Introduction to Epidemiological Modeling: Basic Models and Their Properties

Alun Lloyd

March 16, 2007

Contents

1 Classification of Infections

In a typical infectious disease course, we might talk about different types of infectious disease agents separately (for instance viruses, bacteria, protozoa ... Then we would focus on different families of viruses, bacteria, ...) The discussion would focus on the details of infections.

Instead, we adopt a population biology viewpoint, as laid out by Anderson and May in their 1979 paper. Infectious agents are divided into two classes, microparasites and macroparasites, partly based on the type of model that one would need to describe the transmission of the infection.

Microparasites (e.g. viruses, bacteria, protozoa)

- Direct reproduction within host
- often of small size
- often have small generation time
- hosts often develop immunity to infection (possibly temporary)
- duration of infection is usually short compared to host lifespan
- which means typically transient infections

They claim that one doesn't need to account for the severity of infection in order to model transmission. This means we can divide the host population into a small number of classes, based on their infection status. (Susceptible/ infectious/ recovered). In this way, microparasite infections are naturally described by compartmental models.

They take this last statement as the operational definition of the term microparasite.

One point to make is that microparasite infections can be **directly** or **indirectly** transmitted via some intermediate host (e.g. mosquitoes in the case of malaria).

Macroparasites (e.g. parasitic helminths– which include nematodes, flukes and tapeworms– and arthropods)

- No (or slow) direct reproduction within the host (e.g. produce eggs and larvae which then pass into the environment)
- typically larger in size
- have a longer generation time
- the host's immune response tends to be short-lived (e.g. if we remove parasite by drug treatment, one isn't immune for very long)

- The number of parasites within a host is important: it influences (for example) rate of egg production, level of immune response.

For macroparasitic infections, we need to account for the distribution of parasites amongst hosts. This distribution is often highly clumped: a few individuals have a large parasite burden, most have a low burden. The average burden is a poor representation of the population.

The micro/macro parasite distinction leads to two quite different ways of formulating models.

1.1 How Useful is the Micro/Macroparasite Distinction?

Anderson and May present this as a deliberate oversimplification; one made with a view that emphasizes population biology aspects, rather than disease details.

They argue that one can start with a simple model description and then add in more details if the model is shown to be lacking in some way.

In many ways the micro/macro parasite distinction is unsatisfactory (e.g. for infections with a long and variable timecourse of infection, such as HIV), but it's a starting point.

2 Timecourse of a Microparasite Infection

As an example, we consider an acute viral infection.

We assume that infection (successful introduction of the virus) occurs at $t = 0$. Initially, the amount of virus within an individual (the virus load) grows exponentially. Over time, this growth slows (perhaps as the number of cells that the virus can infect drops or as an immune response builds up). The virus reaches a peak level before falling back, eventually going to zero (or a very low level).

Notice that even for a simple infection, including a detailed description of the timecourse of infection would be quite involved. Using a small number of disease states simplifies our model considerably. (Often assumptions are made on mathematical grounds as well as biological ones.)

- There is a time delay between acquisition of infection and the occurrence of disease symptoms: we call this delay the **incubation period**.
- There is also a time delay between a person becoming infected and their becoming infectious: we call this the **latent period**.
- The **generation time** is the sum of the latent and infectious periods.

- The **serial interval** is the time taken from the observation of symptoms in one individual to the observation of symptoms in a second individual who is directly infected by the first.

The incubation and latent periods may differ: the time periods over which a person shows symptoms and is infectious need not be the same.

This is an important point: we distinguish between **infection** and **disease**. Epidemiological data might often reflect the numbers of symptomatic individuals (the **prevalence of disease**), or the number of new cases of people exhibiting symptoms (the **incidence of disease**), whereas the model might be formulated in terms of the number of infectious individuals (**the prevalence of infectiousness**).

As far as a rapid response to an outbreak of an infection is concerned, any time-lag between individuals becoming infectious and showing symptoms can be significant in terms of attempts to minimize the spread of a disease.

You can have **subclinical infections** and **disease carriers**. It may be possible have an infection (and possibly transmit infection) even if the infection doesn't lead to symptoms. It may also be possible for people to be infected for long periods of time (or even for the rest of their lives) without showing symptoms.

3 Describing the Population

Our model must include some description of the population of interest. Each individual in the population is described by their **state**— those features that determine their potential to acquire and transmit infection.

At the very least, the state of an individual reflects their disease status, such as whether they are are susceptible to the disease, infectious or recovered from infection. Depending on the level of detail that our model employs, other features— such as age or spatial location– may also be required in order to specify the states of individuals.

- Many epidemiological models employ highly simplified descriptions of the disease process and so only distinguish between a small number of possible states.

The model must keep track of the states of each of its members. Two quite different approaches have been employed:

- **Population-level** models keep track of the total numbers of individuals in each state.

- **Individual-level** models explicitly model each individual, keeping track of the state of each member of the population.

Population-level approaches have tended to dominate the epidemiological literature. They often generate simple models, such as a set of coupled ordinary differential equations, which are amenable both to mathematical analysis and numerical simulation. Most of our attention will be focused on population-level models.

Individual-level models tend to be more complex and are often better suited for simulation than analysis. Since the model accounts for each individual as a separate entity, even simulation can be difficult: they can be quite computationally intensive (both in terms of time and memory). This is becoming less of an issue as computers become more powerful.

- Numerous computer packages are available to simplify the task of model simulation and the visualization of simulation results. For instance, MATLAB can be used to simulate population-level models that result in coupled ODEs, while SWARM (http://www.swarm.org) provides an individual-based simulation environment.

4 A Simple Compartmental Model: The SI model

As we mentioned earlier, compartmental Models subdivide the population into classes depending on their infection transmission status: for instance, individuals might be 'susceptible', 'infectious' or 'recovered'. As mentioned before, these classes need not correspond exactly to whether people are exhibiting disease symptoms.

Our first model assumes that

- individuals are either susceptible or infectious
- individuals immediately become infectious and never recover (or die).
- the population is **closed** (i.e. that it consists of a fixed number of individuals: births, deaths, immigration and emigration are ignored).

This situation is known as the **simple epidemic**.

Figure 1 illustrates the movement of individuals between infection classes in the SI model.

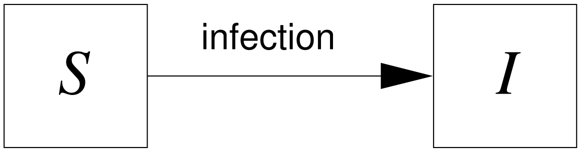

Figure 1: Flowchart showing movement between classes in the simple epidemic model.

If we treat the numbers of susceptible and infectious individuals as continuously varying quantities and employ a deterministic description of the movement between classes, we can describe this compartmental model (the SI model) by the following pair of coupled ordinary differential equations:

$$\dot{S} = -\lambda S \tag{1}$$
$$\dot{I} = \lambda S. \tag{2}$$

Here, S and I denote the number of susceptible and infectious individuals, respectively. The dots denote time derivatives (so that the right hand sides give the rates of change of S and I, respectively). The quantity λ is known as the **force of infection**: it is the (per-capita) rate at which susceptible individuals acquire infection.

The force of infection depends on:

- the rate at which individuals make contacts (the **contact rate**)
- the probability that a given encounter is with an infectious individual
- the probability that a given encounter leads to transmission of the infection (we call such contacts "effective contacts")

The simplest description of transmission employs the **mass action** assumption. ('Mass-action' originates in the kinetic theory of gases.) This assumes that the population is **well-mixed**— any two individuals are equally likely to encounter each other. The rate at which susceptibles acquire infection is proportional to both the density of susceptibles and infectives, leading to:

$$\dot{S} = -\beta SI/N \tag{3}$$
$$\dot{I} = \beta SI/N. \tag{4}$$

Here, N is the population size: I/N is the fraction of individuals that are infectious. β is known as the infection parameter.

It should be pointed out that some authors (notably, Anderson and May) take the infection term to be βSI. This assumption has been dubbed 'pseudo mass-action' and leads to some important differences. Most people consider the $\beta SI/N$ to be more appropriate. This issue has reappeared in the literature several times over the last few decades, and we shall discuss it in more detail later on.

A comment on different notations: Some authors (e.g. Anderson and May) use X, Y and Z in this model. Sometimes (but not always) lower case letters are used to denote fractions. The word 'density' is sometimes used for fraction and can lead to confusion. The use of R is problematic: we shall see later that the quantity R has another meaning in epidemiological theory.

4.1 Behavior of the Model

Since we assume that the population is closed, i.e. we have a fixed number of individuals, we have that $S + I = N$. This means that we can reduce the pair of differential equations (3) and (4) to the single equation

$$\dot{I} \quad = \quad \beta(N - I)I/N. \tag{5}$$

Writing $y = I/N$, we get

$$\dot{y} \quad = \quad \beta(1 - y)y. \tag{6}$$

This differential equation describes **logistic growth**, and has the explicit solution

$$y(t) = \frac{1}{1 + Ce^{-\beta t}}. \tag{7}$$

C is a constant whose value depends on the initial fraction of infectives. Typically, we would imagine that the initial fraction of infectives is low: most of the initial population would be susceptibles.

Logistic growth is well-known and understood: initially, the fraction of infectives increases exponentially, but the growth rate slows as the number of susceptibles drops. Eventually, everyone becomes infected.

The **epidemic curve** describes the rate at which new infectives appear and is given by given by dI/dt. In the SI model, there is a simple relationship between incidence and prevalence, obtained via equation 6.

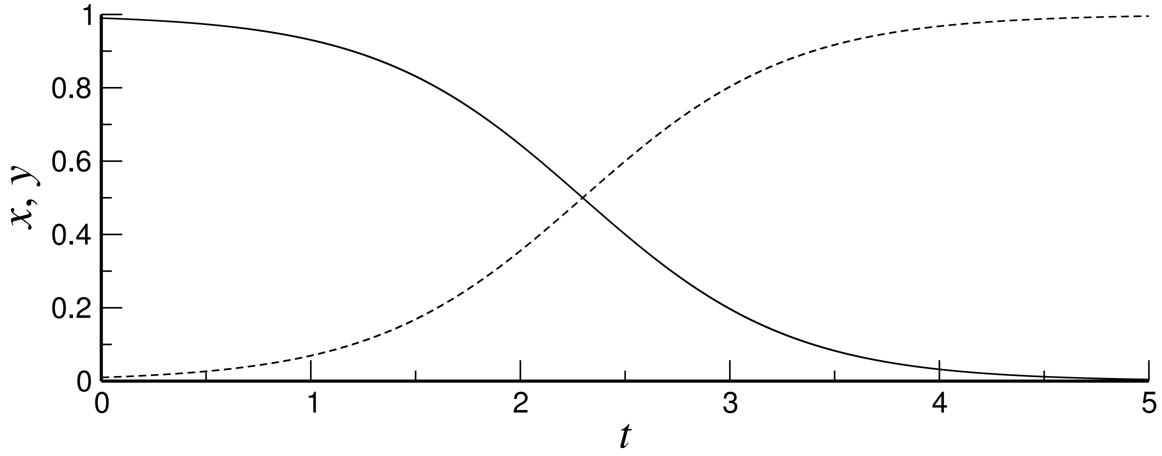

Figure 2: Dynamics of the simple epidemic. We plot the fraction of the population that is susceptible ($x = S/N$, solid line) and the infectious fraction ($y = I/N$, dashed line) against time. Notice that $x + y = 1$. We chose $\beta = 2$ day^{-1}.

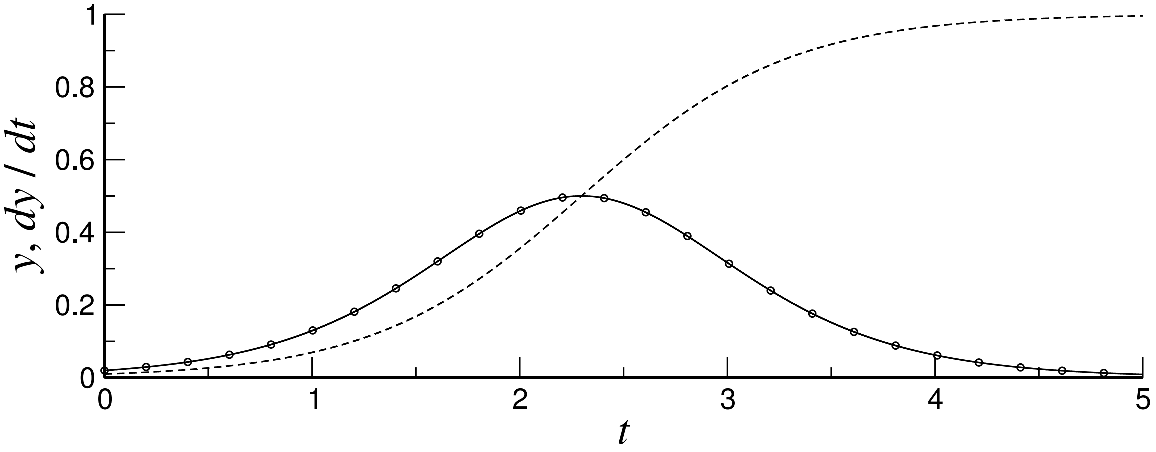

Figure 3: Epidemic curve for the simple epidemic. In addition to the infectious fraction ($y = I/N$, dashed line), we plot the rate of change of y (solid line with circles). dy/dt is proportional to the rate at which new cases appear (the incidence), while y tells us how many infectives there are (the prevalence). Parameter value as before.

We might ask when the entire population becomes infected. The value of I approaches N but never actually reaches it. This is because we have employed a continuous (or deterministic) description of the population. In reality, the number of infectives is a discrete (integer) quantity and so I would become equal to N at some point. We shall later discuss a class of models that account for the fact that the population is made up of individuals.

4.2 Qualitiative Analyses of Epidemic Models

The deterministic simple epidemic is one of the few instances in which we can derive an explicit solution to an epidemic model. In general, the **nonlinear** transmission term, $\beta SI/N$, is the one that makes the analysis non-trivial. (Seen another way, it is the term that makes the behavior of epidemic models interesting!). It is helpful to have alternate analysis techniques, which perhaps give **qualitative** information (what sorts of behaviors can occur, rather than giving exact numbers), as well as having an **intuitive** feeling for what is going on.

As an example, we often look for **equilibrium points**, where the numbers of individuals in each class remains constant over time. For the SI model, equation 6 shows that there are two steady states: $y = 0$ (no-one infected) and $y = 1$ (everyone infected).

Once we have found equilibrium points, we might ask if they are **stable** or **unstable**. It is easiest to think about **local stability**: if we are at an equilibrium and make a **small** change to the number of infectives, does y return to the equilibrium point or move away? It is often helpful to plot the **direction field** of the system. For the SI model, this is just a line with arrows showing whether y increases or decreases at any given point.

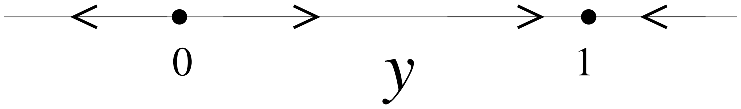

Figure 4: Direction field for the SI model. The arrows show the direction in which y moves: y will increase if it lies between 0 and 1.

Following the arrows, we see that $y = 0$ is an unstable equilibrium point: if we move a small distance away from zero, the arrows take us further and further away. On the other hand, $y = 1$ is a stable equilibrium point: if we move a small distance away then the arrows take us back to $y = 1$.

5 Describing Recovery from Infection and Disease Outbreaks: The SIR Model in a Closed Population

Typically, people do not remain infectious: they recover or die. We can model this by including a 'removed' class in the model, leading to an SIR model.

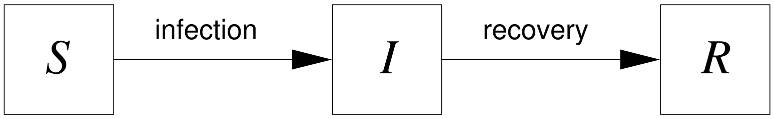

Figure 5: Flowchart showing movement between classes in the SIR model.

We have to describe the I to R transition in some way. The simplest assumption takes the recovery (removal) term to be proportional to the number of infective individuals:

$$\begin{aligned}\dot{S} &= -\beta SI/N &(8)\\ \dot{I} &= \beta SI/N - \gamma I &(9)\\ \dot{R} &= \gamma I. &(10)\end{aligned}$$

Again, we consider a closed population, so $S + I + R = N$. We usually consider the initial number of susceptibles to be close to N.

This model is often called the Kermack and McKendrick model as it appeared in their 1927 paper. It is also called the **general epidemic** model. (Although this SIR model is often called THE Kermack and McKendrick model, it has been pointed out that the 1927 paper goes beyond this model, discussing a more general framework that employs fewer assumptions.)

It's worth pausing to think about the assumption made regarding the recovery term. Having a constant recovery rate means that the distribution of infectious periods is exponential with mean $1/\gamma$. Biologically, this assumption corresponds to the chance of recovery being independent of the time since infection. In most cases this is far from realistic, but it considerably simplifies the formulation of the model. (Otherwise, one needs to keep track of when each infective acquired the infection. We shall return to this complication later in the course.) Mathematically, this simplification results from the **memoryless property** of exponential distributions or Poisson processes.

In the SI model we saw that there was a simple relationship between incidence and prevalence, but this is not the case for the SIR model. The rate of change of the number of infectives depends on the difference between the rate at which new infections arise (the incidence) and the rate at which individuals recover. In the special case when the number of infectives is constant there is a linear relationship between prevalence and incidence.

5.1 Behavior of the SIR Model

For the SIR model, the equation for the rate of change of S shows that the only equilibrium points correspond to an entirely susceptible population or an entirely recovered population.

Numerically generated solutions of the model show two sorts of behavior, illustrated in figure 6.

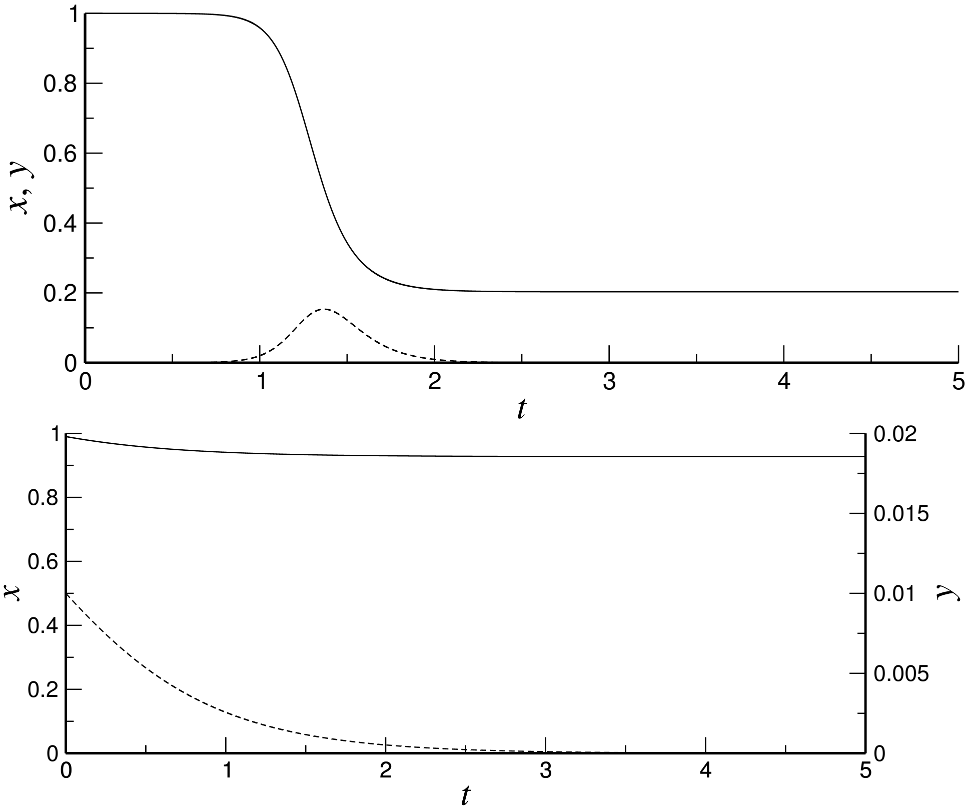

Figure 6: Dynamics of the SIR model in a closed population. (a) Self limiting epidemic, (b) no epidemic. In both cases we plot the fraction of the population that is susceptible, x, (solid line) and the fraction that is infective, y, (broken line).

We either see an epidemic, which appears to be self-limiting, or else we see an exponential decay in the number of infectives.

We can't solve the SIR model explicitly for S and I in terms of time, but we can get a good feel for what is going on by looking at the equation for the rate of change of I:

$$\dot{I} \;=\; I(\beta S/N - \gamma). \tag{11}$$

The number of infectives increases or decreases, according to whether $\beta S/N$ is greater than or less than γ. The (instantaneous) per-capita growth rate of the infective population is $\beta S/N - \gamma$. The rate at which each infective gives rise to new infectives is given by $\beta S/N$, while the rate at which each infective recovers is γ.

5.2 Initial Behavior of the SIR Model

During the initial period (before an epidemic– if it ever happens– has had time to take off), the number of susceptibles will be close to N and so the initial growth rate, which we denote by r, is given by $\beta - \gamma$. This relationship between the initial growth rate of the SIR model and the epidemiological parameters is often written in the following way:

$$\begin{aligned} r &= \beta - \gamma \\ &= \gamma(R_0 - 1), \end{aligned} \tag{12}$$

where R_0 is given by β/γ.

The parameter combination R_0 is called the **basic reproductive number** of the infection. An epidemic can only take off if R_0 is greater than one: in this case we initially see exponential growth in the number of infectives. Otherwise, we see exponential decay in I.

R_0 has an intuitive biological interpretation: it is the average number of secondary infections caused by the introduction of a single infectious individual into an otherwise totally susceptible population (over the course of their infection).

During the initial phase of the epidemic (when $S \approx N$), the rate at which each infective gives rise to new infections is given by β. The average duration of infectiousness (which we write as τ) is $1/\gamma$, so the average number of secondary infections is the product of β and $1/\gamma$.

Notice that equation (12) provides a simple relationship between r, the initial growth rate of the epidemic, and R_0. This has been used to estimate the value of R_0 from epidemic data. (We will come back to this later.)

The initial phase of the epidemic is easily understood because the system is (to a very good approximation) **linear**. The assumption that $S \approx N$ turns the nonlinear transmission term $\beta SI/N$ into a linear one, βI. This is our first example of **linearization**, a mathematical technique that is frequently used to simplify analysis.

5.3 Behavior as the Epidemic Progresses

Assuming that $R_0 > 1$, what happens as the epidemic proceeds? The number of susceptibles drops and so the rate at which new infections arises decreases. Eventually, S drops far enough (more precisely, below N/R_0) that the rate at which individuals recover exceeds the rate at which new infections occur. Consequently, the number of infectives falls (see figure 6 and equation 11). The epidemic is self-limiting because the susceptible population is depleted, although it does not (as we shall confirm below) approach zero.

