Y. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Trans. Inf. Sys. Sec.*, 10, 08.

[41] D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini Jr. FluTE. a publicly available stochastic influenza epidemic simulation model. *PLoS Comp. Bio.*, 6, 2010.

[42] V. Colizza and A. Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.*, 99, 2007.

[43] E. L. Corbett, C. J. Watt, J. Walker, and D. Maher *et al.* The growing burden of tuberculosis - global trends and interactions with the HIV epidemic. *Arch. Intern. Med*, 163, 2003.

[44] C. R. Courtney, L. Agyingi, F. Arlette, C. Stephanie, and A. Bladine *et al.* Monitoring HIV-1 group M subtypes in Yaoundé, Cameroon reveals broad genetic diversity and a novel CRF02-AGF2 infection. *AIDS Res. Hum. Retrov.*, 32, 2016.

[45] D. Cromer, T. E. Schlub, R. P. Smyth, A. J. Grimm, and A. Chopra *et al.* HIV-1 mutation and recombination rates are different in macrophages and T-cells. *Viruses*, 8, 2016.

[46] F. Darabi Sahneh, C. Scoglio, and P. Van Mieghem. Generalised epidemic mean-field model for spreading processes over multilayer complex networks. *IEEE/ACM Trans. Net.*, 21, 2013.

[47] J. C. de Jong, G. F. Rimmelzwaan, R. A. M. Fouchier, and A. D. M. E. Osterhaus. Influenza virus: a master of methamorphosis. *J. Infect.*, 40, 2000.

[48] M. C. M. de Jong, O. Diekmann, and H. Heesterbeek. How does transmission of infection depend on population size? In D. Mollison, editor, *Epidemic Models: their structure and relation to data.* Cambridge University Press, 1995.

[49] J. C. de Roode, M. E. H. Helinski, M. A. Anwar, and A. F. Read. Dynamics of multiple infection and within-host competition in genetically diverse malaria infections. *Am. Naturalist*, 166, 2005.

[50] J. C. de Roode, A. F. Read, B. H. K. Chan, and M. J. Mackinnon. Rodent malaria parasites suffer from the presence of con-specific clones in three-clone *Plasmodium chabaudi* infections. *Parasitology*, 127, 2003.

[51] C. Dibble and P. G. Feidman. The GeoGraph 3d computational laboratory network on terrain landscapes for RePast. *J. Artif. Soc. Soc. Sim.*, 7, 2004.

[52] E. Dion, L. Vanschalkwyk, and E. F. Lambin. The landscape epidemiology of foot-and-mouth disease in South Africa: a spatially explicit multi-agent simulation. *Eco. Model.*, 222, 2011.

[53] W. Duan, Z. Cao, Y. Ge, and X. Qiu. Modeling and simulation for the spread of H1N1 influenza in school using artificial societies. *Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics*, 2011.

[54] W. Duan, Z. Fan, P. Zhang, G. Guo, and X. Qiu. Mathematical and computational approaches to epidemic modelling: a comprehensive review. *Front. Comp. Sci.*, 9, 2015.

[55] R. I. McD. Dunbar. Neocortex size as a constraint on group size in primates. *J. Hum. Evo.*, 22, 1992.

[56] J. B. Dunham. An agent-based spatially explicit epidemiological model in MASON. *J. Artif. Soc. Soc. Sim.*, 9, 2005.

[57] D. W. Dyer. Uncommons maths v. 1.2.3: Random number generators, probability distributions, combinatorics and statistics for Java., 2012.

[58] K. T. D. Eames and M. J. Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proc. Natl. Acad. Sci. USA*, 99, 2002.

[59] D. D. D. Earn. A light introduction to modelling recurrent epidemics. In F. Brauer, P. van den Driessche, and J. Wu, editors, *Mathematical epidemiology*, chapter 1. Springer, Berlin, 2008.

[60] M. Engelberg. clojure.data/priority-map.clj v. 0.0.7, 2013.

[61] P. Erdös and A. Renyi. On random graphs. *Pub. Math.*, 6, 1959.

[62] P. Erdös and A. Renyi. On the evolution of random graphs. *Pub. Math. Inst. Hung. Acad. Sci.*, 5, 1960.

[63] P. Erdös and A. Renyi. On the strength of connectedness of a random graph. *Acta Math. Sci. Hung.*, 12, 1961.

[64] S. G. Eubank, G. H. Kumar, M. V. Marathé, A. Srinivasan, Z. Toroczkai, and N. Wang. Modeling disease outbreaks in realistic urban social networks. *Nature*, 429, 2004.

[65] S. C. Ferreira, R. Castellano, and R. Pastor-Satorras. Epidemic thresholds of the susceptible-infected-susceptible model on networks: A comparison of numerical and theoretical results. *Phys. Rev. E*, 86, 2012.

[66] P. E. Fine and J. A. Clarkson. Measles in England and Wales – I: An analysis of factors underlying seasonal patterns. *Int. J. Epidemiol.*, 11, 1982.

[67] J. Flint, V. Racaniello, G. Rall, and A. M. Skalka. *Principles of Virology, 4th ed.* American Society of for Microbiology, 2015.

[68] C. Fraser. HIV recombination: what is the impact on antiretroviral therapy? *J. R. Soc. Interface*, 2, 2005.

[69] E. E. Freeman, H. A. Weiss, J. R. Glynn, P. L. Cross, and J. A. Whitworth *et al.* Herpes simplex virus 2 infection increases HIV acquisition in men and women: Systematic review and meta-analysis of longitudinal studies. *AIDS*, 20, 2006.

[70] S. Funk and V. A. A. Jansen. Interacting epidemics on overlap networks. *Phys. Rev. E*, 81, 2010.

[71] G. P. Garnett, J. P. Hughes, R. M. Anderson, B. P. Stoner, and S. O. Aral *et al.* Sexual mixing patterns of patients attending sexually transmitted diseases clinics. *Sex. Trans. Dis.*, 23, 1996.

[72] V. V. Gasunov, R. A. Neher, and A. S. Perelson. Mathematical modeling of escape of HIV from cytotoxic T lymphocyte responses. *J. Stat. Mech. Theor. Exper.*, 2013, 2013.

[73] R. J. Gifford, T. de Oliveira, A. Rambaut, O. G. Pybus, and D. Dunn *et al.* Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1. *J. Virol.*, 81, 2007.

[74] E. Gilbert. Random graphs. *Ann. Math. Stat.*, 30, 1959.

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

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

[77] J. R. Gog, L. Pellis, J. L. N. Wood, A. R. McLean, N. Arinaminpathy, and J. O. Lloyd-Smith. Seven challenges in modeling pathogen dynamics within-host and across scales. *Epidemics*, 10, 2015.

[78] S. Gomez, A. Arenas, J. Borge-Holthoefer, S. Meloni, and Y. Moreno. Discrete-time Markov chain approach to contact-based disease spreading in complex networks. *Europhys. Lett.*, 2010.

[79] B. Gonçalves, N. Perra, and A. Vespignani. Modeling users' activity on twitter networks: Validation of Dunbar's number. *PLoS ONE*, 6, 8.

[80] M. Graham and T. House. Dynamics of stochastic epidemics on heterogeneous networks. *J. Math. Bio.*, 68, 2014.

[81] P. Grassberger, Li. Chen, F. Ghanbarnejad, and W. Cai. Phase transitions in cooperative coinfections: Simulation results for networks and lattices. *Phys. Rev. E*, 93, 2016.

[82] B. Grenfell and J. Harwood. (Meta)population dynamics of infectious diseases. *Trends. Eco. Evo.*, 12, 1997.

[83] B. T. Grenfell and B. M. Bolker. Chaos and biological complexity in measles dynamics. *Proc. R. Soc. Lond. B*, 251, 1993.

[84] M. Hecker, D. Qui, K. Marquardt, G. Bein, and H. Hackstein. Continuous cytomegalovirus seroconversion in a large group of healthy blood donors. *Vox. Sang.*, 86, 2004.

[85] J. T. Herbeck, J. E. Mittler, G. S. Gottlieb, and J. I. Mullins. An HIV epidemic model based on viral load dynamics: value in assessing epirical trends in HIV virulence and community viral load. *PLoS Comp. Bio.*, 10, 2014.

[86] T. Hladish, E. Melamud, L. A. Barrera, A. Galvani, and L. A. Meyers. EpiFire: and open-source C++ library and application for contact network epidemiology. *BMC Bioinf.*, 13, 2012.

[87] S. Hochman and K. Kim. The impact of HIV and malaria coinfection: what is known and suggested venues for further study. *Interdisc. Perspect. Inf. Dis.*, 2009, 2009.

[88] R. E. Hope Simpson. Infectiousness of communicable diseases in the household: (measles, cihckenpox and mumps). *Lancet*, 260, 1952.