Notice that, because the infection term is proportional to the fraction of the population that remains susceptible, the average number of secondary infections at any point in time (often known as the general reproductive number, and denoted by R_t) is given by the product of R_0 and the susceptible fraction. So we have

$$R_t = R_0(S/N). \tag{13}$$

Notice that the general reproductive number falls off linearly as S decreases. We can recast equation 11 in the following way:

$$\begin{aligned} \dot{I} &= I\left(\frac{\beta S}{N} - \gamma\right) \\ &= \gamma I\left(\frac{\beta S}{\gamma N} - 1\right) \\ &= \gamma I(R_t - 1). \end{aligned} \tag{14}$$

We see that the number of infectives increases if $R_t > 1$ and decreases if $R_t < 1$. Figure 6 can be reinterpreted in terms of R_t. When the infection is introduced into an entirely susceptible population, we have $R_t = R_0$ and so R_t is greater than one (since we assumed $R_0 > 1$). As the epidemic proceeds, the number of susceptibles falls and so R_t decreases (figure 7). At some point, S decreases by enough that R_t passes through one and so the number of infectives falls off. Notice that the peak prevalence of infection over the epidemic coincides with R_t equalling one.

The behavior of the model can also be depicted graphically by plotting $I(t)$ against $S(t)$ (or, in terms of fractions, $y(t)$ against $x(t)$). This is known as a **phase plot**: the curves denote trajectories of the system, with different curves corresponding to different initial conditions. One could put arrows on these curves to denote the direction in which one moves as time passes.

5.4 Analysis: The Size of an Epidemic

As we mentioned, we cannot solve explicitly for S and I in terms of time. But we can make some progress by eliminating time. This will lead to an expression for the **total size** of the

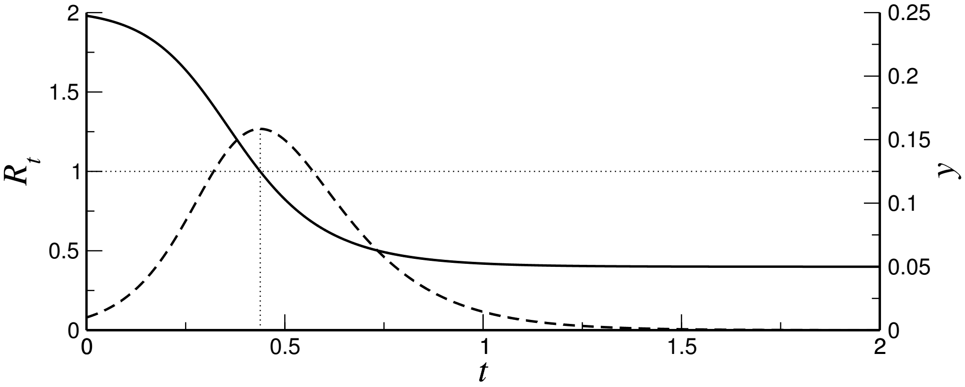

Figure 7: The dynamics of the SIR model in terms of the general reproductive number R_t, assuming that $R_0 > 1$. The solid line depicts R_t and the dashed line shows the infective fraction, y. The light dotted lines highlight the maximum point that occurs when $R_t = 1$.

epidemic: that is, the fraction of individuals who experience infection over the entire timecourse of the epidemic. Remember that at large times, the number of individuals who are infectious goes to zero: everyone is then either susceptible (i.e. have never experienced the infection) or recovered (i.e. have experienced the infection at some point).

Dividing the equations for $\dot{S}$ by the equation for $\dot{I}$, and making use of the Chain Rule to show that $(dS/dt)/(dI/dt) = dS/dI$, we have

$$\begin{aligned}
\frac{\dot{S}}{\dot{I}} &= \frac{-\beta SI/N}{\beta SI/N - \gamma I} \\
&= \frac{-1}{1 - \gamma N/(\beta S)} \\
\frac{dS}{dI} &= \frac{-1}{1 - N/(R_0 S)}. \qquad (15)
\end{aligned}$$

(We get the right hand side of the second line by dividing top and bottom of the fraction by $\beta SI/N$ and the third line by substituting $R_0 = \beta/\gamma$.)

Multiplying both sides by $1 - N/(R_0 S)$ gives

$$\left(1 - \frac{N}{R_0 S}\right)\frac{dS}{dI} = -1. \qquad (16)$$

We integrate both sides of this equation with respect to I, noting that the dS/dI on the left hand side means we end up integrating with respect to S.

$$\int \left(1 - \frac{N}{R_0 S}\right) dS = \int -1 dI, \qquad (17)$$

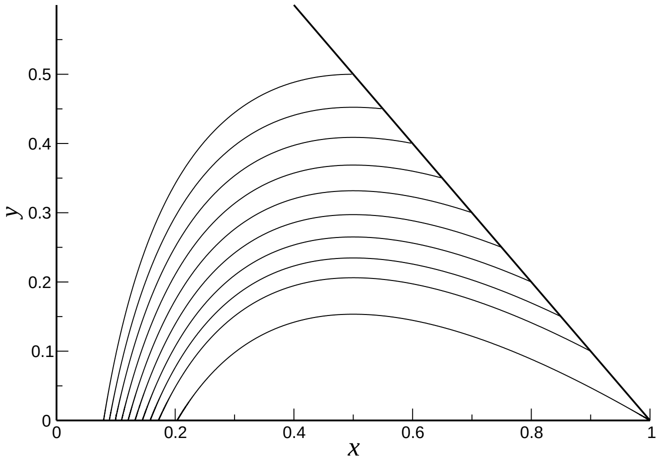

Figure 8: Plots of y versus x for the SIR model, starting at different initial conditions. The straight line represents $x + y = 1$. Our initial conditions will be on this line if we imagine that we start with a population that consists of just susceptibles and infectives (no recovereds at first).

which gives

$$S - \frac{N}{R_0} \ln S = -I + c, \tag{18}$$

where c is the constant of integration.

This gives a relationship between S and I that holds as an epidemic (started from some set of initial conditions) proceeds. Taking different initial conditions (i.e. different values of c) we get a family of curves traced out in the (S, I) or, in terms of fractions, (x, y) plane. See figure 8.

Equation (18) means that the quantity $S - \frac{N}{R_0} \ln S + I$ is constant along each of these curves. This is what is known as a **conserved quantity**.

We can use this conserved quantity to find the size of an epidemic. We equate its value at the start of the epidemic to its value after the epidemic has passed (i.e. as t approaches infinity). Initially, almost the entire population would be susceptible and there would be a small number of infectives. We take the limiting case, namely that $S = N$ and $I = 0$ (of course, we would need there to be some infectives in the population...) At large times, the number of infectives approaches zero. On the figure above, this corresponds to following the bottom-most curve from right to left.

Equating the value of the conserved quantity at $t = 0$ and at large times (i.e. $t = \infty$) gives

$$N - \frac{N}{R_0}\ln N \quad = \quad S(\infty) - \frac{N}{R_0}\ln S(\infty). \tag{19}$$

Rearranging gives

$$N - S(\infty) \quad = \quad \frac{N}{R_0}\ln\frac{N}{S(\infty)}. \tag{20}$$

Multiply both sides of this equation by $-R_0/N$ and (for the right hand side) make use of the fact that $-\ln x = \ln(1/x)$:

$$-R_0\left(1 - \frac{S(\infty)}{N}\right) \quad = \quad \ln\frac{S(\infty)}{N}. \tag{21}$$

Taking e to the power of both sides (to remove the logarithm) gives

$$\exp\left(-R_0\left\{1 - \frac{S(\infty)}{N}\right\}\right) \quad = \quad \frac{S(\infty)}{N}. \tag{22}$$

As we mentioned above, at the end of the epidemic we only have recovered individuals and susceptible individuals. So $S(\infty)/N$ represents the fraction of the population who escaped infection. One minus this quantity denotes the fraction of the population who did not escape infection, i.e. the epidemic size. Denoting the epidemic size by y_{total}, we finally have that

$$\exp\left(-R_0\, y_{\text{total}}\right) \quad = \quad 1 - y_{\text{total}}. \tag{23}$$

This **transcendental equation** relates the total size of the epidemic to the basic reproductive number. Unfortunately, there is no simple explicit expression for y_{total} in terms of R_0. We can solve this equation numerically (using, for instance, matlab or maple). Alternatively, figure 9 illustrates a graphical method in which we plot curves depicting the two sides of the equation and look for points of intersection.

When the basic reproductive number is less than one, the only solution has a zero epidemic size. When $R_0 > 1$, there is a positive solution, corresponding to an outbreak of the infection.

Figure 10 shows a plot of the epidemic size as a function of R_0.

5.5 Comparing the Model to Reality

This model can be fitted to real-world data. Kermack and McKendrick (1927) compared it to data from an epidemic of plague in Bombay in 1905/6. They argue that nearly all infected people died

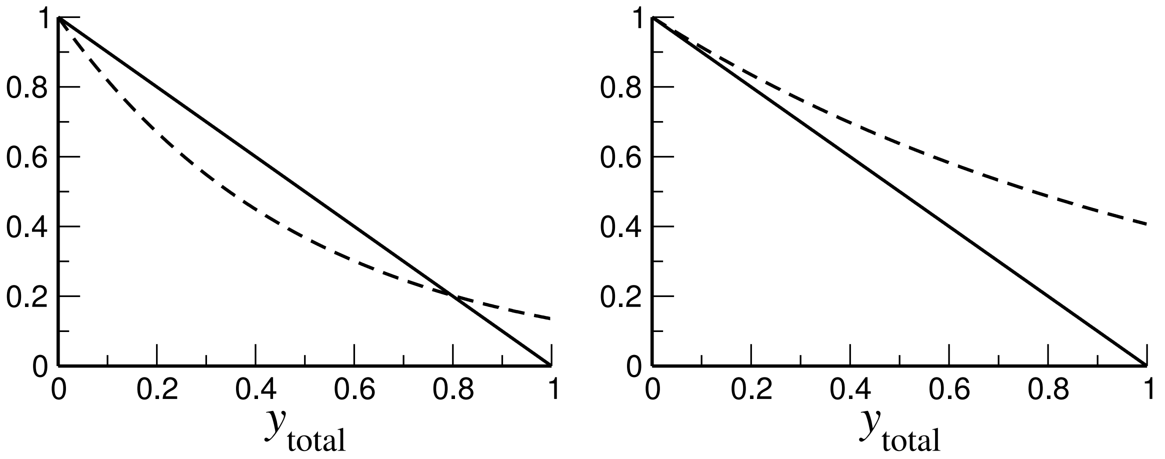

Figure 9: Plots of $1 - y_{\text{total}}$ (solid line) and $\exp(-R_0\, y_{\text{total}})$ (dashed line) versus y_{total}. The left panel is for a case in which $R_0 > 1$ and the right panel for a case when $R_0 < 1$.

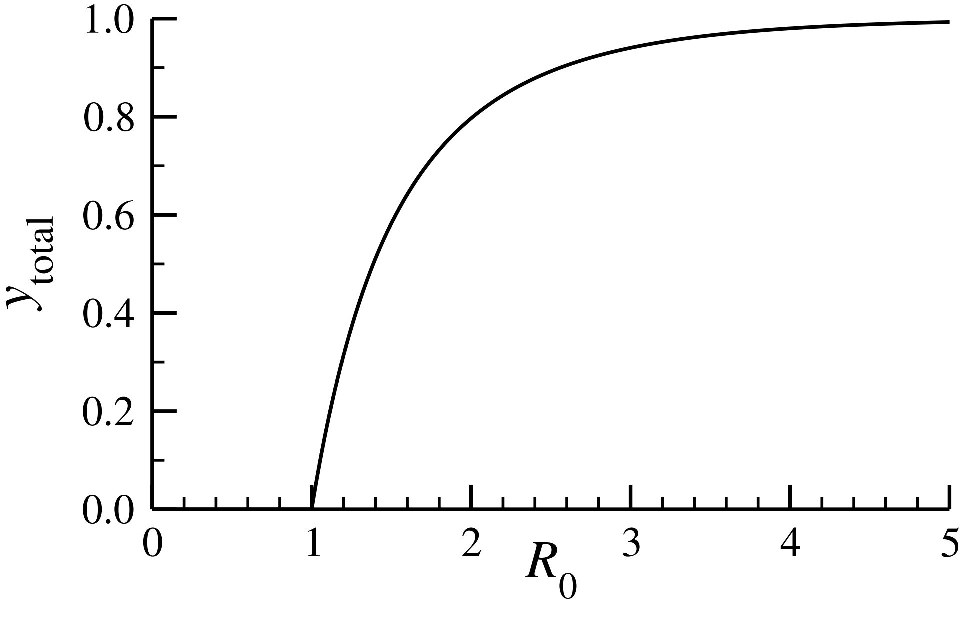

Figure 10: Plot of final epidemic size as a function of R_0.

in the epidemic, so that the equation for the rate of change of removals in fact tells us the death rate.

Dividing the $\dot{S}$ equation by the $\dot{R}$ equation, and using the Chain Rule, tells us that $dS/dR = -R_0S/N$. This equation describes an exponential decay in S as R increases: $S = S(0)\exp(-R_0R/N)$. (We are more used to seeing such a differential equation in terms of the independent variable t, rather than R.)

The equation for dR/dt is then

$$\begin{aligned} \frac{dR}{dt} &= \gamma I \\ &= \gamma(N - S - R) \\ &= \gamma(N - S(0)\exp(-R_0R/N) - R). \end{aligned} \tag{24}$$

So we have an equation for the rate of change of R in terms of R alone. But we can't solve this directly, mainly because we have an R appearing in an exponential term.

What Kermack and McKendrick did was assume that the quantity R_0R/N is small and expanded the exponential term (see next section for brief details). This assumption corresponds to the epidemic being small. The resulting ODE has an exact solution in terms of the hyperbolic secant function.

5.5.1 Small Outbreak Analysis *

Expanding the exponential in equation 24 in terms of the small quantity R_0R/N gives

$$e^{-R_0R/N} \approx 1 - \frac{R_0R}{N} + \frac{1}{2}\left(\frac{R_0R}{N}\right)^2. \tag{25}$$

Substituting this into the differential equation (equation 24) for R gives

$$\frac{dR}{dt} = \gamma\left\{N - S(0) + \left(\frac{R_0S(0)}{N} - 1\right)R - \left(\frac{S(0)R_0^2}{2N^2}\right)R^2\right\}. \tag{26}$$

This differential equation can be solved, making use of the (not so well-known) result that the solution of the differential equation $dx/dt = k(a^2 - x^2)$ is $x(t) = a\tanh(akt + C)$, where C is the constant of integration and tanh is the hyperbolic tangent function. In order to make use of this result, the right hand side of (26) must be rearranged to remove the linear term in R. This is achieved by completing the square, rewriting the right hand side in the form $\gamma\left\{A + D(R - B)^2\right\}$, where A, B and D are appropriate constants. After some unpleasant algebra, the following solution is obtained

$$R(t) = \frac{N^2}{S(0)R_0^2}\left\{\left(\frac{S(0)R_0}{N} - 1\right) + \alpha \tanh\left(\frac{\alpha\gamma t}{2} - \phi\right)\right\}, \tag{27}$$

where

$$\alpha = \left\{\left(\frac{S(0)R_0}{N} - 1\right)^2 + \frac{2S(0)(N - S(0))R_0^2}{N^2}\right\}^{1/2}, \tag{28}$$

$$\phi = \tanh^{-1}\left(\frac{1}{\alpha}\left\{\frac{S(0)R_0}{N} - 1\right\}\right). \tag{29}$$

Since the number of deaths per week was of interest, the relevant model quantity is not R, but its rate of change, dR/dt. We differentiate the above solution to get

$$\frac{dR}{dt} = \frac{\gamma\alpha^2 N^2}{2S_0 R_0^2}\,\mathrm{sech}^2\left(\frac{\alpha\gamma t}{2} - \phi\right). \tag{30}$$

This involves three parameter combinations, $(\gamma\alpha^2N^2)/(2S_0R_0^2)$, $\alpha\gamma/2$ and ϕ.

Remark: if you have Murray's 'Mathematical Biology', which contains an account of this analysis... there are a few typos in the new edition (these may also be present in the old edition). His equation for $R(t)$ has a factor r^2 at the front: this should be a ρ^2. In his equation for ϕ, the factor $1/\alpha$ should be inside the $\tanh^{-1}$ function.

The book by Daley and Gani includes a more detailed description of the analysis, but the step in which one completes the square is still somewhat unclear: at least to my mind, it appears to come out of nowhere. Unfortunately, some mathematical analyses look more like conjuring tricks than one would like ...

5.5.2 Fitting to Data from the Bombay Outbreak

Kermack and McKendrick compared the behavior of the model to data from an outbreak of plague in Bombay. They argued that almost everyone who became infected died, and so 'removals' in the model correspond to deaths, rather than recoveries. Since the dataset gave the number of deaths per week, the relevant model quantity to look at is dR/dt. The analysis of the previous section gives an expression for dR/dt as a function of time in terms of three parameters (or parameter combinations).

Kermack and McKendrick fitted these three parameter combinations using their dataset. Notice that we might be able to do a little better than this: if we had some information about the biology of the infection, we might be able to provide estimates of some parameters (e.g. the average duration of infectiousness) without resorting to such a fitting procedure.

So there are three parameters to estimate. How they did this was unclear to me from their paper. It also wasn't clear to me if the epidemic was indeed 'small'.

5.5.3 Another Model Fitting Example

Murray's 'Mathematical Biology' book describes the fitting of the SIR model to an influenza outbreak in a boarding school in England. Numbers of infectives are shown, together with a best-fit SIR model. When they became ill, boys were confined to bed.

How was the fitting done? Again, this is not entirely clear. Probably by integrating the SIR model from the given initial conditions, varying parameters, until the error between the data and model was made as small as possible. ('Least squares fit'?) Least squares is commonly used, but is not a particularly sophisticated approach.

6 Persistence of Infection: The SIR Model with Demography and Endemic Infections

In order for infections to persist, the susceptible population must be replenished. This could be due to loss of immunity (recovered individuals become susceptible after some time– as in the so-called SIRS model) or by births.

We make a simple set of assumptions:

- there is a constant per-capita birth rate μ (this means that rate at which susceptibles enter the population is μN).
- the infection is non-fatal (that means that death rates are independent of disease status)
- the death rate is constant and equals the birth rate (this means that the population size will remain constant).

The constant death rate assumption (which echoes the assumption we made about recovery) means that the distribution of lifespans is exponential, with average lifespan, L, equal to $1/\mu$. This assumption is known as type II mortality. (This distribution is unlikely to be realistic for populations in developed countries, where lifespans tend to be more closely distributed around the average. This is better modeled by what is known as type I mortality, in which it is assumed that everyone has an identical lifespan.)

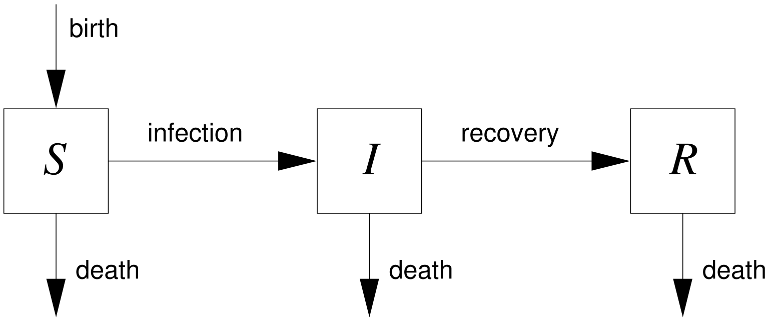

Figure 11: Flowchart showing movement between classes in the SIR model with demography.

The model can be described by the following set of differential equations

$$\dot{S} = \mu N - \beta SI/N - \mu S \quad (31)$$
$$\dot{I} = \beta SI/N - (\gamma + \mu)I \quad (32)$$
$$\dot{R} = \gamma I - \mu R, \quad (33)$$

with $S + I + R = N$. Again notice that our assumption of a constant population size helps us to simplify the model because $R = N - S - I$: we need only keep track of S and I.

The average duration of infectiousness, τ, is given by $1/(\gamma+\mu)$. Notice that this expression accounts for the background mortality (some infectious individuals die before recovering) and so its value is slightly lower than in the earlier model. If the death rate is small compared to the recovery rate, the average duration of infection differs only slightly from $1/\gamma$.

6.1 Behavior of the Model

To find R_0 for this model we can employ the same argument that we used for the SIR epidemic in a closed population. Since the infection term is $\beta SI/N$, the rate at which new infections arise in an entirely susceptible population ($S = N$) is βI, or β per infectious individual. Since the average duration of infectiousness is $1/(\gamma + \mu)$, the average number of secondary infections due to a single infectious individual in an otherwise entirely susceptible population is $\beta/(\gamma + \mu)$. This is very similar to the earlier expression, with the minor modification due to the reduced average duration of infectiousness.

Just as in the closed population SIR model, $R_0 = 1$ determines a threshold condition for the invasion of the infection into an entirely susceptible population.

6.2 Equilibrium Behavior

At an equilibrium, the levels of S, I and R remain constant: their time derivatives equal zero. We sometimes denote equilibrium values with asterisks: (S^*, I^*, R^*). The most direct way to find these equilibrium points is to set the right hand sides of equations 31-33 equal to zero and solve for S, I and R.

The right hand side of equation 32, for the rate of change of I, shows us that either $I = 0$ or $\beta S/N = (\mu + \gamma)$.

- If $I = 0$, then the right hand side of equation 31 tells us that S equals N (and therefore that $R = 0$). This equilibrium has the entire population susceptible: it is the 'infection-free equilibrium'.
- Alternatively, we have $\beta S/N = (\mu+\gamma)$, which can be rearranged as $S/N = 1/R_0$ or $S = N/R_0$. At this equilibrium point, the fraction of the population that is susceptible is given by the reciprocal of the basic reproductive number. Substituting into the right hand side of equation 31 gives $0 = \mu N - \beta I/R_0 - \mu N/R_0$. Rearranging gives an expression for the equilibrium number of infectives: $I = \mu N(R_0 - 1)/\beta$. This can also be written as $\mu N(1 - 1/R_0)/(\mu + \gamma)$

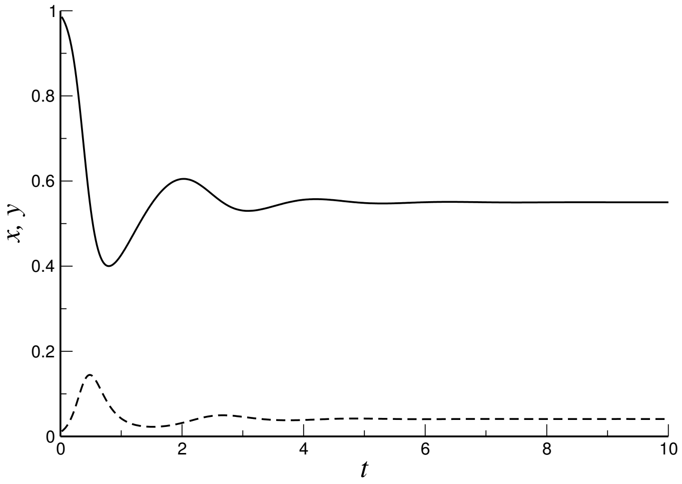

Figure 12: Susceptible and infective fractions (x and y) in the SIR model with demography in the case when $R_0 > 1$. Notice that the number of susceptibles recovers after the first outbreak as births replenish the susceptible pool. This enables the number of infectives to again increase. Notice that whenever the number of infectives hits a maximum or minimum, R_t must equal one. At these points we have $dI/dt = 0$. These points do not coincide with the peaks and troughs in the number of susceptibles ($dS/dt = 0$). The only places where **both** dS/dt and dI/dt are zero are the equilibrium points.

> or $N(1 - 1/R_0)(\tau/L)$. (Recall that τ is the average duration of infectiousness and L the average lifespan. Notice that, because we have relationships like $L = 1/\mu$ we can often express quantities such as equilibrium levels in several equivalent ways. At first sight, some of these might not appear to be equivalent!)

The first of these equilibria, $(N, 0, 0)$, is also an equilibrium point of the SIR model in a closed population. (Recall that in the closed population model, there was more than one infection-free equilibrium: any point of the form $(S, 0, N - S)$ is an equilibrium point.)

The second equilibrium is new, having no counterpart in the closed population model. Notice that this equilibrium only makes sense biologically when R_0 is greater than one. (When R_0 is less than one, S is greater than N and I is negative.) This equilibrium corresponds to infection being maintained in the population at some level: we call it the ‘endemic equilibrium’. The condition $R_0 > 1$ is needed in order for this equilibrium to exist (in a biologically meaningful way).

When R_0 is greater than one, typical trajectories of the model approach the endemic equilibrium, as shown in figure 12.

Figure 13 shows the phase portraits of the system in the $R_0 < 1$ and $R_0 > 1$ cases. This figure should be compared to figure 8.

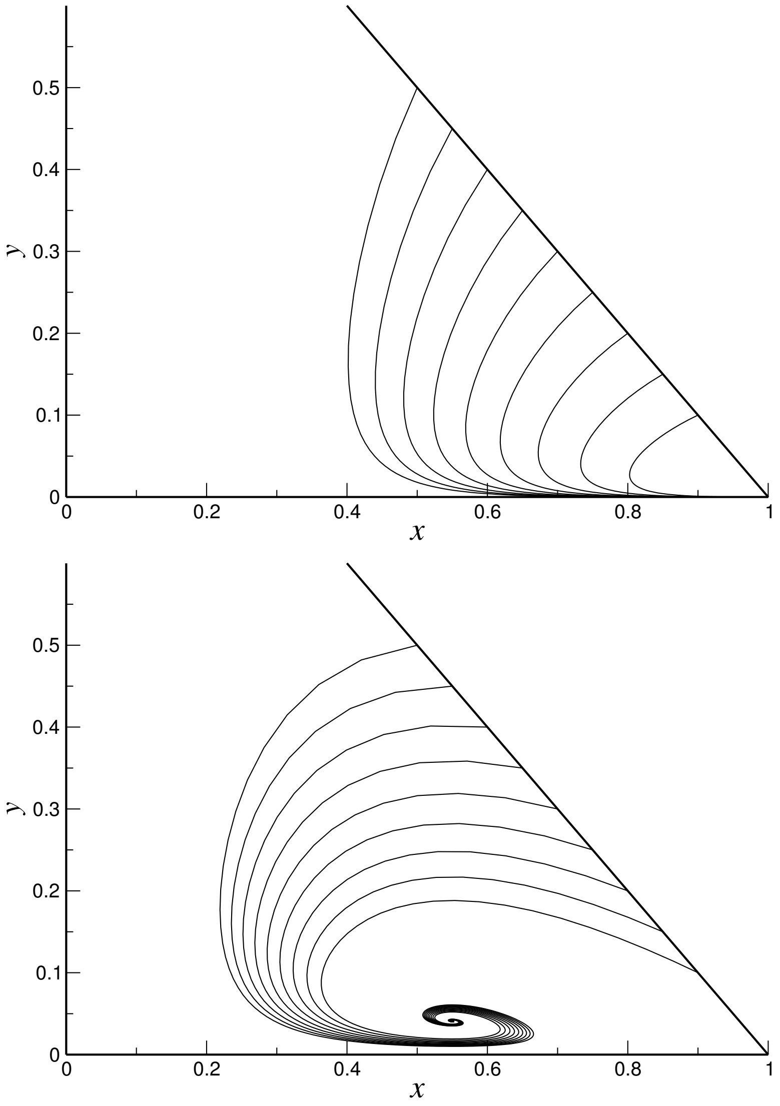

Figure 13: Phase portraits in the SIR model with demography when $R_0 < 1$ (upper figure) and $R_0 > 1$ (lower figure). The below-threshold behavior exhibits a stable infection free equilibrium: births lead to the population returning to the state with everyone susceptible. The above-threshold behavior shows the unstable infection free equilibrium: as soon as infection is introduced, the system moves away from the state with everyone susceptible. Trajectories then spiral towards the endemic equilibrium.

At the endemic equilibrium, the rate at which individuals move into the S class (due to births) balances the rate at which individuals leave the S class (due to infection and deaths). The rate at which individuals move into the I class (due to infection) balances the rate at which individuals leave the I class (due to recovery and deaths). Notice that each of these processes is still going on (so that individuals are still moving between classes) but the rates of entry and exit balance. It might take some time to reach this state... there may be some **transient** behavior.

We can get some intuitive feeling for what is going on at the endemic equilibrium by considering the number of infectives. At an equilibrium, this number neither goes up nor goes down, so the general reproductive number must equal one: each infectious individual leads to exactly one secondary infection. Since we have that $R_t = R_0(S/N)$, we immediately see that $R_t = 1$ implies $S = N/R_0$. (Notice that this does not give us the complete picture: we can have R_t equal to one but not be at an equilibrium. In order for the system to be at an equilibrium, the rate of change of S must also be zero.)

At the endemic equilibrium, the fraction of the population that can be found in the susceptible class is $1/R_0$. Looked at from an individual's viewpoint, this means that, on average, each individual spends $1/R_0$ of their lifespan in the susceptible class. If (as defined earlier) the average lifespan is L years, this means the average age at which they leave the susceptible class is L/R_0 years. If we assume that most individuals leave the susceptible class because they acquire infection (rather than because they die), we see that the average age at which individuals acquire infection, denoted by A, is equal to L/R_0. (This argument can be extended to account for the chance of susceptibles dying, but we do not pursue this here. See, for instance, the discussion on page 45 of Diekmann and Heesterbeek.)

The important point that emerges from this analysis is that $R_0 = 1$ defines a threshold condition for both invasion of the infection and for persistence of the infection (i.e. there being an endemic equilibrium).

6.3 Equilibrium Stability

We use local stability analysis to investigate whether the system moves towards (stable) or away from (unstable) a given equilibrium point. Local stability of an equilibrium can be assessed using standard techniques.

The Jacobian matrix (i.e. the matrix of partial derivatives) is given by

$$J = \begin{pmatrix} -\mu - \beta I/N & -\beta S/N \\ \beta I/N & \beta S/N - (\mu + \gamma) \end{pmatrix}. \tag{34}$$

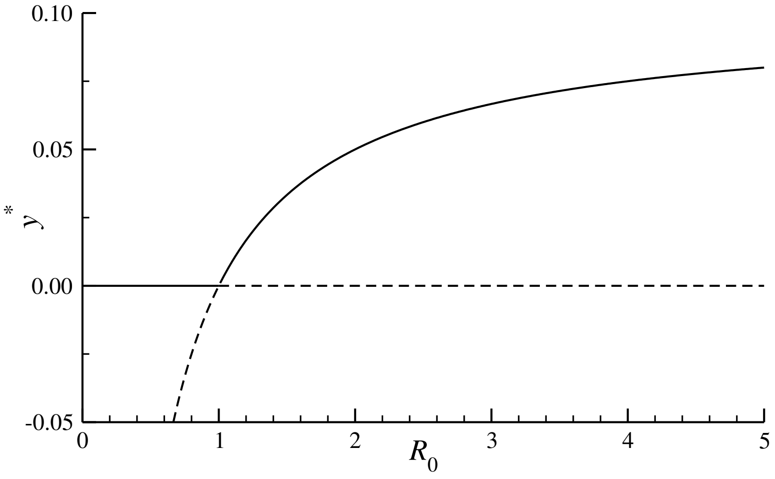

Figure 14: Fraction of the population that is infectious at the infection free and endemic equilibria. The linestyle denotes the stability (solid=stable, broken=unstable) of the equilibrium.

Notice that we only need to consider the 2-dimensional (S, I) system since we can eliminate R using $R = N - S - I$. (Otherwise, we would have to examine a 3×3 matrix: the constant population size assumption makes our life a lot simpler!)

For the infection-free equilibrium, substituting $(S, I) = (N, 0)$, we have

$$J = \begin{pmatrix} -\mu & -\beta \\ 0 & \beta - (\mu + \gamma) \end{pmatrix}. \tag{35}$$

The eigenvalues of this matrix are particularly easy to calculate. A standard result tells us that if one of the off-diagonal elements of a 2×2 matrix is zero, then the eigenvalues are just given by the entries found on the leading diagonal. The eigenvalues are $-\mu$ and $\beta - (\mu + \gamma)$. The second of these can be written as $(\mu + \gamma)(R_0 - 1)$. We see that this equilibrium is a stable node (both eigenvalues are negative) when R_0 is less than one, but is a saddle (one positive and one negative eigenvalue) when R_0 is greater than one. The second eigenvalue is zero when R_0 equals one.

There is a structural reason why there cannot be oscillations around this equilibrium: the value of I cannot become negative.

For the endemic equilibrium, one can show that

$$J = \begin{pmatrix} -\mu R_0 & -\beta / R_0 \\ \mu(R_0 - 1) & 0 \end{pmatrix}. \tag{36}$$

Notice that the trace of this matrix is $-\mu R_0$ and that the determinant is equal to $\beta\mu(R_0 - 1)/R_0$. The trace is negative and the sign of the determinant depends on whether R_0 is greater than one or less than one. Thus the stability of the endemic equilibrium is determined by R_0 being greater than or less than one. When R_0 is greater than one, we can easily see that the approach to the equilibrium is oscillatory (i.e. the eigenvalues are complex) unless R_0 is close to one. (This argument is fleshed out in the following section.)

This analysis completes the discussion of the equilibria of the SIR model that accompanied figure (14). When R_0 passes through one, the model undergoes a qualitative change of behavior: such changes are known as **bifurcations**. The two equilibrium points change stability: one becomes unstable and the other becomes stable: this is known as a **transcritical bifurcation** or **exchange of stability** bifurcation. This bifurcation is an example of a **local bifurcation**: behavior is only affected in the neighborhood of the equilibria. A result from nonlinear dynamics shows that local bifurcations involve situations when the real part of an eigenvalue of the linearization equals zero. Looking back at our eigenvalue expressions, we see that this is indeed the case when R_0 equals one.

6.3.1 A Closer Look at the Eigenvalues

After a little algebra, one obtains the following quadratic for the eigenvalues of the Jacobian matrix (36):

$$\lambda^2 + \mu R_0 \lambda + \mu\beta(1 - 1/R_0) = 0. \tag{37}$$

This can be simplified to

$$\lambda^2 + \frac{1}{A}\lambda + \frac{\mu(R_0 - 1)}{\tau} = 0, \tag{38}$$

where A is the average age at infection and τ is the average duration of infectiousness (corrected for mortality). (Recall $R_0 = \beta\tau$.)

This is quite messy to work with, so we consider a special case, which applies to many of the **childhood diseases** which can be modeled using the SIR model. For many of these diseases, there is a separation of timescales between the duration of infectiousness (lasts on the order of days), the average age at infection (on the order of a few years) and the average lifespan (on the order of tens of years). Such diseases have large values of R_0 (corresponding to a separation between mean age at infection and the average lifespan) and so we can approximate $R_0 - 1$ by R_0. Using the relationship $R_0 = L/A$, where L is the average lifespan (which equals $1/\mu$), this then leads to

$$\lambda^2 + \frac{1}{A}\lambda + \frac{1}{A\tau} = 0. \tag{39}$$

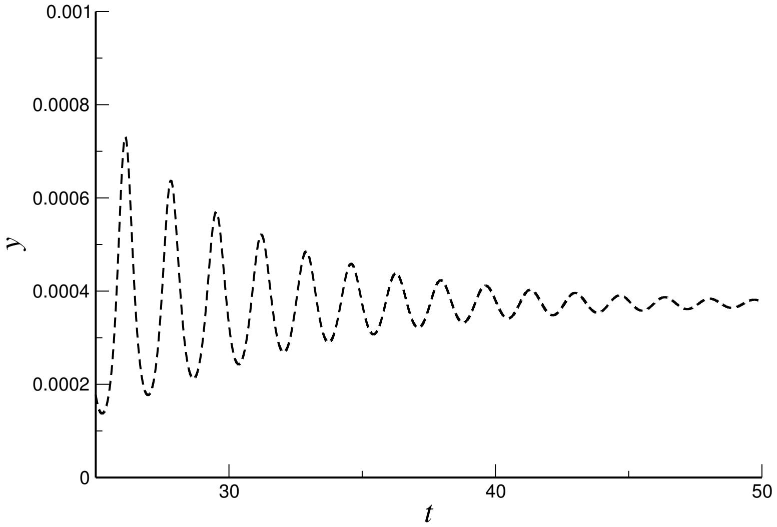

Figure 15: Approach of the SIR model to its endemic equilibrium. We plot the variable y, the fraction of the population that is infectious. Parameter values were chosen to be appropriate for measles: $\mu = 0.02$ year^{-1}, $\gamma = 50$ year^{-1} and $\beta = 750$year^{-1}.

This quadratic has roots

$$\lambda = -\frac{1}{2A} \pm \frac{1}{2}\sqrt{\frac{1}{A^2} - \frac{4}{A\tau}}. \tag{40}$$

Since $A >> \tau$, we have that

$$\lambda \approx -\frac{1}{2A} \pm \frac{i}{\sqrt{A\tau}}. \tag{41}$$

Thus the equilibrium point is an attracting spiral, and so we see damped oscillations towards the endemic equilibrium. The relaxation time is $2A$ and the period of oscillation $2\pi\sqrt{A\tau}$. Given the separation of timescales, the relaxation time is much longer than the period of oscillation, so these are weakly damped oscillations (see figure 15).

6.4 Behavior Away From the Equilibrium

We might wonder how well the linearized model describes the behavior of the model when we are not close to the equilibrium. Let's think about introducing a small number of infectives into an otherwise susceptible population and follow what happens.

The initial behavior of the SIR model with demography closely follows that of the SIR model for the corresponding closed population. This should not be surprising since the demographic processes operate on a slow timescale (on the order of years). At first, we see a single epidemic accompanied by a marked depletion in the number of susceptibles (figure 16).

Looking over a longer timescale, we see the replenishment of the susceptible pool (figure 17). In fact, the number of susceptibles increases roughly linearly for some time. Since the number of infectives has fallen to a very low level following the first outbreak, the dynamics of the susceptible population is well approximated by $dS/dt = \mu N - \mu S$. If only a small fraction of the population is susceptible, then $dS/dt \approx \mu N$.

Eventually, the susceptible population reaches the threshold level required for a disease outbreak and the infective population again starts to increase, leading to a second outbreak.

This process continually repeats, with successive outbreaks becoming less severe as the system approaches the endemic equilibrium via these damped oscillations. We notice that the time between successive outbreaks decreases (and eventually approaches the value predicted by the linear analysis).

In order to better understand this process, it is helpful to replot the behavior versus time graph with the number of infectives on a logarithmic scale. We see that the first outbreak leads to a very severe epidemic during which the number of susceptibles is considerably depleted. Following the outbreak, the number of susceptibles grows roughly linearly over time, but given the low susceptible fraction that is left after the epidemic, it takes a number of years for R_t to increase to the level at which the number of infectives can again increase. During this time, the infective fraction continues to decrease.

Once R_t reaches one, the number of infectives starts to increase. It takes a considerable time for this increase to become noticeable when I (or y) is plotted on a linear scale given the extremely low level to which I falls following the first outbreak. But eventually the second outbreak becomes apparent. This second outbreak is less severe because there are far fewer susceptibles around. Consequently, both S and I do not fall to such low levels following the second outbreak: the third outbreak can therefore occur after a shorter inter-epidemic interval.

One striking feature of the timecourse of the logged plot is the extremely low level to which I (or y) falls following the first outbreak. The number of infectives is far below a single person! We shall later return to discuss this within the context of stochastic modeling.

7 Control of Infection

In the examples we have seen, an epidemic cannot occur, or an infection cannot persist, if R_0 is less than one. We aim to control infection by reducing R_0. Since R_0 is given by the product of

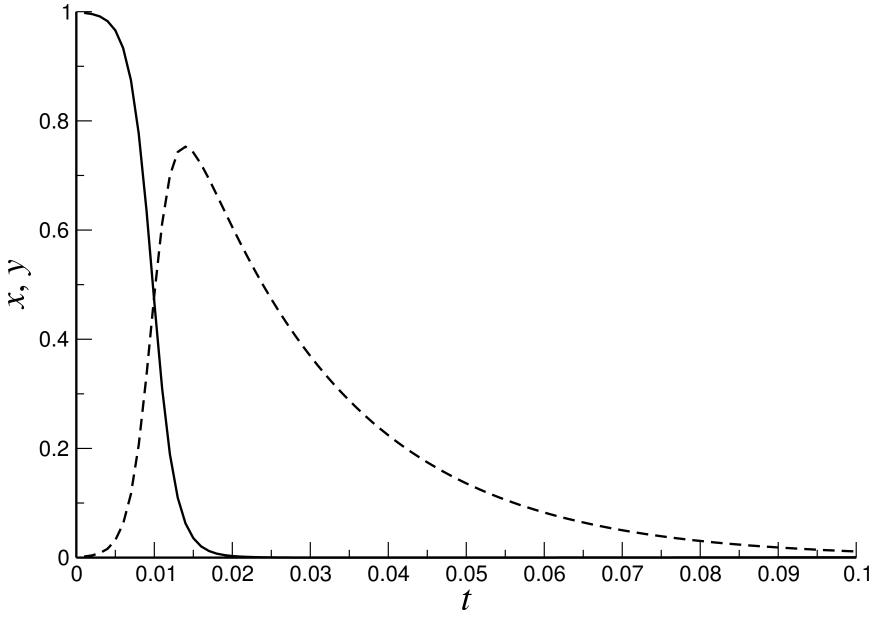

Figure 16: Initial behavior of the SIR model with demography when a small number of infectives are introduced into an otherwise susceptible population. Notice the scale on the time axis: the initial outbreak occurs and dies away very quickly. Parameter values are as in the previous figure.

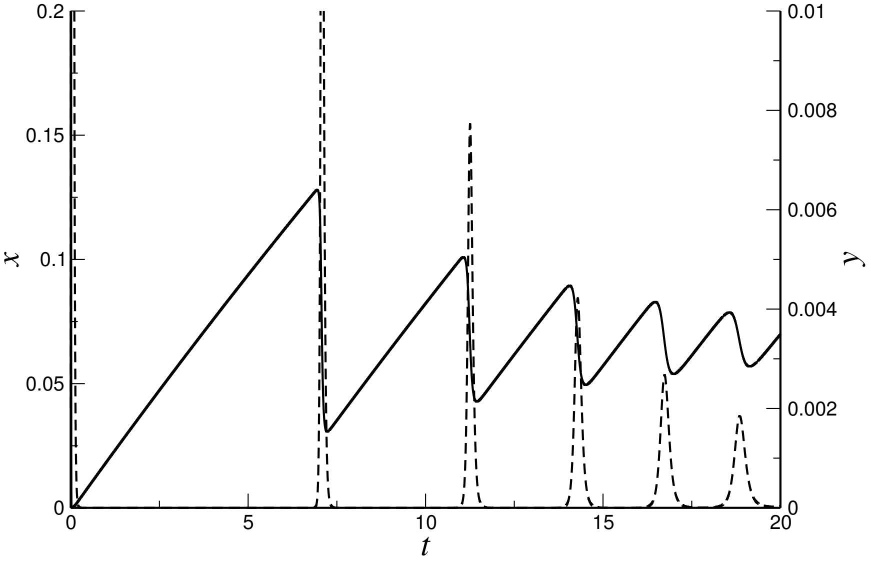

Figure 17: Longer-term behavior of the SIR model. Compare the scale on the time axis with the previous figure. Parameter values are as in the previous figure.

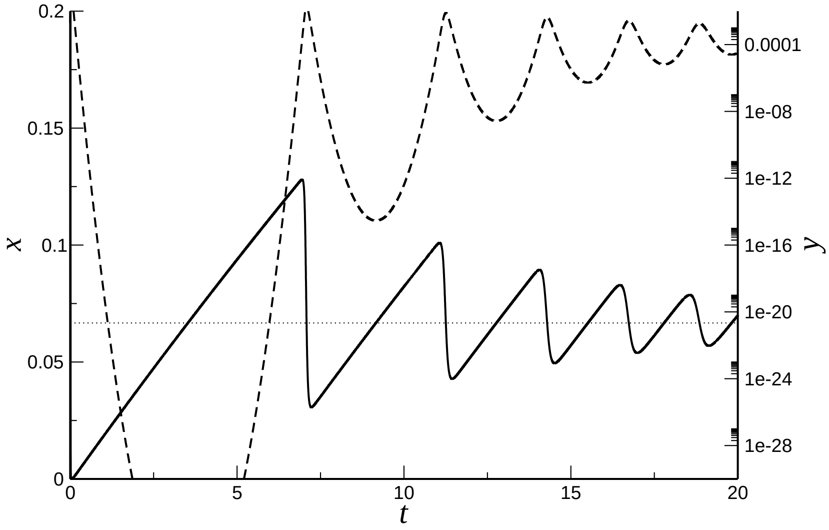

Figure 18: Longer-term behavior of the SIR model, with the infective fraction plotted on a logarithmic scale. The dotted horizontal line denotes the susceptible fraction at which R_t equals one. Parameter values are as in the previous figure.

β (which represents both infectiousness and the rate at which contacts are made) and the average duration of infectiousness, there are many ways by which R_0 could be reduced:

- Shorten the infectious period
- Reduce the number of contacts that infectives make
- Reduce the infectiousness of infectives.

Drug treatment of infectives could achieve aims 1 and 3, and quarantine (or inducing behavioral changes) could achieve aim 2.

The quantity R_0 refers to an entirely susceptible population. More generally, we think about R_t, where $R_t = R_0(S/N)$. If R_t is less than one, the number of infectives will decrease over time. From the definition of R_t, we see that R_t will always be less than one if the number of susceptibles is less than N/R_0. (Alternatively, the fraction of the population that is susceptible is less than $1/R_0$.)

The important upshot of this is that we can achieve control by reducing the number of susceptibles below N/R_0. **Notice that we do not need $S = 0$, just $S < N/R_0$.** We can achieve control by vaccinating a certain proportion of the population, p_c, or higher. This is called the **critical vaccination proportion**, and we see that $p_c = 1 - 1/R_0$, since vaccinating this proportion leaves just $1/R_0$ of the population who could possibly be in the susceptible class.