[89] R. E. Hope Simpson. The role of season in the epidemiology of influenza. *J. Hyg. (Lond.)*, 86, 1981.

[90] T. House and M. J. Keeling. Epidemic prediction and control in clustered populations. *J. Theor. Bio.*, 272, 2011.

[91] T. House and M. J. Keeling. Insights from unifying modern approximations to infections on networks. *J. R. Soc. Interface*, 8, 2011.

[92] W. L. Ince, A. Gueye-Mbaye, J. R. Bennink, and J. W. Yewdell. Reassortment complements spontaneous mutation in influenza A virus NP and M1 genes to accelerate adaptation to a new host. *J. Virol.*, 87, 2013.

[93] A. Jung, R. Maier, J. P. Vartanian, G. Bocharov, and V. Jung *et al.* Multiply infected spleen cells in HIV patients. *Nature*, 418, 2002.

[94] B. Karrer and M. E. J. Newman. Message passing approach for general epidemic models. *Phys. Rev. E*, 82, 2010.

[95] B. Karrer and M. E. J. Newman. Competing epidemics on complex networks. *Phys. Rev. E*, 84, 2011.

[96] B. F. Keele, E. E. Giorgi, J. F. Salazar-Gonzalez, J. M. Decker, and K. T. Pham. Indetification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci. USA*, 105, 2008.

[97] M. J. Keeling. The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. Lond. B*, 266, 1999.

[98] M. J. Keeling and K. T. D. Eames. Networks and epidemic models. *J. R. Soc. Interface*, 2, 2005.

[99] M. J. Keeling, D. A. Rand, and A. J. Morris. Correlation models for childhood epidemics. *Proc. R. Soc. Lond. B*, 264, 1997.

[100] K. S. Kemal, C. M. Ramirez, Burger. H., B. Foley, and D. Mayers *et al.* Recombination between variants from genital tract and plasma: evolution of multidrug-resistant HIV type 1. *AIDS Res. Hum. Retrov.*, 28, 2012.

[101] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. London A*, 115, 1927.

[102] H. Kesten. *Percolation theory for mathematicians.* Birkäuser, 1982.

[103] P. D. Kirwan, C. Chau, A. E. Brown, O. N. Gill, and V. C. Delpech *et al.* HIV in the UK - 2016 report. 2016.

[104] I. E. Kiwelu, V. Novitsky, L. Margolin, Baca J., R. Manongi, and N. Sam *et al.* Frequent intra-subtype recombination among HIV-1 circulating in Tanzania. *PLoS ONE*, 8, 2013.

[105] A. Kleinman. Random recombination and evolution of drug resistance. *Stats. Med.*, 34, 2015.

[106] N. Lee, P. K. S. Chan, D. S. C. Hui, T. H. Rainer, and E. Wong *et al.* Viral loads and duration of viral shedding in adult patients hospitalized with influenza. *J. Inf. Dis.*, 200, 2009.

[107] J. Lessler, N. G. Reich, R. Brookmeyer, T. M. Perl, K. E. Nelson, and D. A. T. Cummings. Incubations of acute respiratory viral infections: a systematic review. *Lancet Inf. Dis.*, 9, 09.

[108] D. N. Levy, G. M. Aldrovandi, O. Kutsch, and G. M. Shaw. Dynamics of HIV recombination in its natural target cells. *Proc. Natl. Acad. Sci. USA*, 101, 2004.

[109] M. Lichtner, P. Cicconi, S. Vita, A. Cozzi-Lepri, and M. Galli *et al.* Cytomegalovirus coinfection is associated with an increased risk of severe non-AIDS-defining events in a large cohort of HIV-infected patients. *J. Inf. Dis.*, 211, 2015.

[110] F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg. The web of human sexual contacts. *Nature*, 411, 2001.

[111] F. Liu, W. T. A. Enanoria, J. Zipprich, S. Blumberg, K. Harriman, and S. F. Ackley *et al.* The role of vaccination coverage, individual behaviours, and the public health response in the control of measles epidemics: an agent-based simulation for California. *BMC Pub. Health*, 15, 2015.

[112] A. L. Lloyd. Realistic distributions of infectious periods in epidemic models: Changing patterns of persistence and dynamics. *Theor. Pop. Bio.*, 60, 2001.

[113] S. J. Lycett, G. Baillie, G. Coulter, S. Bhatt, and P. Kellam *et al.* Estimating reassortment rates in co-circulating Eurasian swine influenza viruses. *J. Gen. Virol.*, 93, 2012.