If we vaccinate a proportion p that is greater than p_c, the infection cannot invade or persist. Notice that not everyone need be vaccinated: this phenomenon is known as **herd immunity**. Even if you aren't vaccinated, you can gain an indirect benefit if sufficiently many others are. (Notice that this can be a problem if there are some risks associated with the vaccine, since my selfish choice would be to not get vaccinated, instead relying on herd immunity to protect me. Clearly this cannot work if everyone were to come to a similar conclusion.)

Notice that the critical vaccination proportion increases as R_0 increases. Unsurprisingly, it becomes more difficult to control infections that spread more readily. Looking at a table of R_0 values then explains why certain disease eradication campaigns (such as smallpox) have been more successful than others (such as measles). It is much more difficult to achieve the highest vaccination levels, especially if the vaccine is not 100% effective.

	R_0	p_c
smallpox	4-5	80%
polio	5-7	86%
chickenpox	8-10	90%
measles	15	93%

Table 1: Approximate values of R_0 for four infectious diseases, and corresponding critical vaccination proportions.

Some comments are in order here: these values were culled from Anderson and May, and are only approximate. Values of R_0 depend on the population setting: they might be quite different in developed or developing countries, and may well depend on the spatial scale at which one looks (think about how much more likely transmission would be within a hospital setting).

It's worth thinking about how these estimates of R_0 might have been obtained. While its easy to write down expressions for R_0 in terms of the parameters that appear in the model, it's often not so easy to estimate some of these parameters, in particular β. R_0 is usually estimated indirectly, for instance interpreting the initial behavior of an epidemic using a relationship like $r = \gamma(R_0 - 1)$. Alternatively, the epidemic size or 'average age at infection' relationships could be used. Another approach is the model fitting technique, where one tweaks the model parameters so that the behavior of the model most closely fits some observed epidemic data.

7.1 Modeling Control

Within the SIR-type framework, we typically assume that individuals are vaccinated at birth. If we assume that a fraction p of the population are successfully vaccinated at birth, then only $(1-p)$ of newborns enter the S class. The rest go straight into the R class. We have

$$\begin{aligned}\dot{S} &= \mu(1-p)N - \beta SI/N - \mu S && (42)\\ \dot{I} &= \beta SI/N - (\gamma+\mu)I && (43)\\ \dot{R} &= \gamma I - \mu R + \mu pN. && (44)\end{aligned}$$

with $S + I + R = N$.

As before, we can solve for the endemic equilibrium: this will still exist and be stable as long as we don't vaccinate more than p_c. An interesting observation is that the dI/dt equation has not changed: setting $dI/dt = 0$ again gives $S^* = N/R_0$ (assuming that we are not at the infection free equilibrium, $I = 0$). At first it might seem counterintuitive that we have the same number of susceptibles at equilibrium, even though we are vaccinating a fraction p of the population. (Thinking a little, though, we see that this is the way it has to be, since we still need $R_t = 1$ at the equilibrium.)

The observation that there must still be the same number of susceptibles at the endemic equilibrium leads to an important result: the average age at which individuals acquire infection increases in a partially vaccinated population. (Remember that in the unvaccinated case, we argued that since a fraction $1/R_0$ of the population could be found in the S class, the average age at which individuals left the S class, A, must equal L/R_0.) In the vaccinated case, only $(1-p)$ of the population ever enters the S class. So the average age at which **they** become infected, which we write as A', must equal $A/(1-p)$. (To see this, imagine that only half of newborns entered the S class. Since we know that the equilibrium number of individuals in the S class remains the same, each of them must stay there twice as long as they did in the unvaccinated situation. This doubles the mean age at infection. We see that to get the mean age at infection, we divided A by the fraction of the population (1/2) that entered the S class. Generalizing, we get the result claimed.)

The increase in the average age at infection can be a problem for infections that have more serious consequences in older individuals (e.g. rubella, which can lead to birth defects if a pregnant woman becomes infected). Before starting mass vaccination campaigns, the potential for problems in a partially vaccinated population must be recognized and addressed. Once the critical vaccination proportion is reached, and control is achieved, these problems will vanish.

Notice that the increase in the average age at infection has implications for the dynamics of the infection near the endemic equilibrium. Earlier, for an SIR model describing childhood diseases, we saw damped oscillations, with damping time equal to $2A$ and oscillation period given by $2\pi\sqrt{A\tau}$. As the population is vaccinated, A increases, which increases the damping time (the oscillations become less strongly damped) and increases the period of the oscillations.

7.2 Other Ways to Model Control

Control can be modeled in other ways, for instance one could imagine that β becomes reduced or γ increased. Another way to model vaccination might be to assume that susceptibles are moved into the recovered class at some rate (e.g. there might be a $-\alpha S$ term in the dS/dt equation).

An interesting alternative appears in a model for polio vaccination. The oral polio vaccine (OPV) involves administering a weakened (**attenuated**) strain of the virus. This leads to a mild infection that does not lead to disease. We call the naturally occurring— disease causing— virus the **wild-type** virus. We assume that infection with either virus leads to permanent immunity to both forms of the virus. An important point is that both strains of the virus can be transmitted (although not necessarily at the same rate) between individuals: susceptible individuals can acquire the attenuated strain. The circulation of the attenuated strain gives us additional vaccinations 'for free'.

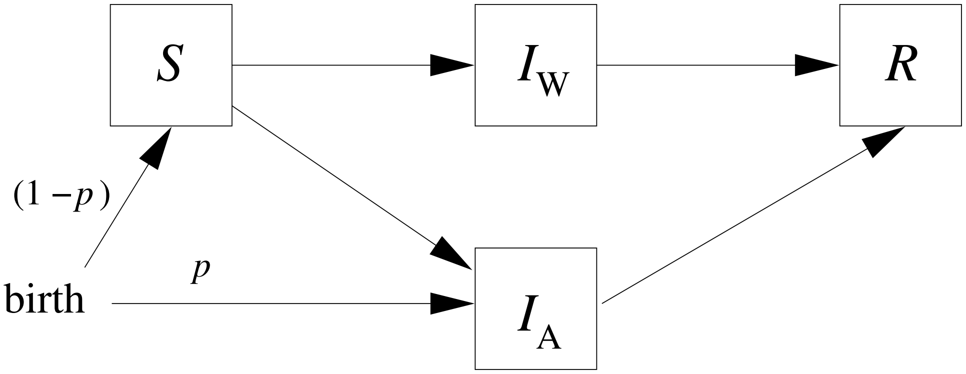

Figure 19: Flow diagram for the oral polio vaccination model. To simplify the picture, the arrows denoting background mortality are omitted.

We assume that a fraction p of newborns are vaccinated with OPV: they move into the attentuated virus infective class, I_A. The remaining fraction of newborns enter the susceptible class. They can become infected with either strain of virus, with transmission parameters given by β_W and β_A (for wild-type and attenuated strains, respectively). Notice that there are two arrows leaving the S class, leading to the I_W and I_A classes. We allow the recovery rates for the two strains to differ, and write γ_W and γ_A.

We then have the following equations:

$$\begin{aligned}
\dot{S} &= \mu(1-p)N - \beta_W S I_W/N - \beta_A S I_A/N - \mu S && (45)\\
\dot{I}_W &= \beta_W S I_W/N - (\gamma_W + \mu) I_W && (46)\\
\dot{I}_A &= \beta_A S I_A/N - (\gamma_A + \mu) I_A + \mu N p. && (47)
\end{aligned}$$

We can define the basic reproductive numbers for the two strains of the virus:

$$R_0^W = \frac{\beta_W}{\gamma_W + \mu} \qquad\qquad R_0^A = \frac{\beta_A}{\gamma_A + \mu}. \tag{48}$$

The two virus strains compete for susceptibles: the behavior of the system reflects this competition. If either R_0^A or R_0^W were less than one, the particular strain could not circulate by itself. (Notice that the attenuated strain would persist because we are constantly reintroducing it by means of the vaccination.) So, the interesting cases involve both R_0^A and R_0^W being greater than one.

If R_0^A is greater than R_0^W, i.e. the attenuated strain spreads more effectively, the attenuated strain drives the wild-type to extinction. The endemic equilibrium has $I_W = 0$ and I_A positive. In ecology, this phenomenon is known as **competitive exclusion**.

If R_0^W is greater than R_0^A and there is no vaccination, the wild-type virus would win the competition between strains, driving the attenuated virus to extinction. But the attenuated strain is being constantly introduced by vaccination. Vaccination also reduces the pool of susceptibles that could acquire the wild-type strain, reducing the transmission potential of the strain. Thus we might imagine that the critical vaccination fraction is lower than in the simple vaccination model. This is indeed the case, and analysis of the model equations shows that

$$p_c = \left(1 - \frac{1}{R_0^W}\right)\left(1 - \frac{R_0^A}{R_0^W}\right). \tag{49}$$

As R_0^A approaches R_0^W, we see that the critical vaccination fraction approaches zero, reflecting the outcome of the competition discussed above.

The interaction between two virus strains that confer complete (or partial) immunity to each other is known as **epidemiological interference**.

8 The SEIR Model

One unrealistic feature of the models considered so far is that individuals become infectious immediately upon infection. In this section we extend the SIR model with demography to account for a latent period between an individual acquiring infection and becoming infectious. We add a **exposed** class to the model.

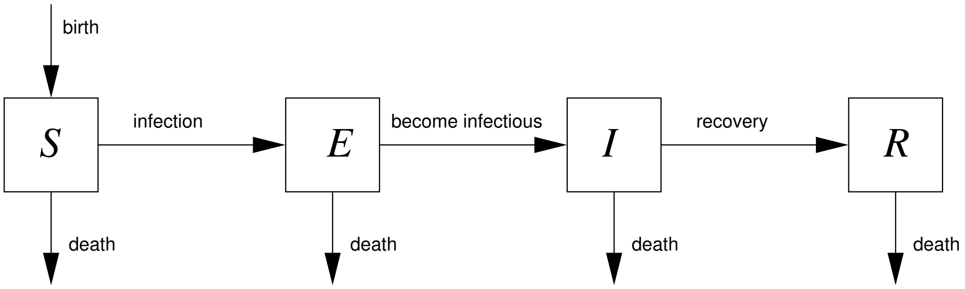

Figure 20: Flowchart showing movement between classes in the SEIR model.

If we assume an exponentially distributed latent period, with average duration of latency $1/\sigma$, we have the following set of differential equations

$$\begin{aligned}
\dot{S} &= \mu N - \beta SI/N - \mu S && (50)\\
\dot{E} &= \beta SI/N - (\sigma + \mu)E && (51)\\
\dot{I} &= \sigma E - (\gamma + \mu)I && (52)\\
\dot{R} &= \gamma I - \mu R. && (53)
\end{aligned}$$

As usual, we assume a constant population size, so we have $S + E + I + R = N$.

The average duration of infectiousness, τ, accounting for mortality is $1/(\gamma + \mu)$, and the average duration of latency, τ', accounting for mortality is $1/(\sigma + \mu)$. As before, if γ and σ are large compared to μ (i.e. latency and infectiousness are of short duration compared to the lifespan), these quantities are approximately equal to $1/\gamma$ and $1/\sigma$, respectively.

Notice that there are two possible ways of leaving the exposed class: not all individuals who acquire infection go on to become infectious: a small fraction dies before doing so. Comparing the two rates at which individuals leave the exposed class, we see that the probability P of going on to become infectious is $P = \sigma/(\sigma+\mu)$. Since σ is much bigger than μ, P is very close to one. (One can write $P = (1 + \mu/\sigma)^{-1}$ and expand in the small quantity μ/σ, using the binomial approximation: $(1 + x)^n \approx 1 + nx$ if x is small.)

To find R_0 for this model we employ similar arguments to before, and find that $R_0 = \beta\tau P$: the product of the average rate at which new infections arise in an entirely susceptible population, the average duration of infectiousness, and the factor P. This last factor decreases R_0 slightly to account for the fact that not all individuals that acquire infection go on to become infectious. In terms of the parameters that appear in the model, we have

$$R_0 = \frac{\beta}{\gamma + \mu}\frac{\sigma}{\sigma + \mu}. \tag{54}$$

As we've seen before, $R_0 = 1$ determines a threshold condition for the invasion of the infection into an entirely susceptible population and for the persistence of the infection at an endemic level.

8.1 Initial Behavior

The exposed class introduces a delay between acquisition of infection and transmission of infection. This slows the initial growth of an epidemic, as illustrated in the following figure.

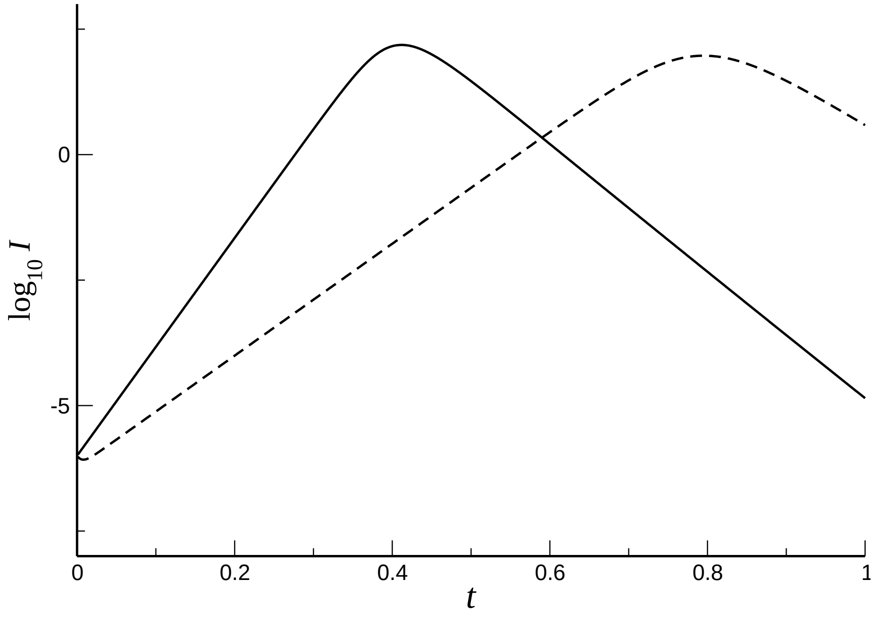

Figure 21: Comparison between initial epidemic behaviors in SIR model (solid curve) and SEIR model (dashed curve). The delay between infection and the start of infectiousness in the SEIR model slows the initial growth of the epidemic. Parameter values are as follows: $\mu = 0.02$ year^{-1}, $\gamma = 50$ year^{-1}, $\sigma = 80$ year^{-1} and $\beta = 100$ year^{-1}. This corresponds to an infection with infectious period equal to about 7.3 days and exposed period equal to about 4.6 days.

We can calculate the initial growth rate by linearizing the model, remembering that linearizing about the infection-free equilibrium essentially corresponds to setting $S = N$. Unlike the SIR model, for which we obtained a single equation for the rate of change of I, the corresponding linearization of the SEIR model gives a pair of coupled linear differential equations:

$$\begin{aligned}
\dot{E} &= \beta I - (\sigma + \mu)E & (55)\\
\dot{I} &= \sigma E - (\gamma + \mu)I. & (56)
\end{aligned}$$

We find the growth rate by examining the eigenvalues of the Jacobian matrix,

$$J = \begin{pmatrix} -(\sigma + \mu) & -\beta \\ \sigma & -(\gamma + \mu) \end{pmatrix}. \qquad (57)$$

The eigenvalues, r, satisfy $\det(J - rI) = 0$: this gives a quadratic equation that can be manipulated as follows:

$$\begin{aligned}
0 &= (\mu + \sigma + r)(\mu + \gamma + r) - \sigma\beta, \\
&= r^2 + (\mu + \sigma + \mu + \gamma)r + (\mu + \sigma)(\mu + \gamma) - \sigma\beta, \\
&= r^2 + \left(\frac{1}{\tau'} + \frac{1}{\tau}\right) r + (\mu + \sigma)(\mu + \gamma)\left\{1 - \frac{\sigma\beta}{(\mu + \sigma)(\mu + \gamma)}\right\}, \\
0 &= r^2 + \left(\frac{1}{\tau'} + \frac{1}{\tau}\right) r + \frac{1}{\tau'\tau}(1 - R_0). & (58)
\end{aligned}$$

It's easy to see that the stability of the infection free equilibrium— and hence whether an epidemic grows or decays— depends on whether R_0 is greater than or less than one.

Unlike the SIR case, we do not have a simple expression for r in terms of R_0. Instead, we have a quadratic equation that links the two. (Notice that this could be rearranged to give R_0 in terms of r and the average durations of latency and infectiousness.) We could solve to get r in terms of R_0, but this is a little messy.

We can get something simpler if we consider a special case in which R_0 is not much larger than one, so that r remains small. We assume that r^2 is small compared to the other two terms that appear in the quadratic (58). Dropping the quadratic term, we get a linear equation which gives us

$$r \approx \frac{1}{\tau + \tau'}(R_0 - 1). \qquad (59)$$

Comparing to the expression we got in the SIR case, $r = (R_0 - 1)/\tau$, we see that the effect of the latent period is to reduce the rate at which the epidemic grows, by an amount that depends on the relative durations of the latent and infectious periods. If we compared an SEIR and an SIR epidemic, both of which had the same value of R_0, the SEIR epidemic would grow more slowly.

This observation has an important implication for attempts to estimate R_0 based on the rate at which an epidemic grows during its early stages. The use of the SIR model to infer r from R_0 will tend to underestimate the true value of R_0 if there is a latent period. The slower growth rate of an SEIR epidemic– due to the latent period– would be interpreted as arising from a lower value of R_0 if one was thinking using the SIR framework. Notice that this mis-estimation of R_0 is in an unfortunate direction: it leads us to believe that R_0 is smaller than it really is, which will give us more optimistic estimates of the vaccine coverage needed to control the infection. If we designed a control policy based on the critical vaccination coverage predicted using our estimated value of R_0, we would not achieve control.

8.2 Endemic Equilibrium Behavior

If R_0 is greater than one, we have the following endemic equilibrium

$$\begin{aligned} S^* &= \frac{N}{R_0} && (60) \\ E^* &= \mu N \left(1 - \frac{1}{R_0}\right) \frac{1}{\mu + \sigma} && (61) \\ I^* &= \frac{\mu N}{\mu + \gamma} \left(1 - \frac{1}{R_0}\right) \frac{\sigma}{\sigma + \mu}. && (62) \end{aligned}$$

The Jacobian matrix is given by

$$J = \begin{pmatrix} -\beta I^*/N - \mu & 0 & -\beta S^*/N \\ \beta I^*/N & -(\sigma + \mu) & \beta S^*/N \\ 0 & \sigma & -(\gamma + \mu) \end{pmatrix}. \quad (63)$$

To find the eigenvalues, we solve $\det(J - \lambda I) = 0$, using the expression for the determinant of a 3×3 matrix: $\det A = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{31}a_{23}) + a_{13}(a_{21}a_{32} - a_{31}a_{22}))$. This leads to a cubic equation, and after some simplification, we get

$$\lambda^3 + (\gamma + \sigma + \mu R_0 + 2\mu)\lambda^2 + \mu R_0(\gamma + \sigma + 2\mu)\lambda + \mu(R_0 - 1)(\gamma + \mu)(\sigma + \mu) = 0. \quad (64)$$

This is much more messy to work with than the quadratic that we got in the SIR case. Again, we think about special cases, particularly the one in which there is some separation of timescales. If the average durations of latency and infectiousness are both short compared to the mean age at infection and the average lifespan, we have that both σ and γ are much greater than μ and μR_0. We can then obtain the following approximate version of (64)

$$\lambda^3 + (\gamma + \sigma)\lambda^2 + \mu R_0(\gamma + \sigma)\lambda + \mu(R_0 - 1)\gamma\sigma = 0, \quad (65)$$

which can be re-written as

$$\lambda^3 + (\gamma+\sigma)\left\{\lambda^2 + \mu R_0 \lambda + \mu(R_0-1)\frac{\gamma\sigma}{\gamma+\sigma}\right\} = 0. \tag{66}$$

In appendix C of Anderson & May, it is stated that this cubic has an approximate root, $\lambda \approx -(\gamma+\sigma)$ and that the remaining roots are obtained by solving the quadratic between the braces. (This can be demonstrated by examining the limit $\mu \to 0$ in more detail.) Notice that the root $-(\gamma+\sigma)$ represents rapidly decaying behavior as its value is large and negative, and that the new quadratic looks not unlike the one we obtained when thinking about the behavior near the endemic equilibrium of the SIR model. The one difference between the quadratics is that the constant term in the SIR case was given by $\mu(R_0-1)/\tau$, whereas we now have $\mu(R_0-1)/(\tau+\tau')$. Hence we again have damped oscillations with the same damping time but whose period is given by $2\pi\sqrt{A(\tau+\tau')}$. Inclusion of the latent period increases the period of the damped oscillations about the endemic equilibrium.

9 Comparing the SIR Model to Reality

Since we have been discussing the behavior of the SIR model in the setting of a childhood infection, we shall look at data showing the incidence of measles in different locations.

In the data, we see:

- **Sustained oscillations.** Measles epidemics occur with a two-yearly (biennial) period in most cases. In some instances, annual epidemics are seen, while in others, epidemics occur every 3 years (triennial).
- **Some Irregularities.** The peak incidences differ between outbreaks in the same city.
- **Breaks in the chain of infection.** The infection can undergo ‘fadeout’ following an epidemic as the number of infectives falls to low levels in the ‘troughs’ between epidemics.

What does the SIR model do? It exhibits damped oscillations to an endemic equilibrium. The period of these oscillations is given by $2\pi\sqrt{A\tau}$. For measles, with an R_0 of 15 and a 7 day average duration of infectiousness (=7/365 years), the average age at infection is 5 years if we assume a 75 year lifespan. The period of the damped oscillations is then equal to $2\pi\sqrt{5/(7/365)} = 2\pi\sqrt{0.1} \approx$ 1.98 years.

The SIR model correctly predicts the period of the oscillations, but incorrectly predicts that they are damped. We need to look for mechanisms that could lead to the maintenance of oscillatory behavior. It was originally thought that the inclusion of an exposed class could achieve this, but

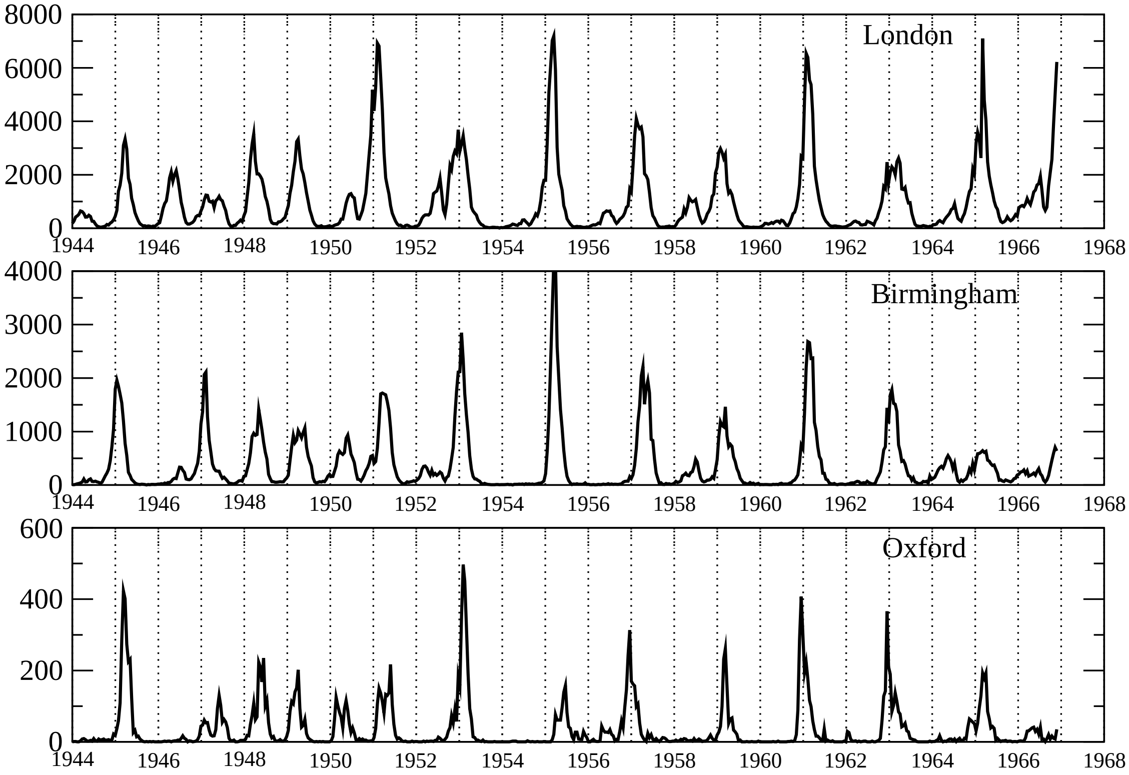

Figure 22: Measles incidence data (fortnightly numbers of cases) for three British cities: London (approx. population size 3 to 5 million), Birmingham (approx. population size 1 million) and Oxford (approx. population size 100 000). The timeseries covers the dates 1944 through 1967: this final date is around the time when mass vaccination against measles was introduced in the U.K.

this is not the case: the SEIR model also exhibits damped oscillations. Two other mechanisms have been suggested: seasonal variations in contact rate (between school terms and vacation) and randomness in the system. We shall examine both of these, but shall consider the impact of randomness first.

Random effects (also known as stochastic effects) are also implicated in the second and third observations that arose from the timeseries (variability in epidemics and fadeout of infection). We remark that another mechanism that can lead to variability between epidemics has been discussed, namely deterministic chaos. We shall return to this when we come back to seasonal variations in the transmission parameter.

10 Demographic Stochasticity in Epidemic Models

The fadeout phenomenon is an inherently stochastic effect. During the inter-epidemic troughs, there are only a few number of infectives. By chance, it is possible for each of these to recover before passing on the infection, interrupting the chain of transmission. Notice that once the infection has undergone fadeout, it can only reoccur if it is reintroduced from elsewhere. This can lead to dynamics that reflect a combination of extinction and recolonization processes: introduction of infection leads to an epidemic, which may lead to fadeout, after which the infection is reintroduced at some point, leading to fadeout, and so on. (Ecologists talk about a similar situation within the context of "metapopulation dynamics".)

An important point about random effects is that their impact depends on population size. More precisely, the impact of stochastic effects depends on the numbers of infectives (since this is typically the smallest subpopulation within the model). If the population is large enough then the probability of fadeout will be small: infection can be maintained without reintroduction from an external source. This is Bartlett's notion of the **critical community size**. Incidence data shows that measles can remain persistent in populations that have at least 300 000 to 500 000 people. The critical community size phenomenon is most clearly seen for island populations (which are naturally isolated, at least in the days before the widespread use of air travel). It is less clear-cut for cities within a country (since they are not so well isolated).

There is no way that out deterministic models can reproduce this effect: if we look at the fractions of the population that are in the different compartments, the deterministic models are independent of N. If R_0 is greater than one then we always have an endemic equilibrium, and this equilibrium is reached for any positive initial number of infectives. The deterministic model does not know that 1/10th of an individual is biologically meaningless. Instead, we turn to stochastic models.

10.1 From Deterministic to Stochastic Models

We have been looking at deterministic models for infection processes. These models treated the numbers in each class (e.g. susceptible or infective) as continuously varying quantities. As an example, take the deterministic SIR model with demography:

$$\frac{dS}{dt} = \mu N - \mu S - \beta SI \tag{67}$$

$$\frac{dI}{dt} = \beta SI - \mu I - \gamma I. \tag{68}$$

This model considers four different processes: infection, recovery, birth and death. The terms appearing in equations (67) and (68) describe the rates at which these different processes occur.

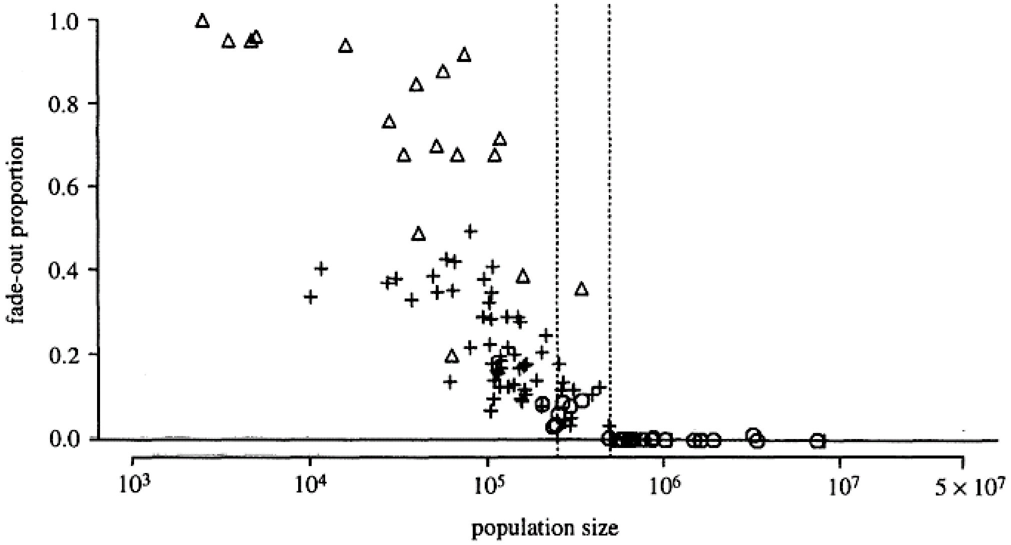

Figure 1. Empirical estimates of critical community size. Horizontal axis shows population size; vertical axis shows 'fade-out proportion', the fraction of months in each sample with no reported cases of measles. The vertical dotted lines represent the empirical estimate of critical community size range (Bartlett 1957, 1960*a*), 250000–500000. Data from Black (1966) (△, islands) Bartlett (1960*a*) (○, U.S. and Canadian cities), and Shaw (1990) (+, British cities; data originally from OPCS (1948–68)). Although there are many possible sources of variation in the data, including different reporting rates, *per capita* birth rates and levels of outside epidemiological contact, all the data fall roughly along the same curve.

Figure 23: The critical community size phenomenon. Figure taken from Bolker & Grenfell (1995), Phil. Trans. R. Soc. Lond. B **348**, 309-320.

Event	Transition		Rate at which event occurs	Probability of transition in time interval $[t, t + dt]$
Birth	$S \to S + 1$		μN	$\mu N dt$
Susceptible death	$S \to S - 1$		μS	$\mu S dt$
Infection	$S \to S - 1,$	$I \to I + 1$	$\beta SI/N$	$(\beta SI/N)dt$
Recovery	$I \to I - 1$		γI	$\gamma I dt$
Infectious death	$I \to I - 1$		μI	$\mu I dt$

Table 2: Possible events in the SIR model, their rates and probabilities of their occurrence over a short time interval

The stochastic formulation recognizes that the population is made up of individuals and that transitions between classes really involve discrete events. These events change the number of susceptibles or infectives one at a time. For instance, an infection event decreases the number of susceptibles by one while increasing the number of infectives by one. This description introduces some randomness into the model: rather than transitions occuring continuously, they occur as discrete events, whose exact times are random. We call such randomness **demographic stochasticity** as it arises as a consquence of having a **finite population size**. Notice that the deterministic model gives us an idea of the **average rates** at which these events occur.

The translation of the model from a deterministic to a stochastic framework is achieved by reinterpreting the rates at which the various processes occur to give the probabilities of the various events happening in an infinitesimal time interval dt. To work with something concrete, we shall take the SIR model with demography. The events, the rates at which they happen and their probabilities are listed in table 2. For instance, the probability of a birth occurring in the time interval $[t, t + dt]$ is $\mu N\, dt$.

More precisely, we assume that the different processes can be described by independent **Poisson processes** whose rates are given by those that appear in the deterministic model. Notice that since we do not keep track of the number of recovered individuals, we need not distinguish between the death of an infective and the recovery of an infective: both simply reduce I by one.

Basic Facts About Poisson Processes

"Events" occur randomly in time in the following sense: there is a constant ρ, known as the rate of occurrence (or intensity) of the process, such that

- pr(one event in $(t, t + dt]) = \rho\, dt + o(dt)$
- pr(no event in $(t, t + dt]) = 1 - \rho\, dt + o(dt)$
- pr(two or more events in $(t, t + dt]) = o(dt)$.

(The notation $o(dt)$ means a quantity that tends to zero faster than dt as $dt \to 0$.)

Events are independent and the behavior does not depend on the history of the system.

Properties

- The distribution of the number of events in $(t, t + h]$ is Poisson, with mean ρh. (ρ is the expected number of events per unit time.)
- The time between events is exponentially distributed with mean $1/\rho$.
- If we have independent Poisson processes, with rates ρ and μ then the sum of the processes is also Poisson, with rate $\rho + \mu$.

10.2 Simulation of the Stochastic Model

The simulation of the stochastic model is relatively straightforward. The basic approach involves simulating each event that occurs in the population and updating the numbers of susceptibles and infectives accordingly. One issue with this technique is that it may be quite time consuming. Thinking about the SIR model, a typical individual has four events happen over the course of their lifetime: they are born, become infected, recover and later die. The total number of events that we must simulate is proportional to the population size: if N is large then this could involve a large amount of computation.

We remark that it is a simple matter to calculate incidence in this stochastic framework since infection events (like the other types of events) are accounted for separately. We simply count the number that occur in whatever time interval is of interest.

The simplest simulation technique makes use of the properties of Poisson processes. Recall that for a Poisson process with constant rate ρ, the time between events is exponentially distributed with mean $1/\rho$. Also recall that if we have two independent Poisson processes, then their sum is also a Poisson process, with rate given by the sum of those of the two separate processes. (Notice that this summed process describes the occurrence of events of either type.) Another property that we will make use of later is that the number of events that occur in a time interval of length h is Poisson distributed with mean ρh.

Returning to the SIR model, we can sum the rates of the four population processes; we call this rate T (for total rate). In the case of the SIR model, T equals $\mu N + \beta SI + \mu S + (\mu + \gamma)I$, and gives the rate at which events (of any type) occur in the population. The time between events, therefore, is exponentially distributed with mean $1/T$. We can, therefore, find the time at which the next event occurs by sampling from an exponential distribution with mean $1/T$.

Now that we know when an event has occurred, how do we find out what type of event occurred? The different rates give the relative probabilities of the occurrence of the different events. Given that an event has occurred, the probability that it is a birth is given by $\mu N/T$, the probability that it is an infection is $\beta SI/T$, etc. If we sample a random number, X, from the uniform distribution on $(0, T)$, we decide that a birth has happened if X is less than μN, an infection if X lies between μN and $\mu N + \beta SI$, and so on. The idea here is that the different events divide up the interval $(0, T)$ into lengths that correspond to their relative probabilities of occurrence.

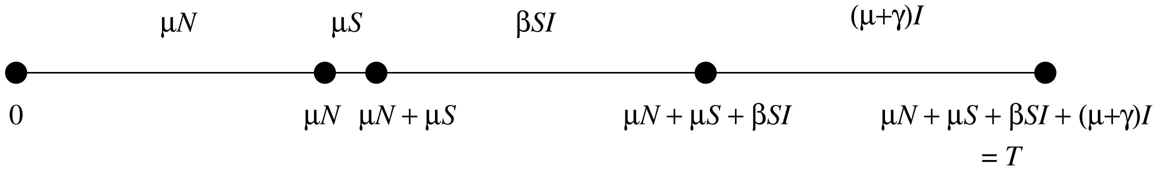
μN
μS
βSI
$(\mu+\gamma)I$
0
μN
$\mu N + \mu S$
$\mu N + \mu S + \beta SI$
$\mu N + \mu S + \beta SI + (\mu+\gamma)I$
$= T$

Figure 24: The division of the interval $(0, T)$ into lengths that give the relative probabilities of the occurrence of the four events in the SIR model with demography

We then update S and I appropriately, and repeat the simulation step until we reach the desired endpoint of the simulation. This so-called Monte Carlo simulation method generates exact realizations of the stochastic model. This simulation technique is sometimes associated with Gillespie, although it has a much longer history of use in the literature.

In general, if we have some set of possible events $\{1, 2, \ldots, n\}$, with corresponding rates $\{r_1, r_2, \ldots, r_n\}$, we first calculate the summed probabilities s_i. (s_1 simply equals r_1, s_2 is given by $r_1 + r_2$, and so on. s_n gives the total rate T.) We sample a number from an exponential distribution with mean $1/s_n$. This will tell us how long this timestep will be, and is the amount that the time variable t must be incremented by at this step. We then generate a random number uniformly from the interval $(0, s_n)$ and find the value of m such that this number lies between s_{m-1} and s_m. This tells us which event has occurred, and so we update the state variables (e.g. S and I) appropriately.

10.2.1 *Practical Issues

In many cases, we can easily generate a uniform random number between 0 and 1, but we might not have routines for more general distributions. A uniform random number between 0 and T is easily obtained by multiplying by T. An exponentially distributed random number can be generated by taking the natural logarithm of the $U(0, 1)$ number, and multiplying by $-1/T$. There is a possible pitfall here: some random number generators could return 0 for the uniform random number. You must check that this either cannot happen or, if zero could be returned, include a loop that would check and ask for another number if this happened. (This problem can be difficult to spot: it can take a very long time before zero is returned: the code can run fine for a long time until this situation is reached at some point.)

There is also an issue concerning the decision as to which event occurred. The simplest way to do this is to compare the second random number in turn to s_1, s_2, s_3, and so on. The problem arises when one of the events with larger m happens: one has to make a whole load of comparisons before reaching the appropriate one. A simple way to lessen this problem is to have the most frequently occurring events near the top of the list. A better, although more complex, way to go is to do something known as a binary search. Choose a number, call it j, close to $n/2$ and compare the uniform random number to s_j. If it is less than s_j then we know that one of the events $1, 2, \ldots, j$ has occurred. If not, then one of $j + 1, \ldots, n$ happened. In one step, we have eliminated half of the possible events. Repeated use of this process will locate the event much more quickly than looking through all possible events in turn.

If n is small, as in the simplest SIR model, then this binary search is easily explicitly coded by hand. In more complex settings, one can use more general code to do the search.

The simulation will spend a lot of time just calculating the rates at each timestep. In many situations, some of these rates may not change at every timestep. For instance, the recovery rate will not change if a susceptible is born. Some time can be saved by only updating rates when they change.

The simulation will also spend a lot of time sampling random numbers: two numbers must be generated at each timestep. Using a fast random number generator will save a lot of time.

10.2.2 *Alternative Exact Simulation Techniques

An alternative simulation technique calculates the times at which each different type of event would happen next. This is achieved by sampling numbers from appropriate exponential distributions. These times are compared and the appropriate event is determined to have happened. This technique is known as the next event method.

At first sight, this would seem to be a slower approach: at each step of the SIR model we would have to generate four exponential random numbers. Three of these numbers appear to be wasted since only one of the events ends up happening. The benefit comes because it turns out that we can **reuse** the remaining random numbers. First, if the rate at which an event occurs does not change as a result of the transition that occurred, we can simply keep the time predicted in the previous step. Second, if the rate changes, we can simply rescale the predicted time by an appropriate factor to account for the change in the rate. (At first sight, one might worry about non-independence of the random numbers from step to step, but one can show that this technique is legitimate.) Write the old rate of the given event as r_{old}, the new rate as r_{new}, the previously predicted time of the occurrence of the event as t' and the current simulation time as t. It is then fairly easy to see that the appropriate updated predicted time for the occurrence of the event is

$$t'_{\text{updated}} = \frac{r_{\text{old}}}{r_{\text{new}}}(t' - t) + t. \tag{69}$$

Subtracting t from both sides of this equation makes its origin clear: the time remaining until the next event $(t'_{\text{updated}} - t)$ is simply a multiple of the time that would have remained if the rate had not changed $(t' - t)$.

As before, in order to find the next event, we have to find the smallest of n numbers. This process can again be streamlined by using binary search techniques.

This improved simulation technique requires just a single random number at each timestep and so is likely to be much faster than the simple method described earlier.

10.2.3 *Alternative Approximate Simulation Techniques

The exact simulation techniques proceed one event at a time. An alternative approach allows for multiple events per timestep, but at the cost of being approximate.

We choose an ‘appropriate’ time step, h. (More on what ‘appropriate’ means later...) For each event, we generate a Poisson random number with mean $r_i h$. This gives the numbers of events of each type that would occur over the time interval **if the rates r_i were constant over time**. In reality, these rates are not constant as they depend on the state variables. The approximation being used is that h is sufficiently small that the r_i do not vary much over the timestep.

A simple use of this method involves choosing h at the outset and employing the same value throughout. More sophisticated approaches might vary h over the course of the simulation, reducing the timestep if the system is in a state for which the r_i vary quickly. For instance, the r_i would be less likely to change by an important amount if both S and I were large than if I was small.

Hybrid approaches are possible between some of these simulation approaches, even going as far as a mixed deterministic and stochastic approach. It might be possible to describe some variables, or parts of phase space, deterministically.

10.3 Impact of Stochasticity on Invasion of an Infection

We already commented that the effects of stochasticity are most noticeable when the number of infectives is small. We would expect that randomness could play an important role during the early stages of the invasion of an infection.

In the presence of randomness, introduction of infection need not lead to a major outbreak, even if R_0 is greater than one. For instance, if there is a single infective, they may recover before passing on the infection. Stochasticity also leads to variability in the timecourse of an epidemic: repeated simulation of a stochastic model, starting from the same initial condition, leads to realizations that have different timecourses. This is in marked contrast to repeated simulation of a deterministic model.

We study these effects in a simple setting by considering the SIR model in a closed population. We plot both the timecourse of realizations of the model and the distribution of outbreak sizes seen when infection is introduced into the population. Here, we generate 1000 realizations of the model at a combination of different R_0 values and nitial numbers of infectives. We take the population size to equal 1000.

Figure (25) illustrates the timecourse of five realizations of the model when R_0 is equal to two and one infective is initially present. We see that extinction occurs quickly in two out of these five realizations. A degree of variability can be seen between the timecourses of the three remaining realizations, although each one leads to a sizeable outbreak.

When R_0 is equal to two and a single infective is initially introduced, we see that roughly half of the introductions lead to major outbreaks (figure 26). When these outbreaks occur, they are fairly large, with around 80% of the population becoming infected. Notice the variability of the sizes of these outbreaks.

Major outbreaks occur less frequently when R_0 is equal to 1.5, with a major outbreak occurring less than 40% of the time. When R_0 is below one, we typically see just a small outbreak, although there are a few realizations in which more than 50 cases are seen (figure 26).

We now turn to a situation in which we initially introduce five infectives (figure 27).

The main difference we see is that the probability of a major outbreak is much larger. Notice that

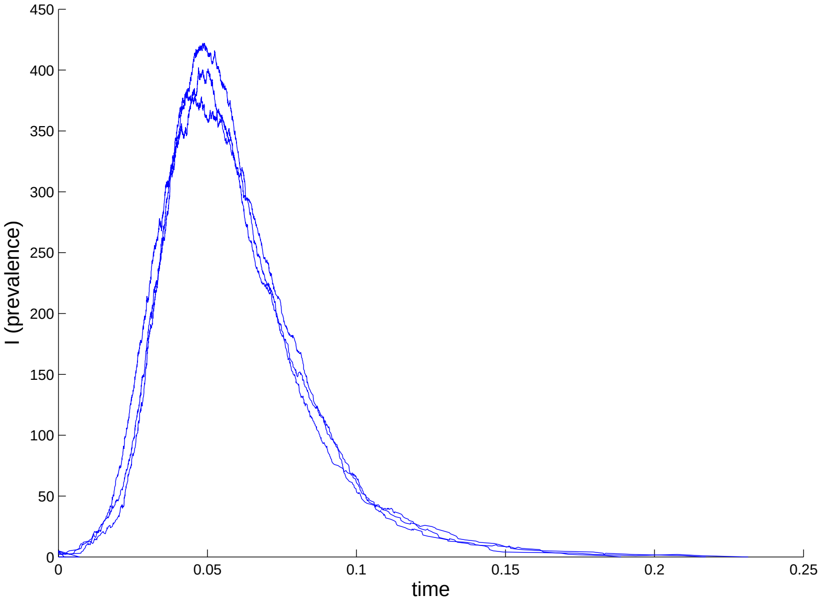

Figure 25: Five realizations of the SIR model with demographic stochasticity in a closed population. One infective is introduced into an otherwise susceptible population at time $t = 0$. Notice that the infection quickly goes extinct in two of the five realizations. Parameter values are as follows: $N = 1000$, $\beta = 200$ and $\gamma = 100$.