[114] W. A. Lynn and S. Lightman. Syphilis and HIV: a dangerous combination. *Lancet Inf. Dis.*, 4, 2004.

[115] V. Marceau, P.-E. Noël, L. Hébert-Dufresne, A. Allard, and L. J. Dubé. Modelling the dynamical interaction between epidemics on overlay networks. *Phys. Rev. E*, 84, 2011.

[116] N. Marshall, L. Priyamavada, Z. Ende, J. Steel, and A. C. Lowen. Influenza virus reassortment occurs with high frequency in the absence of segment mismatch. *PLoS Path.*, 9, 2013.

[117] M. Martcheva and S. S. Pilyugin. The role of coinfection in multidisease dynamics. *SIAM J. Appl. Mat.*, 66, 2006.

[118] A. S. Mata and S. C. Ferreira. Pair quenched mean-field theory for the susceptible-infected-susceptible model on complex networks. *EPL*, 103, 2013.

[119] M. Matsumoto and T. Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Mod. Comp. Sim.*, 8, 1998.

[120] R. M. May and A. L. Lloyd. Infection dynamics on scale-free networks. *Phys. Rev. E*, 66, 2001.

[121] R. M. May and M. A. May. Superinfection, metapopulation dynamics, and the evolution of diversity. *J. Theor. Bio.*, 170, 1994.

[122] R. M. May and M. A. Nowak. Coinfection and the evolution of parasite virulence. *Proc. Bio. Sci.*, 261, 1995.

[123] H. McCallum, N. Barlow, and J. Hone. How should pathogen transmission be modelled? *Trends Eco. Evo.*, 16, 2001.

[124] I. Mena, M.I. Nelson, F. Quezada-Monroy, J. Dutta, and R. Cortes-Fernández *et al.* Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. *eLife*, 5, 2016.

[125] L. A. Meyers, B. Pourbohloul, M. E. J. Newman, D. M. Skowronski, and R. C. Brunham. Network theory and SARS: predicting outbreak diversity. *J. Theor. Bio.*, 232, 2005.

[126] J. C. Miller. A note on a paper by Erik Volz: SIR dynamics in random networks. *J. Math. Bio.*, 62, 2011.

[127] J. C. Miller. Cocirculation of infectious diseases on networks. *Phys. Rev. E*, 87, 2013.

[128] J. C. Miller, A. C. Slim, and E. Volz. Edge-based compartmental modelling for infectious disease spread. *J. R. Soc. Interface*, 9, 2012.

[129] R. Milo, N. Kashtan, S Itskovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences.

[130] Y. Min, J. Hu, W. Wang, Y. Ge, J. Chang, and X. Jin. Diversity of multilayer networks and its impact on collaborating epidemics. *Phys. Rev. E*, 99, 2014.

[131] S. M. Mniszewski, S. Y. D. Valle, P. D. Strond, J. M. Riese, and S. J. Sydoriak. EpiSims simulation of a multicomponent strategy for pandemic influenza. *Proceedings of Spring Simulation Multiconference*, 2008.

[132] K. Modjarrad and S. H. Vermund. Effect of treating co-infections on HIV-1 viral load: a systematic review. *Lancet Inf. Dis.*, 10, 2010.

[133] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Rand. Struct. Alg.*, 6, 1995.

[134] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *Eur. Phys. J. B.*, 26, 2002.

[135] R. A. Neher and T. Leitner. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comp. Bio.*, 6, 2010.

[136] M. I. Nelson, S. E. Detmer, D. E. Wentworth, Y. Tan, and A. Schwartzbard *et al.* Genomic reassortment of influenza A virus in North American swine, 1998-2011. *J. Gen. Virol.*, 93, 2012.

[137] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66, 2002.

[138] M. E. J. Newman. Threshold effects for two pathogens spreading on a network. *Phys. Rev. Lett.*, 95, 2005.

[139] M. E. J. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[140] M. E. J. Newman and C. R. Ferrario. Interacting epidemics and coinfection on contact networks. *PLoS ONE*, 8, 2013.

[141] P.-A. Noël, A. Allard, L. Hébert-Dufresne, V. Marceau, and L. J. Dubé. Propagation on networks: An exact alternative perspective. *Phys. Rev. E*, 85, 2012.

[142] J. Parker and J. M. Epstein. A distributed platform for glocal-scale agent-based models of disease transmission. *ACM Trans. Mod. Comp. Sim.*, 22, 2011.