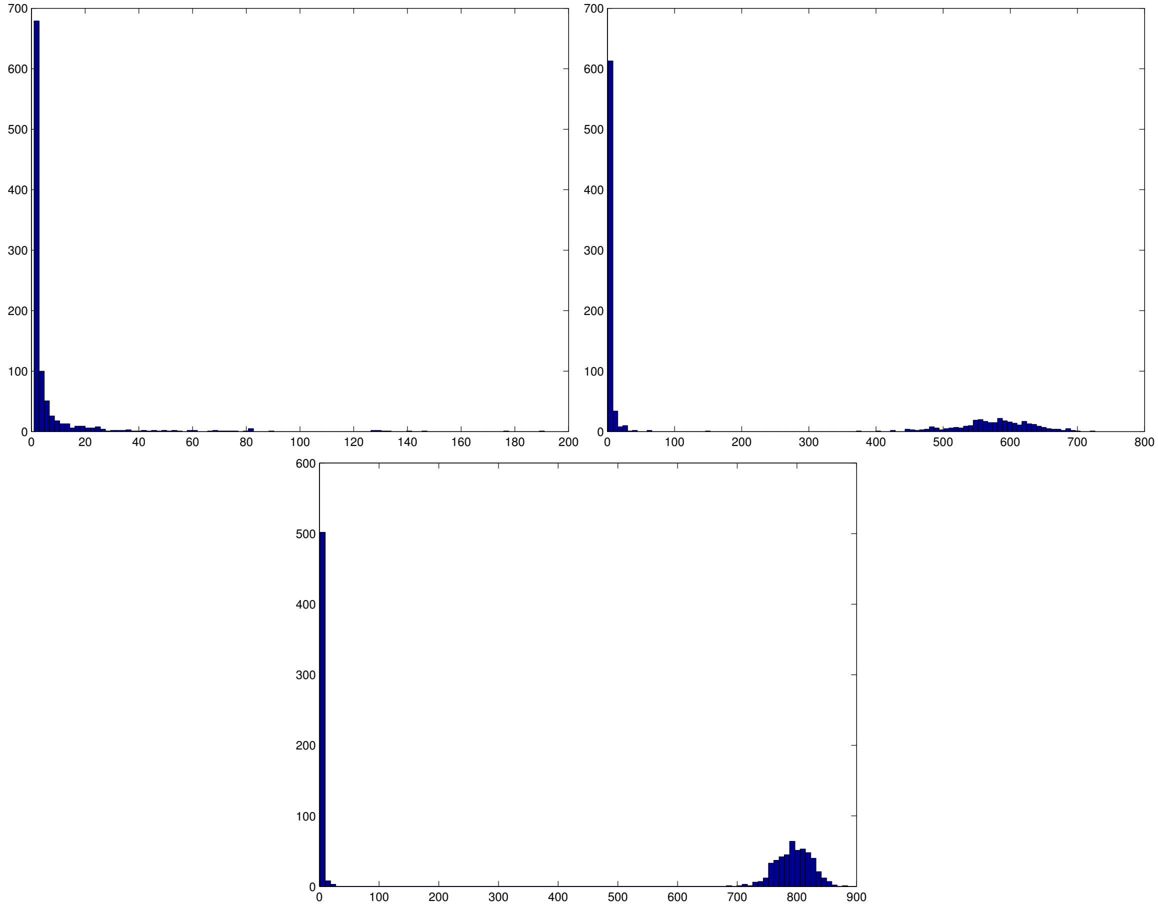

Figure 26: Outbreak size distributions with one infective introduced into a population of 1000. R_0 taken to be 0.9, 1.5 and 2.0. Notice that the graphs have different scales on their axes.

some substantially-sized outbreaks are seen even in the case where $R_0 = 0.9$. We also see that the variability in the outbreak sizes is much smaller when $R_0 = 2$ compared to $R_0 = 1.5$.

Compared to the deterministic situation, there is not such a clear-cut threshold result. We see that below threshold, one has a 'J-shaped' distribution. As one moves above threshold, one gets a bimodal distribution, with the weight of the lower mode decreasing as one moves further from the threshold.

Some authors talk about a 'transition region' between the two situations. The range of R_0 values that fall in this region (the 'width' of the region) depends on the number of initial introductions: we move more quickly from one situation to the other as $I(0)$ is increased. Figure 28 perhaps illustrates behavior in the transition region:

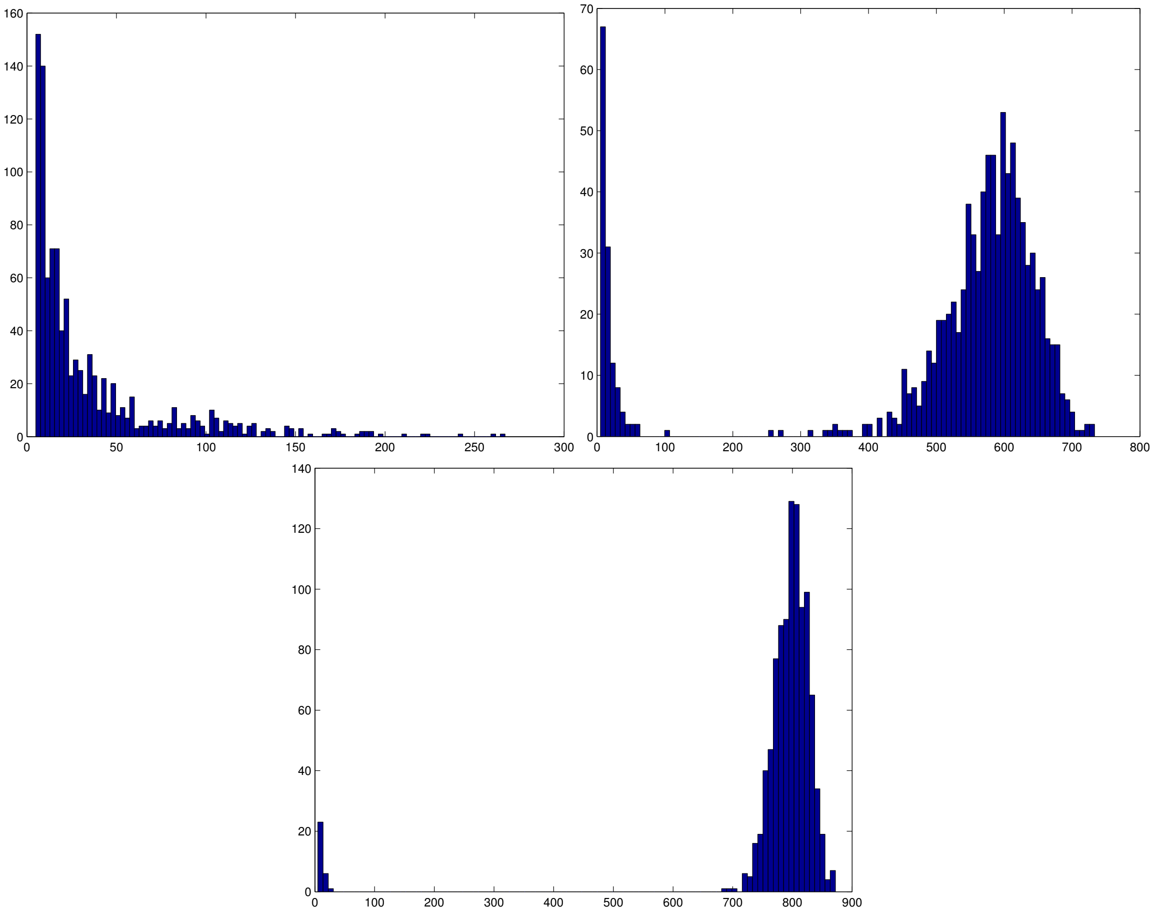

Figure 27: Outbreak size distributions with five infectives introduced into a population of 1000. R_0 taken to be 0.9, 1.5 and 2.0. Notice that the graphs have different scales on their axes.

10.4 Impact of Stochasticity on Equilibrium Behavior

In the SIR model with demography, stochastic effects prevent the system from settling into its endemic equilibrium. Taking parameter values that are appropriate for measles, we see that individual realizations of the model continue to fluctuate around the equilibrium with period roughly equal to the oscillatory period predicted by the linear analysis. The presence of randomness means that these fluctuations are not as regular as the oscillations seen in the deterministic model.

We can estimate the impact of stochasticity by calculating the variability seen within a single realization of the model. A convenient way to express this variability is in terms of the coefficient of variation (cv), given by the standard deviation of the fluctuations divided by the mean. We generate a long timeseries, discard some transient period (so that the particular initial conditions chosen have little effect) and then use the remainder of the series to estimate variability.

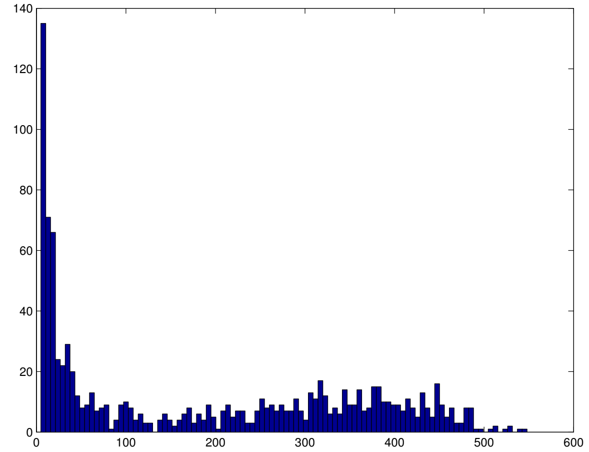

Figure 28: Outbreak size distributions with five infectives introduced into a population of 1000. R_0 taken to be 1.2.

An alternative way to estimate variability is to examine the behavior of a collection of realizations. After discarding some initial transient, we can calculate the standard deviation over the collection of realizations at a single point in time and hence the coefficient of variation. It turns out that these two approaches give the same answer: an average taken over time for a single realization equals an average taken at a single time point over a collection of realizations.

This measure of variability gives some insight into the probability of fadeout. Fadeout occurs when a realization wanders far enough away from the mean that the number of infectives hits zero. Fadeout is more likely to occur if variability is high, as trajectories wander further from the mean in such cases. If the variability is roughly the same size as the mean (i.e. cv comparable to 1) then fadeouts are likely to occur.

The variability that is seen within and between realizations depends on the population size. This effect can be illustrated by calculating variability at a number of different population sizes. We see that the coefficient of variation scales as $N^{-1/2}$: this behavior is reminiscent of the Central Limit Theorem of statistics.

We see that the coefficient of variation is sizeable compared to one for population sizes on the order of millions. The simple SIR model gives us a rough idea of the critical community size, although it overestimates its value. The simple model is slightly more prone to fadeout than the real world.

10.5 Brief Discussion of Analyses for Stochastic Models

Analyses of stochastic models are typically much more difficult than those of deterministic models. We need to describe how the probability distribution of S and I develops over time. We write

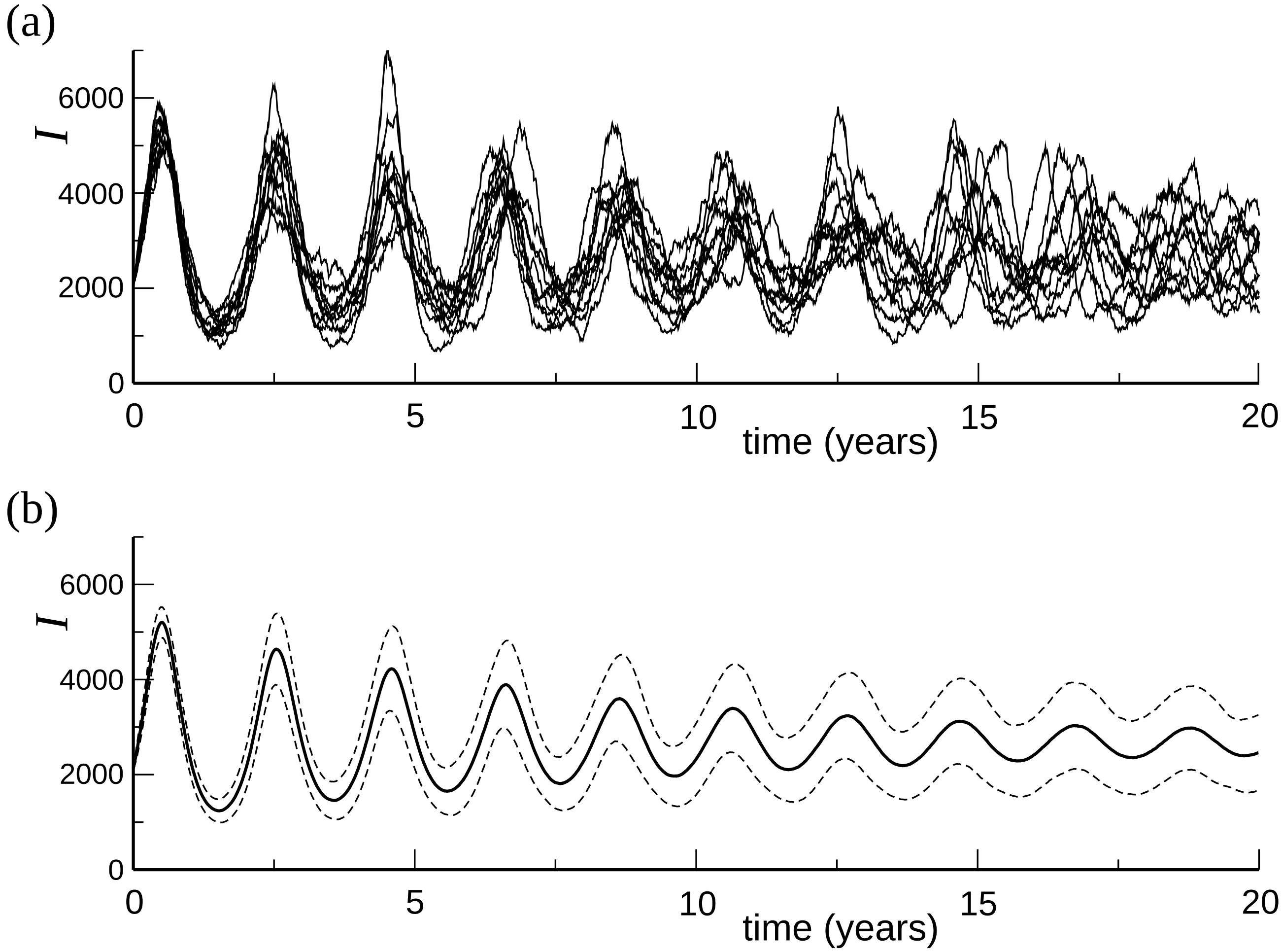

Figure 29: (a) Numbers of infectives seen in ten realizations of the stochastic SIR model for $N = 10^7$. Initially, the numbers of susceptibles and infectives were taken to be close to the equilibrium, (S^*, I^*), of the corresponding deterministic model, with $S = 1.05S^*$ and $I = 0.8I^*$.
(b) Estimates of the average (solid line) and standard deviation (illustrated as mean ± standard deviation) of the number of infectives seen in the stochastic model, obtained by averaging over 1000 realizations of the model, each of which were started at the same initial condition.
Parameter values are as follows: $\mu = 1/70$ year^{-1}, $\gamma = 50$ year^{-1}, $\beta = 750$ year^{-1}.

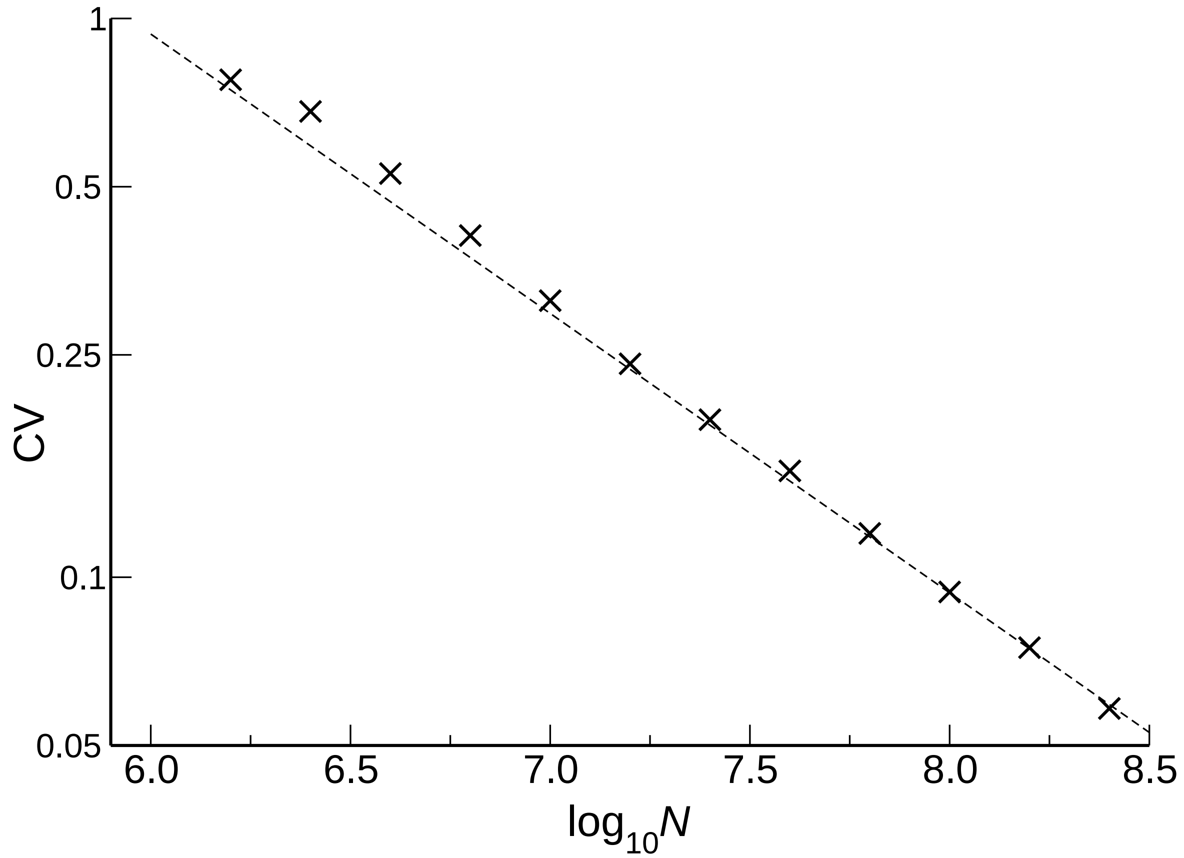

Figure 30: Variability seen about the equilibrium of the stochastic SIR model. Crosses denote variability estimated from model simulations. Notice the logged scales on both axes. Dashed line corresponds to a $N^{-1/2}$ scaling relationship. Parameters as in the previous figure.

$p_t(s,i)$ for the probability of finding the system with s susceptibles and i infectives at time t. For simplicity, we will ignore deaths for now, but it's easy to add them back in.

Let's look over a short time interval, from t to $t+dt$. How could we end up at (s,i) at the end of the time interval? From the Poisson process definitions, we know that there is essentially zero chance that two or more events occur over the interval. So, we could either have had one event or no events:

- We might have started off at $(s-1,i)$ and had a birth.
- We could have started off at $(s+1,i-1)$ and had an infection.
- We could have started off at $(s,i+1)$ and had a recovery.
- We could have started off at (s,i) and had no event.

Hence, we can write

$$\begin{aligned} p_{t+dt}(s,i) &= \text{prob(birth)}p_t(s-1,i) + \text{prob(infection)}p_t(s+1,i-1) + \text{prob(infection)}p_t(s,i+1) \\ &\quad +\text{prob(no event)}p_t(s,i) \qquad (70) \\ &= \mu N dt\, p_t(s-1,i) + (\beta/N)(s+1)(i-1)dt\, p_t(s+1,i-1) + \gamma(i+1)dt\, p_t(s,i+1) \\ &\quad +(1-\mu N dt - (\beta/N)sidt - \gamma idt)p_t(s,i). \qquad (71) \end{aligned}$$

We can rearrange this, subtracting $p_t(s,i)$ from both sides. We then notice that every term on the resulting right hand side has a factor dt, so we can divide both sides by dt and let dt approach zero.

We then have

$$\begin{aligned} \frac{d}{dt}\, p_t(s,i) &= \mu N\, p_t(s-1,i) + (\beta/N)(s+1)(i-1)\, p_t(s+1,i-1) + \gamma(i+1)\, p_t(s,i+1) \\ &- (\mu N + (\beta/N)si + \gamma i)p_t(s,i). \qquad (72) \end{aligned}$$

We have to be a little careful about boundary cases: we define $p_t(s,i)$ to be zero whenever s or i is negative. This makes sense biologically: for instance we cannot have a birth take us from $s=-1$ to $s=0$!

Equation (72) describes a set of coupled differential equations that describe how the probability distribution of S and I changes over time. This is known as the Kolmogorov forward equation. (In the physics and chemistry literatures, the Kolmogorov forward equation is usually called the Master equation.)

The Kolmogorov equations (72) are linear in $p_t(s,i)$, and so can be written as a matrix differential equation

$$\frac{d}{dt}\mathbf{p}_t = A\mathbf{p}_t. \tag{73}$$

(Some book-keeping is required to arrange the values of (s,i) into a single vector.)

The solution of this system of equations is simply $\mathbf{p}_t = \exp(At)\mathbf{p}_0$. Alternatively, the system could be integrated numerically. Given the number of equations involved in this system, analytic results are perhaps difficult to come by, although expressions have been derived for the probability distribution describing outbreak sizes in a closed population of small size. ('Small' here means perhaps less than 10 individuals!)

10.5.1 Linear Case: Estimation of the Probability of Invasion

As usual, things are easier when our model is linear, so we consider initial behavior. In the deterministic picture we have

$$dI/dt = \beta I - \gamma I. \tag{74}$$

In the stochastic picture, we have infections arising at rate βI and recoveries occurring at rate γI. As ever, we do not have to worry about the numbers of susceptibles in the linear setting. This means that each infective gives rise to secondary infections at rate β, regardless of the number of infectives.

This is a linear birth death process, commonly studied in stochastic process theory courses. We shall now outline some of this theory. We can derive the result of interest using either this continuous time stochastic process or using a closely connected discrete time process. We shall make use of the second approach, using **branching process theory**. Either approach makes heavy use of **probability generating functions**.

Probability Generating Functions

Suppose that we have a random variable, X, that takes non-negative integer values. The probability density function (pdf) of X gives the probabilities, p_k, that X takes the value k:

$$p_k = \mathrm{Prob}(X = k). \tag{75}$$

The probability generating function, $G(s)$, is a convenient way of summarizing the pdf of X and is defined as the following power series in the variable s

$$G(s) = \sum_{k=0}^{\infty} p_k s^k. \tag{76}$$

The probability generating function can, therefore, be written as an expectation

$$G(s) = \mathrm{E}(s^X) \tag{77}$$

and has several well known properties:

- $G(0) = p_0$
- $G(1) = \sum_{k=0}^{\infty} p_k = 1$
- $G'(s) = \sum_{k=0}^{\infty} kp_k s^{k-1}$, so $G'(1) = \sum_{k=0}^{\infty} kp_k = \mathrm{E}(X)$.
- In a similar way, higher derivatives of $G(s)$ can be used to find expressions for higher moments (such as the variance) of the distribution of X. For instance, $\mathrm{Var}(X) = G''(1) + G'(1) - \{G'(1)\}^2$.

Notice that in cases where X can take an infinite number of different values, the generating function is an infinite sum of terms and so the proof of some of the above results requires us to consider the convergence of the sum.

Example: Poisson Distribution. For a Poisson distribution with mean μ, we have that $p_k = \mu^k e^{-\mu}/k!$ and so

$$\begin{aligned} G(s) &= \sum_{k=0}^{\infty} s^k \frac{\mu^k}{k!} e^{-\mu} \\ &= e^{-\mu} \sum_{k=0}^{\infty} \frac{(\mu s)^k}{k!} \\ &= e^{-\mu} e^{\mu s} \\ &= e^{\mu(s-1)}. \end{aligned} \tag{78}$$

The branching process formulation considers the chains of infection that start with a single infective. In the language of branching processes, the secondary infections due to a given infective are called the offspring of an individual. (The branching process was originally developed to ask the question of whether a given family name would go extinct over time: much of the terminology of this theory derives from this genealogical context.) Even though the epidemic is really a continuous time process, we can create a discrete time process in terms of generations (see figure 31). These 'generations' refer to whether an infection is directly due to the original infective (first generation), their offspring (second generation), and so on. Notice that most of the temporal information is lost in this description: the temporal order of infections can differ from their order in terms of generations.

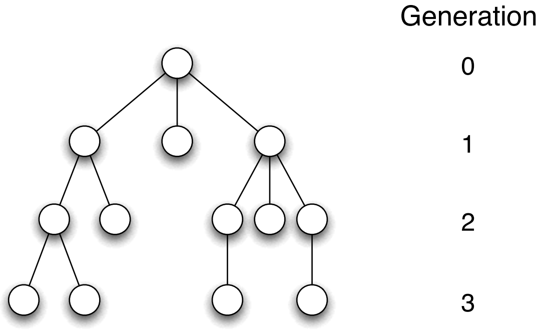

Figure 31: Branching process description of the initial stages of an outbreak.

We assume that the numbers of offspring (secondary infections) of each infective can be described by independent identically distributed random variables, X. This implies that the fate of different lineages (chains of infection) are independent. Notice how these assumptions are reliant on the linear description of the initial behavior of the epidemic: if the depletion of susceptibles were being accounted for then the average number of secondary infections would decrease over time. The numbers of offspring of different infectives would neither be independent nor identically distributed.

The main result of branching process theory states that the probability of the eventual extinction of the process, starting with one infective individual in generation zero, is given by the smallest non-negative root of the equation

$$G(s) = s, \tag{79}$$

where $G(s)$ is the probability generating function of the offspring distribution. Furthermore, the theory states that the probability of eventual extinction is 1 if the average number of offspring (secondary infections) is less than one, and is less than one if the average number of offspring is greater than one.

Notice the strong analogy between this result and the $R_0 = 1$ threshold from the deterministic setting. Also notice that the independence assumptions can be used to relate the probability of ultimate extinction starting from I_0 infectives to the probability of extinction starting from a single infective.

We need to calculate the probability generating function of the offspring distribution. If the infectious period was exactly t time units, the assumption that secondary infections arise at constant rate β would imply a Poisson distribution of the number of secondary infections, with mean βt. In other words, the probability of there being k secondary infections would be $(\beta t)^k e^{-\beta t}/k!$. The offspring distribution for an arbitrary infectious period distribution, $f(t)$ can be obtained by conditioning on t

$$P(X = k) \quad = \quad \int_0^\infty f(t) \frac{(\beta t)^k}{k!} e^{-\beta t}\, dt. \tag{80}$$

For the exponentially distributed infectious period, with mean $1/\gamma$, we have $f(t) = \gamma \exp(-\gamma t)$. The pgf is thus given by

$$\begin{aligned}
G(s) &= \sum_{k=0}^{\infty} s^k \int_0^\infty \gamma e^{-\gamma t} \frac{(\beta t)^k}{k!} e^{-\beta t}\, dt \\
&= \gamma \sum_{k=0}^{\infty} \int_0^\infty \frac{(\beta s t)^k}{k!} e^{-(\beta+\gamma)t}\, dt \\
&= \gamma \int_0^\infty \sum_{k=0}^{\infty} \frac{(\beta s t)^k}{k!} e^{-(\beta+\gamma)t}\, dt \\
&= \gamma \int_0^\infty e^{\beta s t} e^{-(\beta+\gamma)t}\, dt \\
&= \gamma \int_0^\infty e^{t\{\beta(s-1)-\gamma\}}\, dt \\
&= \frac{\gamma}{\gamma - \beta(s-1)}.
\end{aligned} \tag{81}$$

Hence, since the probability of ultimate extinction starting from one infective satisfies $G(s) = s$, we have that

$$\begin{aligned}
& s = \frac{\gamma}{\gamma - \beta(s-1)} \\
\Rightarrow\quad & \gamma = \{\gamma - \beta(s-1)\}s \\
\Rightarrow\quad & \gamma(1-s) = -\beta s(s-1) \\
\Rightarrow\quad & s = 1 \quad \text{or} \quad s = \gamma/\beta.
\end{aligned} \tag{82}$$

Thus we have that the probability of eventual extinction in the linear birth-death process is

$$\text{Prob(eventual extinction)} = \begin{cases} 1 & \text{if } \beta < \gamma \\ \left(\frac{\gamma}{\beta}\right)^{I(0)} & \text{if } \beta > \gamma \end{cases} . \tag{83}$$

This result says that we can still have extinction, even if β is greater than γ (i.e. $R_0 > 1$). Notice that in this case, the average number of infectives goes to infinity, even though individual realizations may go extinct. Extinction is more likely if β/γ is small or if $I(0)$ is small.

The linear birth-death process can be seen as an upper bound to the epidemic process, since in the nonlinear epidemic process the ‘birth’ (infection) term will be smaller as S is depleted. We can reinterpret extinction in the birth-death process as meaning that only a small outbreak has occurred (recall that extinction is much more likely to occur when the number of infectives is small). So we have

$$\begin{aligned} \text{prob(small outbreak)} &= 1 \qquad \text{if } R_0 < 1 \\ \text{prob(small outbreak)} &= (1/R_0)^{I(0)} \qquad \text{if } R_0 > 1 \\ \text{prob(large outbreak)} &= 1 - (1/R_0)^{I(0)} \qquad \text{if } R_0 > 1. \end{aligned} \tag{84}$$

Given that analyses are more difficult to undertake in stochastic settings, this sort of approach— in which one compares the behavior of a stochastic model to a simpler one which can be analyzed— is frequently used.

10.5.2 Equations for Moments of the Stochastic Process

The Kolomogorov equation can be used to find out how the mean of the distribution changes over time, or indeed how higher order moments of the distribution (e.g. the variance) change. As an example, the average value of I, known as the **expected value**, is given by

$$\mathrm{E}(I) = \sum_i i \ \mathrm{prob}(I = i), \tag{85}$$

and so we have

$$\frac{d}{dt}\mathrm{E}(I) = \sum_i i \ \frac{d}{dt}\mathrm{prob}(I = i). \tag{86}$$

We can therefore obtain the rate of change of the average number of infectives by multiplying the right hand side of (72) by i and summing over all possible values of i. This is not so difficult, but is a little fiddly. An important issue that would arise is that the equation for the average number of infectives will involve higher order moments (e.g. the covariance between S and I). This is a consequence of the nonlinear transmission term, keeping in mind that $\mathrm{E}(SI) = \mathrm{E}(S)\mathrm{E}(I) + \mathrm{cov}(S, I)$.

Another approach that is commonly seen is to use the Kolmogorov equation to derive a partial differential equation for the probability generating function, $G(t_1, t_2) = \sum_{s,i} t_1^s t_2^i p_t(s, i)$. In most cases, the pde cannot be solved analytically.

10.5.3 Alternative Derivation of Moment Equations

If we have some function f of the random variables S and I, then we write the average value of f (called the **expected value** of f) as $\mathrm{E}\,(f(S, I))$. The rate at which this average changes is given by

$$\frac{d}{dt}\mathrm{E}\,(f(S,I)) \;=\; \mathrm{E}\Big(\sum_{\text{events}} \text{rate of event} \times \text{change in } f \text{ due to event}\Big). \qquad (87)$$

For instance, we could think about how the average value of I changes. If there is an infection event, I increases by one. If there is a recovery or an infective death, I decreases by one. The corresponding rates are $\beta SI/N$ and $(\mu + \gamma)I$, so we have

$$\frac{d}{dt}\mathrm{E}(I) \;=\; \mathrm{E}\Big(\frac{\beta SI}{N} \times 1 + (\gamma + \mu)I \times (-1)\Big) \qquad (88)$$

$$\;=\; \frac{\beta}{N}\mathrm{E}(SI) - (\gamma + \mu)\mathrm{E}(I). \qquad (89)$$

Notice that this looks similar to the corresponding deterministic equation, except that $\mathrm{E}(SI)$ appears instead of SI.

We remark that in the stochastic setting, $\mathrm{E}(SI) = \mathrm{E}(S)\mathrm{E}(I) + \mathrm{cov}(S, I)$. The average value of a product is not, in general, equal to the product of the average values as one has to account for the covariance between variables. The equation for the average number of infectives involves a **second order moment**, namely the covariance. This coupling between moment equations of different orders is typical for **nonlinear** models.

We could go on to think about how the average value of I^2 changes, as this would let us calculate the variance of I. If there is an infection event, the value of I^2 increases to $(I + 1)^2$, an increase of $2I + 1$. If there is a recovery or infective death, the value of I^2 decreases to $(I - 1)^2$, a decrease of $2I - 1$. So we have

$$\frac{d}{dt}\mathrm{E}\Big(I^2\Big) \;=\; \mathrm{E}\Big(\frac{\beta SI}{N} \times (2I + 1) + (\gamma + \mu)I \times (-2I + 1)\Big) \qquad (90)$$

$$\;=\; \frac{\beta}{N}\Big(2\mathrm{E}(SI^2) + \mathrm{E}(SI)\Big) + (\gamma + \mu)\Big(1 - 2\mathrm{E}(I)\Big). \qquad (91)$$

We could also generate equations describing how the average values of quantities such as S, S^2 and SI change over time.

Notice that the equations for the second order moments involve third order moments, such as $\mathrm{E}(SI^2)$. This process could be continued indefinitely, giving equations for ever higher order moments. That would not be very convenient to work with, so what we do is **truncate** the system of moment equations at some order. We are typically most interested in the first and second order moments, so we employ some **moment closure** approximation that allows us to approximate third and higher order moments in terms of lower order moments. For instance, one might assume a bivariate normal distribution. The third order central moments of such a distribution (e.g. the skewness) are equal to zero, and this allows us to rewrite terms such as $\mathrm{E}(SI^2)$ in terms of lower order moments. This gives us a closed set of five equations for the averages of S, I, S^2, I^2 and SI.

The deterministic model corresponds to a lower order approximation, in which it is assumed that the variances of S and I, and their covariance, are equal to zero. With those assumptions, one has that the average value of S^2 is just equal to the square of the average value of S, the average value of SI is just the product of the average values of S and I, and so on.

The bivariate normal approximation leads to the following equations (written in terms of variances and covariances)

$$
\begin{aligned}
\frac{dE(S)}{dt} &= \mu N - \mu E(S) - (\beta/N)E(SI) && (92)\\
\frac{dE(I)}{dt} &= (\beta/N)E(SI) - (\gamma+\mu)E(I) && (93)\\
\frac{d\mathrm{Var}(S)}{dt} &= \mu N + \mu\left\{E(S) - 2\mathrm{Var}(S)\right\} + (\beta/N)\{E(SI) - 2E(I)\mathrm{Var}(S) \\
&\quad -2E(S)\mathrm{Cov}(S,I)\} && (94)\\
\frac{d\mathrm{Var}(I)}{dt} &= (\gamma+\mu)\left\{E(I) - 2\mathrm{Var}(I)\right\} + (\beta/N)\{E(SI) + 2E(S)\mathrm{Var}(I) \\
&\quad +2E(I)\mathrm{Cov}(S,I)\} && (95)\\
\frac{d\mathrm{Cov}(S,I)}{dt} &= -(\gamma+2\mu)\mathrm{Cov}(S,I) - (\beta/N)\{E(SI) - E(I)[\mathrm{Var}(S) - \mathrm{Cov}(S,I)] \\
&\quad +E(S)\left[\mathrm{Var}(I) - \mathrm{Cov}(S,I)\right]\}. && (96)
\end{aligned}
$$

Even these are not nice to work with... but they do allow us to get an idea of how the variance between realizations develops over time. As we've mentioned before, this is an important quantity to discuss when it comes to making predictions as it gives some idea of how far away from the mean we might expect to see individual realizations of the stochastic process. We can figure out how important stochastic effects are likely to be— in particular, we can address the question of how well the deterministic (average) behavior describes the system.

11 Seasonality in Transmission

Demographic stochasticity can maintain fluctuations in incidence, but these aren't as regular as the oscillations we saw in real-world data. For childhood infections, at least, variations in transmission across the year can give rise to more regular oscillatory behavior. Levels of transmission tend to be higher during school terms, when many children children spend a lot of time in close proximity to each other. Transmission is lower during vacations.

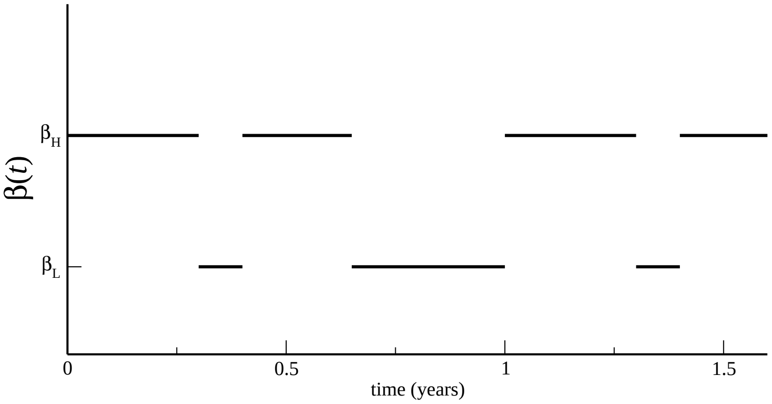

Figure 32: Seasonal variations in transmission: discrete level representation, depicting two school terms and vacations per year.

These seasonal variations in transmission can be modeled by allowing the transmission parameter, β, to be a function of time. In particular, we take $\beta(t)$ to be a function with period one year, representing annual forcing of our model. We could use a function whose level changes discontinuously between a high (term-time) and low (vacation-time) level (figure 32). Notice that we choose not to model high frequency variations in transmission: we do not distinguish between weekdays and weekends (or even daytime and night-time).

If we are only interested in broad-scale patterns, it turns out that the exact form of the seasonal function is not so important. Consequently, many authors choose to employ a sinusoidal description of seasonality

$$\beta(t) = \beta_0(1 + \beta_1 \cos 2\pi t), \tag{97}$$

where β_0 is the baseline level of transmission and β_1 determines the amplitude of the seasonal variation (we often talk about the 'strength' of seasonality).

11.1 Impact of Seasonality

We shall examine the impact of seasonality on the deterministic SIR model (with demography). We are thinking about the setting of measles, for which the unforced model exhibited damped oscillations with natural period close to two years. It is much more difficult to analyze the seasonally forced model than the unforced model. Perturbation analysis can be used, although this is typically only useful in the setting where β_1 is small. Instead, we shall use numerical simulations to get an idea of the behavior, simply simulating the model for different choices of parameters and initial conditions. A more thorough approach uses numerical bifurcation analysis.

The introduction of seasonality gives rise to multiannual oscillations in the model. If β_1 is small, then we see stable annual oscillations around what would have been the endemic equilibrium of the SIR model (figure 33). The endemic equilibrium is no longer stable. These oscillations are known as "passive oscillations", since the system is simply responding to the changing value of β. An important observation is that the presence of seasonality imposes a definite phase on the oscillatory solutions: peaks and troughs of prevalence occur at particular times of the year. The unforced model is left unchanged if we shift the origin of time (i.e. replace t by $t + t_0$, where t_0 is a constant). This means that the unforced model does not impose a particular phase on its damped oscillations: the timing of the peaks and troughs depends on the initial conditions chosen.

As β_1 increases, the amplitude of the oscillations increases (figure 34). As β_1 is increased further, the qualitative behavior of the system undergoes a change: instead of annual oscillations, we see biennial oscillations (35). This change is known as a period doubling bifurcation. (We remark that the annual oscillation is still there, but it has now become unstable so we do not see it in simulations. Unless we start off exactly on the annual solution, our simulation will approach the stable period two solution.) Close to the period doubling bifurcation, there is only a small asymmetry between the behavior in odd and even years.

As seasonality is further strengthened, the biennial oscillations become increasingly asymmetric: the outbreak sizes in odd and even years differ considerably (figure 36). These solutions look quite like the biennial patterns we saw in incidence data, although they are more regular (remember that this is a deterministic model).

Notice that β_1 is relatively small, but seasonality has a major impact. The impact of seasonality depends on the relationship between the natural period of the model and the annual period of seasonality. If the natural period were close to one year, then the resonance between the frequencies would lead to large amplitude oscillations. In our case, the fact that the natural period is close to two years (twice the seasonal period) means that it is relatively easy to excite a biennial oscillation. What we are seeing is an interaction between the (damped) oscillations of the unforced model and the seasonality. Seasonality leads to the maintenance of oscillations, in a way that is familiar from

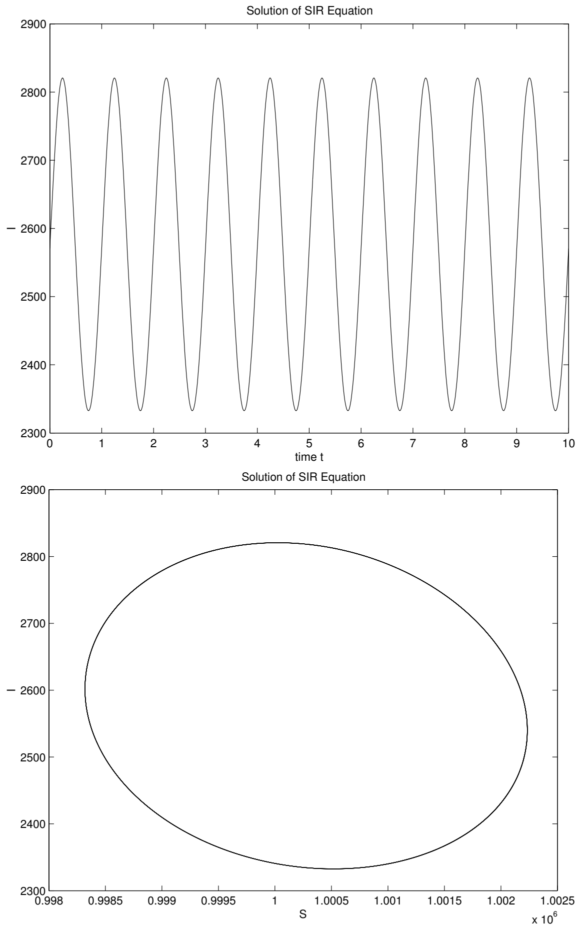

Figure 33: Small amplitude annual oscillations in the seasonally forced SIR model: prevalence versus time and phase plot (I vs S). Parameter values are as follows: $\beta_1 = 0.01$, $N = 10^7$, $\beta_0 = 500$, $\gamma = 50$ and $\mu = 1/70$.

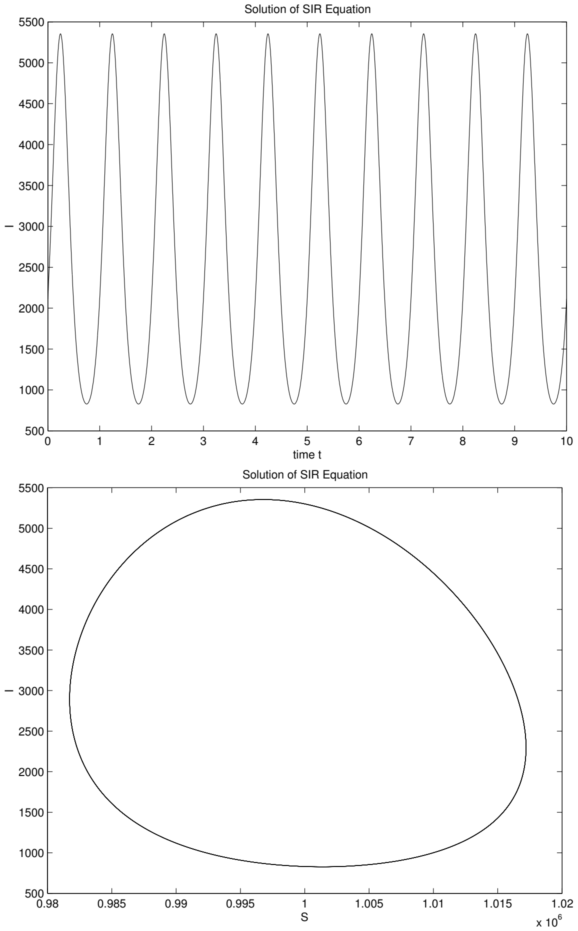

Figure 34: Larger amplitude annual oscillations in the seasonally forced SIR model seen when $\beta_1 = 0.10$: prevalence versus time and phase plot (I vs S). Other parameter values are as in the previous figure.

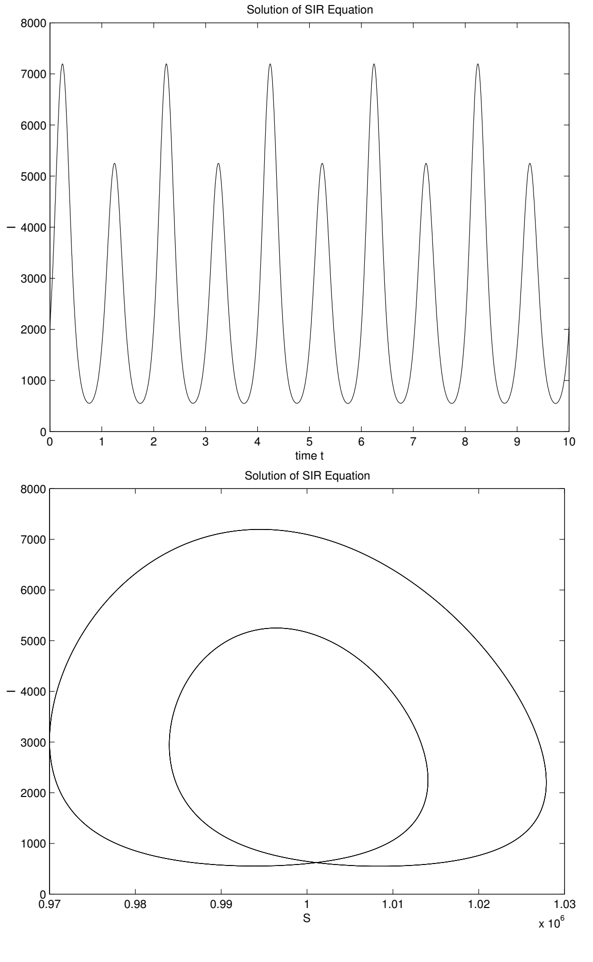

Figure 35: Biennial oscillations in the seasonally forced SIR model seen when $\beta_1 = 0.15$: prevalence versus time and phase plot (I vs S). Other parameter values are as in the previous figure.

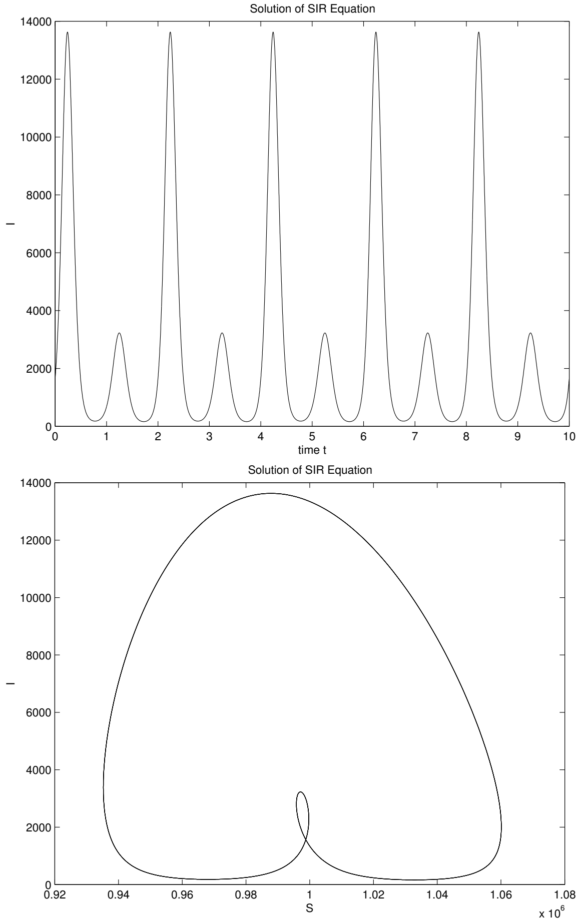

Figure 36: Highly asymmetric biennial oscillations in the seasonally forced SIR model seen when $\beta_1 = 0.20$: prevalence versus time and phase plot (I vs S). Other parameter values are as in the previous figure.

elementary treatments of forced oscillatory systems (see figure 37), or to anyone who has pushed a child on a swing.