[143] R. Pastor-Satorras and A. Castellano. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86, 2001.

[144] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Rev. Mod. Phys.*, 87, 2015.

[145] R. Pastor-Satorras and A. Vespignani. Epidemic dynamics and endemic states in complex networks. *Phys. Rev. E*, 63, 2001.

[146] L. Pellis, F. G. Ball, and P. Trapman. Reproduction numbers for epidemic models with households and other social structures. I. definition and calculation of R_0. *Math. Biosci.*, 235, 2012.

[147] M. Pérez-Losada, M. Arenas, J. C. Galán, F. Palero, and González-Candela. Recombination in viruses: mechanisms, methods of study and evolutionary consequences. *Infect. Gen. Evo.*, 30, 2015.

[148] K. V. N. Prasad and P. Mohamed Ali. Incubation period of leprosy. *Indian J. Med. Res.*, 55, 1967.

[149] F. Radicchi and C. Castellano. Beyond the locally treelike approximation for percolation on real networks. *Phys. Rev. E*, 93, 2016.

[150] A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, and E. C. Holmes. The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453, 2009.

[151] D. A. Rasmussen, R. Kouyos, H. F. Güthard, and T. Stadler. Phylodynamics on local sexual contact networks. *PLoS Comp. Bio.*, 13, 2017.

[152] M. Recker, O. G. Pybus, S. Nee, and S. Gupta. The generation of influenza outbreaks by a network of host immune responses against a limited set of antigenic types. *Proc. Natl. Acad. Sci. USA*, 104, 2007.

[153] D. L. Robertson, P. M. Sharp, F. E. McCutchan, and B. H. Hahn. Recombination in HIV-1. *Nature*, 374, 1995.

[154] M. A. Rodgers, E. Wilkinson, A. Vallari, C. McArthur, and L. Sthreshley *et al.* Sensitive next-generation sequencing method reveals deep genetic diversity of HIV-1 in the Democratic Republic of the Congo. *J. Virol.*, 91, 2017.

[155] A. Rodriguez-Carvajal, K. A. Crandall, and D. Posada. Recombination favors the evolution of drug resistance in HIV-1 during antiretroviral therapy. *Infect. Gen. Evo.*, 7, 2007.

[156] L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes and martingales.* Cambridge University Press, 2000.

[157] J. Sanz, C.-Y. Xia, S. Meloni, and Y. Moreno. Dynamics of interacting diseases. *Phys. Rev. X*, 4, 2014.

[158] E. Sartori, A. Calistri, C. Salata, C. del Vecchio, and G. Palù. Herpes simplex virus type 2 infection increase human immunodeficiency virus type 1 entry into human primary macrophages. *Virology*, 8, 2011.

[159] N. Schwartz and L. Stone. Exact epidemic analysis for the star topology. *Phys. Rev. E*, 87, 2013.

[160] K. J Sharkey. Deterministic epidemiological models at the individual level. *J. Math. Bio.*, 57, 2008.

[161] K. J. Sharkey. Deterministic epidemic models on contact networks: Correlations and unbiological terms. *Theor. Pop. Bio.*, 79, 2011.

[162] P. M. Sharp and B. H. Hahn. Origins of HIV and AIDS pandemic. *Cold Spring Harb. Perspect. Med.*, 1, 2011.

[163] D. Shirner, A. G. Rodrigo, D. C. Nickle, and J. I. Mullins. Pervasive genomic recombination of HIV-1 *in vivo*. *Genetics*, 167, 2004.

[164] M. Shreshta, S. V. Scarpino, and C. Moore. Message-passing approach for recurrent-state epidemic models on networks. *Phys. Rev. E*, 92, 2015.

[165] P. Simon, M. Taylor, and I. Kiss. Exact epidemic models on graphs using graph-automorphism driven lumping. *J. Math. Bio.*, 62, 2011.

[166] G. J. D. Smith, V. Dhanasekaran, J. Bahl, S. Lycett, and S. Worobey *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, 459, 2009.

[167] R. P. Smyth, T. E. Schlub, A. J. Grimm, C. Waugh, and P. Ellenberg *et al.* Identifying recombination hot spots in the HIV-1 genome. *J. Virol.*, 88, 2014.

[168] A. Sobel Leonard, M. T. McClain, G. J. D. Smith, D. E. Wentworth, and R. A. Halpin *et al.* The effective rate of influenza reassortment is limited during human infection. *PLoS Path.*, 13, 2017.

[169] J. Steel and A. C. Lowen. Influenza A virus reassortment. In R. W. Compans and M. B. A. Oldstone, editors, *Influenza pathogenesis and control - volume I*, chapter 16. Springer, Berlin, 2014.

[170] G. W. Suryavanshi and N. M. Dixit. Emergence of recombinant forms of HIV: Dynamics and scaling. *PLoS Comp. Bio.*, 3, 2007.

[171] H. Susi, B. Barrès, P. F. Vale, and A.-L. Laine. Co-infection alters population dynamics of infectious disease. *Nature Comm.*, 6, 2015.

[172] H. Tao, J. Steel, and A. C. Lowen. Intrahost dynamics of influenza virus reassortment. *J. Virol.*, 88, 2014.

[173] V. Trifonov, H. Khiabanian, and R. Rabadan. Geographic dependence, surveillance and origins of the 2009 influenza A (H1N1) virus. *N. Eng. J. Med*, 361, 2009.

[174] P. van de Perre, M. Segondy, V. Foulongne, A. Ouedraogo, and I. Konate *et al.* Herpes simplex virus and HIV-1: Deciphering viral synergy. *Lancet Inf. Dis.*, 8, 2008.

[175] P. Van Mieghem. Exact Markovian SIR and SIS epidemics on networks and an upper bound for the epidemic threshold. *arXiv:1402.1731*, 2014.

[176] P. Van Mieghem. *Performance Analysis of Complex Networks and Systems.* Cambridge University Press, 2014.

[177] P. Van Mieghem, J. Omic, and R. Kooji. Virus spread in networks. *IEEE/ACM Trans. Net.*, 17, 2008.

[178] D. A. Vasco, H. J. Wearing, and P. Rohani. Tracking the dynamics of pathogen interactions: Modelling ecological and immune-mediated processes in a two-pathogen single-host system. *J. Theor. Bio.*, 245, 2007.

[179] E. Volz. SIR dynamics in random networks with heterogeneous connectivity. *J. Math. Bio.*, 56, 2008.

[180] E. Volz and L. A. Meyers. Epidemic thresholds in dynamic contact networks. *J. R. Soc. Interface*, 6, 2009.

[181] E. Volz and L. A. Meyers. Epidemic thresholds in dynamic networks. *J. R. Soc. Interface*, 6, 2009.

[182] W. Wang, M. Tang, H. E. Stanley, and L. A. Braunstein. Unification of theoretical approaches for epidemic spreading on complex networks. *Rep. Prog. Phys.*, 80, 2017.

[183] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 1998.

[184] E. Wilkinson, D. Rasmussen, O. Ratmann, T. Stadler, S. Engelbrecht, and T. de Oliveira. Origin, imports and exports of HIV-1 subtype C in South Africa : a historical perspective. *Inf. Gen. Evo.*, 46, 2016.

[185] E. B. Wilson, C. Bennet, M. Allen, and J. Worcester. Measles and scarlet fever in Providence RI. *Proc. Am. Phil. Soc.*, 80, 1939.

[186] F. Xu, J. A. Schillinger, L. E. Markowitz, M. R. Sternberrg, M. R. Rubin, and M. E. St. Louis. Repeat *Chlamydia trachomatis* infection in women: analysis through a surveillance case registry in Washington state, 1993-1998. *Am. J. Epidemiol.*, 152, 2000.

[187] R.-H. Xu, J.-F. He, M. R. Evans, G.-W. Peng, and H. E. Field *et al.* Epidemiologic clues to SARS origin in China. *Emerg. Inf. Dis.*, 10, 2004.

[188] Yang. Y., P. M. Atkinson, and D. Ettema. Analysis of CDC social control measures using an agent-based simulation of an influenza epidemic in a city. *BMC Inf. Dis.*, 11, 2011.

[189] M. Youssef and C. Scoglio. An individual-based approach to SIR epidemics in contact networks. *J. Theor. Bio.*, 283, 2011.

[190] J. Zhuang, A. E. Jetzt, G. Sun, H. Yu, and G. Klarmann *et al.* Human immunodeficiency virus type 1 recombination rate, fidelity and putative hot spots. *J. Virol.*, 76, 2002.

[191] D. Zinder, T. Bedford, S. Gupta, and M. Pascual. The roles of competition and mutation in shaping antigenic and genetic diversity in influenza. *PLoS Path.*, 9, 2013.