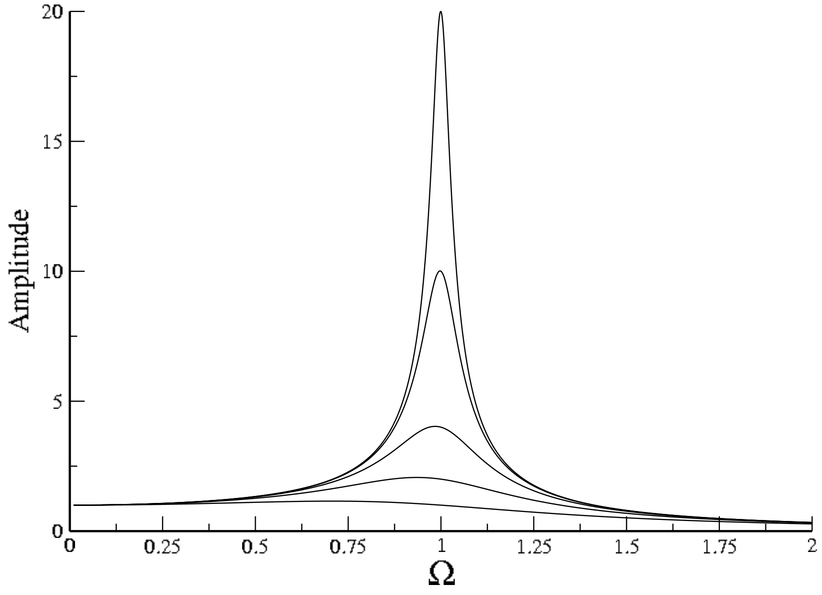

Figure 37: Frequency response curve of the damped linear oscillator subject to periodic forcing, as described by the second order ODE $\ddot{x} + k\dot{x} + \omega_0^2 x = F\cos\Omega t$. The non-negative constant k describes the strength of damping, with $k = 0$ depicting no damping. The non-negative constant F denotes the amplitude of the periodic forcing and Ω the (angular) frequency of the forcing. The natural frequency of the unforced system can be shown to equal $\sqrt{\omega_0^2 - k^2/4}$. For $F > 0$, the long-term behavior of the model is oscillatory with an amplitude that depends on F, ω, k and Ω. The curves on the graph show the oscillatory amplitude (relative to F) as Ω varies. ω is taken to equal one and we illustrate five different values of k: 0.05, 0.1, 0.25, 0.5 and 1.0 (upper curve through lower curve). For smaller values of k (more weak damping), we see a highly peaked curve when the frequency of forcing approaches the natural frequency, illustrating a resonance effect.

If β_1 is increased further, we get additional period doubling bifurcations: we see oscillations of period 4 (figure 38), 8, 16, ... Successive period doubling bifurcations occur with smaller increases in β_1. This period doubling cascade eventually leads to “chaotic” behavior: we get prevalence timeseries that appear to exhibit some degree of randomness (for instance the heights of the peaks are irregular) despite their being generated by a deterministic (non-random) system. This was suggested as a mechanism by which the irregularities that we saw in the timeseries could be generated. (But notice that we know that there is randomness in reality, so we need not invoke this mechanism.) An

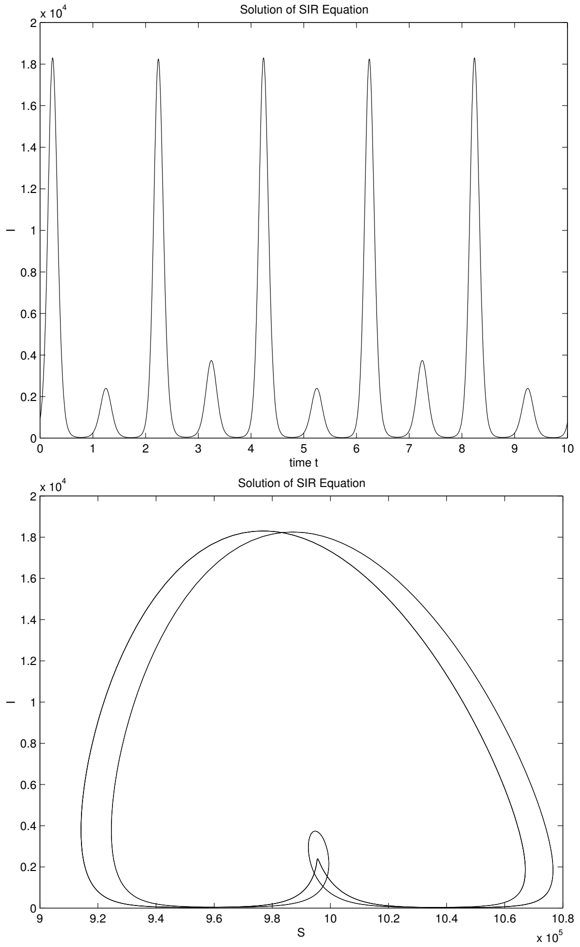

Figure 38: Four-yearly oscillations in the seasonally forced SIR model seen when $\beta_1 = 0.30$: prevalence versus time and phase plot (I vs S). Other parameter values are as in the previous figure.

important feature of chaotic systems is that they exhibit sensitive dependence on initial conditions. If we vary the initial number of infectives, even fractionally, we would get two simulations whose timecourses diverged over time.

The model often exhibits other stable solutions, such as period three behavior, that co-exist with the solution described above (figure 39). (We remark that triennial oscillations typically have large amplitude.) We talk about having "multiple attractors": in such cases, the **qualitative** behavior that the model exhibits will depend on the initial conditions that we choose for the simulation.

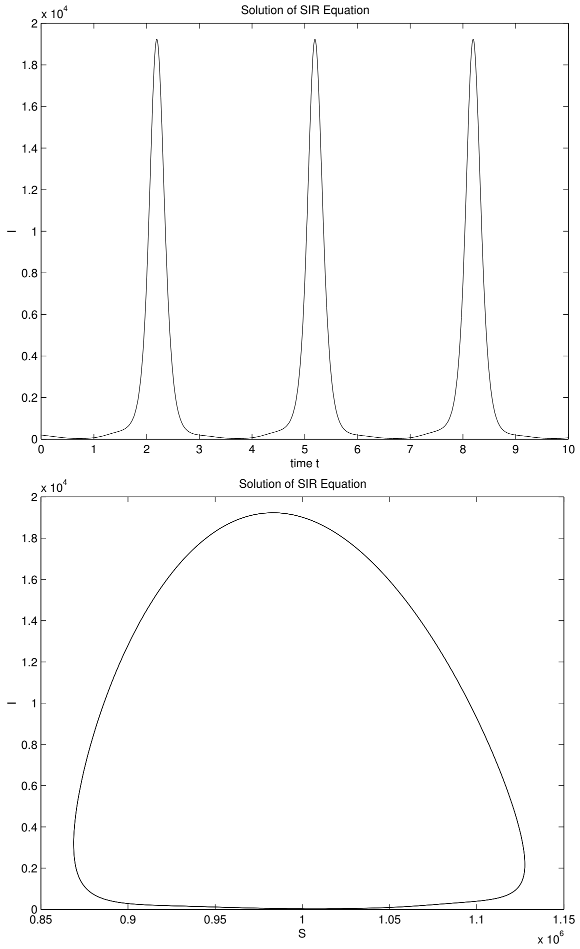

Figure 39: Period three attractor that co-exists with the annual attractor in the seasonally forced SIR model when $\beta_1 = 0.05$: prevalence versus time and phase plot (I vs S). Other parameter values are as in the previous figure.

The behavior of the model could be summarized by plotting the behavior in terms of the strength of seasonality, β_1. (We call such plots bifurcation diagrams.) To highlight the multiannual nature of the oscillations, we plot the value of I at the start of each year for different values of β_1. For a given value of β_1, annual behavior will give us a single point on our plot, biennial behavior will give us two. We also denote the stability of various solutions using solid or broken curves (stable and unstable, respectively). Figure 40 summarizes some of the behavior described above.

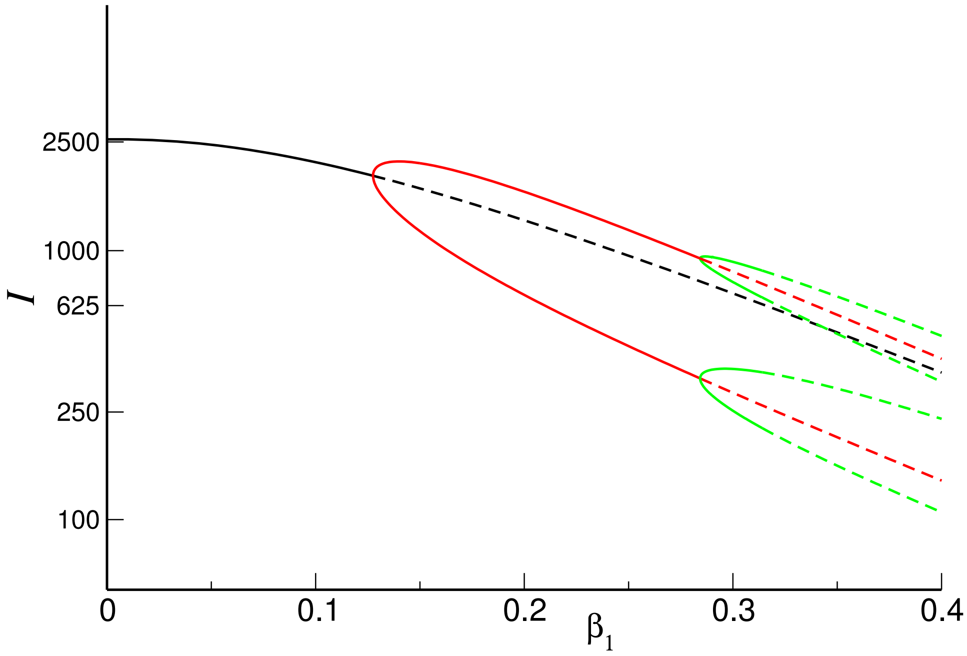

Figure 40: Bifurcation diagram showing the main annual, biennial and four-yearly attractors for the SIR model as β_1 changes. Black curves denote annual behavior, red curves denote biennial and green curves denote four-yearly solutions. Solid curves denote stable oscillations and broken curves denote unstable oscillations. The start of the period doubling cascade is visible. Notice that there are other stable solutions that are not shown on this figure. Other parameter values are as in the previous figure.

We see qualitative changes in behavior at some points on the diagram, for instance where stable biennial behavior replaces stable annual behavior. Such points are known as bifurcation points.

We could also plot behavior as β_0 varies (for a fixed strength of seasonality). This is of interest because of β_0's relation to R_0: we can use changing values of β_0 as a surrogate for either vaccination or changes in birth rate. The interesting observation (figure 41) is that either low or high values of β_0 give rise to annual, rather than biennial cycles. This means that vaccination could take us from biennial to annual behavior (decreasing β_0) and that increasing birth rates tends to lead to annual behavior. This second result is intuitively clear: if the birth rate is high then the susceptible pool is rapidly replenished and so large outbreaks can occur every year.

We could also make plots of β_1 versus β_0 that depict the curves at which various bifurcations

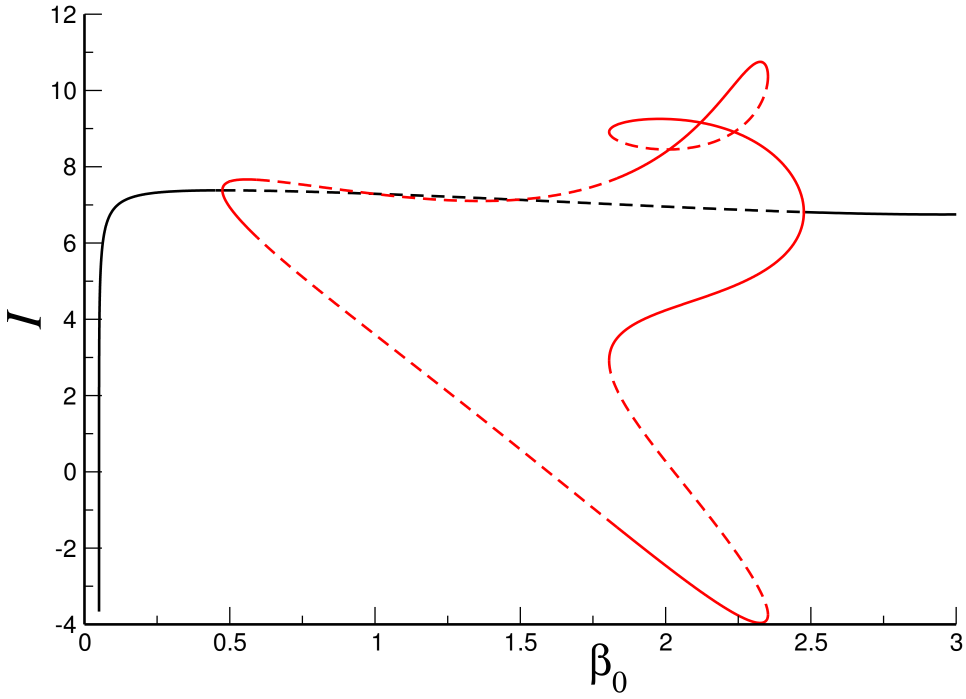

Figure 41: Bifurcation diagram showing the changes in behavior that occur as β_0 varies when β_1 is taken to equal 0.16. Black curves denote annual behavior and red curves denote biennial. Only the main annual and biennial behaviors are shown, other stable solutions also exist but are not depicted here. Notice that there are two co-existing stable biennial solutions towards the right hand side of the biennial regime. Other parameter values are as in the previous figure.

occur. For instance, we could have the curve that gives the β_1 value at which the period doubling bifurcation occurs for various values of β_0.

11.1.1 Seasonality and Stochastic Models

The inclusion of seasonality leads to very low numbers of infectives in the inter-epidemic troughs (i.e. between outbreaks). As in the unforced model, the deterministic nature of the model means that the infection will not go extinct even though the prevalence falls well below a fractional number of individuals.

We earlier discussed the critical community size phenomenon. Seasonality, since it deepens the inter-epidemic troughs, increases the chance of extinction so we need a much larger population size in order to guarantee persistence of the infection in the model. Simple seasonally forced SIR models (or SEIR models) have difficulty in correctly predicting the critical community size: in the presence of seasonality, they considerably overestimate the population sizes required for persistence.

One of the major themes in the development of models for measles has been to reconcile non-trivial dynamics with persistence of the infection and to produce models that can simultaneously reproduce realistic patterns of incidence and persistence.

Seasonally forced stochastic models raise the potential for interaction between nonlinear dynamics and stochasticity. This can lead to some quite intricate dynamics. This area remains largely unexplored.

11.2 Estimating Seasonal Variations in the Transmission Parameter

So far, we have assumed some particular form for the seasonally varying transmission parameter, $\beta(t)$, and examined its impact on prevalence. Is it possible to do the inverse problem: infer the seasonal variations in β from a given disease time series? Several authors (Fine and Clarkson in 1982, and more recently Finkelstädt and Grenfell) have attempted to do this.

What problems arise when one tries to estimate β from incidence (or prevalence) data? We assume that a simple SIR process is sufficient to describe the system. Transmission involves both the numbers of infectives and susceptibles, but typically we will only have data on the former. One way around this problem is to notice that each infection means that S decreases by one and I increases by one. Assuming that we know about every infection event, we can keep track of the number of susceptibles. But our data is subject to under-reporting. If we knew what fraction of cases were reported, we could maybe correct for this.

For measles, it turns out that estimating the level of reporting is not so hard to do. Before vaccination, almost everyone would get measles as a child. (This was confirmed by serological surveys of the population, testing for the presence of measles antibodies.) Consequently, the number of cases over some time period could be related to the number of births over the time period. Another method for estimating the level of under-reporting uses the unfortunate fact that a certain proportion of measles cases lead to severe side-effects. Such children require hospitalization, and so it is possible to standardize the incidence data against the hospitalization rate. In the U.K., roughly 65 percent of all measles cases were reported in the pre-vaccine era.

Another issue concerns the frequency at which the data is sampled: epidemiological data often involves monthly or weekly reports of numbers of cases. This is fortunate for measles researchers since they can make use of the fact that the infectious period of measles is roughly equal to one week. (The Finkelstädt and Grenfell analysis make use of the fact that the sum of the latent and infectious period is approximately two weeks.) This means that one can use a discrete time model, in which infectious individuals are assumed to recover by the next week.

Fine and Clarkson use the following equations:

$$\begin{aligned} C_{t+1} &= C_t S_t r_t && (98) \\ S_{t+1} &= S_t - C_{t+1} + B_t. && (99) \end{aligned}$$

Here C_t denotes the number of cases (infectious individuals) in week t, S_t denotes the number of susceptibles. Notice that the number of cases in the next week is assumed to follow a mass-action type description with transmission parameter r_t. All current cases are assumed to recover by the next week, and the number of susceptibles is depleted by those that are infected but replenished by births, B_t, which they estimate from demographic data. They then estimate r_t by rearranging equation 98. Since the data are noisy, Fine and Clarkson work with the "average biennial measles pattern", which they obtain by averaging the sixteen years of incidence data that they have over a two-yearly window.

It should be kept in mind that it just happens that measles data is particularly amenable to this kind of analysis: estimation of transmission parameters is much more delicate in general.

12 Disease Outbreaks in Small Populations: Chain Binomial Models

Many infections spread rapidly within a family because of the large amount of time that family members spend in close proximity to each other. We will look at a class of models that have been used to describe the transmission of infection in family settings. Since the number of individuals in our household population is small, a stochastic model must be employed.

Chain binomial models assume that the infectious period is short compared to the latent period of the infection. Even if the infection does not conform to this assumption, these models could still be used if individuals were quickly removed from circulation upon becoming symptomatic (keeping in mind the proviso that the duration of latency may differ from the length of the incubation period). Furthermore, it is also assumed that individuals cannot be infected more than once (recovery might lead to permanent immunity, or immunity that lasts longer than the timescale of the outbreak).

In this setting, there are successive generations of infectives in the household (see figure 42) so it is convenient to employ a discrete time model, with time step equal to the latent period. Notice that the successive generations description is unlikely to hold if more than one individual in the household acquires infection from outside the family. So we imagine that there is a single introduction of infection, rather than multiple introductions: this is likely to be the case if within-family transmission is much more probable than between-family transmission, and/or the prevalence of infection is low in the general community.

The course of the outbreak within a family can be described as a chain of infection. We track the number of infectives in each generation, written as i_k. We refer to the initial (introduced) infection as generation zero, and write $i_0 = 1$. The chain ends when there are no infections in some generation: this could be because everyone has already become infected. As an example, we might have the chain $1 \to 1 \to 2$: this means that, following the introduction, there was one infective in the first generation, two infectives in the second generation and the chain then ended. Overall, there were four cases in the household.

The model assumes that contacts are made independently within the family and that each susceptible is equally likely to acquire infection. For a given generation, the probability that a susceptible escapes infection is written as q_i, where i denotes the number of infectives. The chance that the susceptible acquires infection in the generation therefore equals $1 - q_i$, which we write as p_i.

If there are i infectives and s susceptibles remaining, we can calculate the chance that exactly x of them acquire infection as

$$(p_i)^x (q_i)^{s-x} \binom{s}{x}, \tag{100}$$

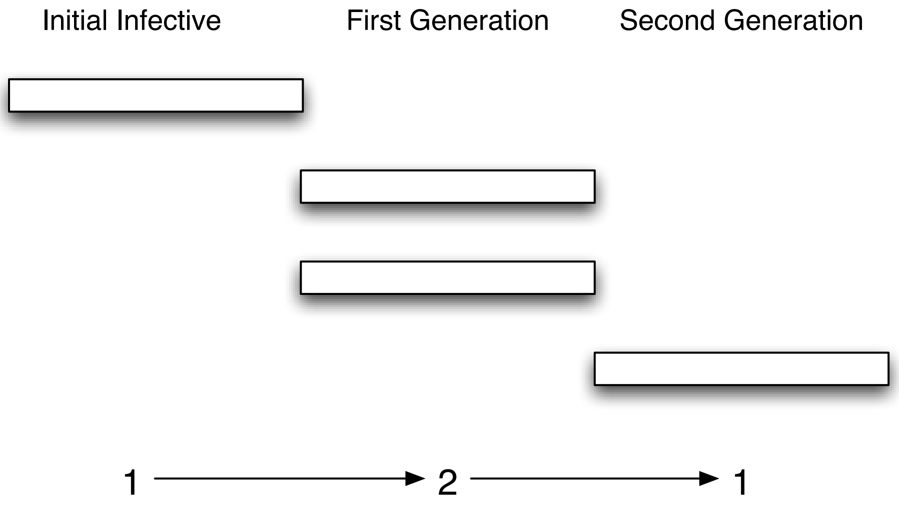

Figure 42: Example chain of infection as considered by a chain binomial model. Following the introduction of a single infective, two infections are seen in the first generation followed by a single infection in the second generation. This chain would be written as 1 → 2 → 1.

where $\binom{s}{x}$ is a binomial coefficient, whose value equals $s!/\{x!(s-x)!\}$. This expression is obtained by multiplying the probability that x individuals become infected, $(p_i)^x$, the probability that the rest (of which there are $s-x$) escape infection, $(q_i)^{s-x}$, and the number of different ways of choosing x individuals out of a total of s people.

We notice that the number of infections in the generation is described by a binomial distribution, with s individuals each of whom have probability of infection equal to p_i. This explains why the model is known as a chain binomial model: at each generation we have a binomial distribution of infectives. This distribution typically differs from generation to generation as the numbers of infectives and susceptibles change over the course of the outbreak.

If the number of infectives in generation t is written as I_t and the number of susceptibles as S_t, then we have that

$$\text{Prob}(I_{t+1} = x \mid S_t = s,\ I_t = i) = \frac{s!}{x!(s-x)!} p_i^x q_i^{s-x} \quad x = 0, 1, \ldots, s. \tag{101}$$

The number of susceptibles remaining can then be calculated as

$$S_{t+1} = S_t - I_{t+1}. \tag{102}$$

The probabilities that different chains occur can then be calculated in a step-by-step fashion.

As a simple example, we consider a household of size three. Since there is a single initial infective,

Chain	Probability of Chain
1	q_1^3
$1 \to 1$	$3q_1^4p_1$
$1 \to 1 \to 1$	$6q_1^4p_1^2$
$1 \to 2$	$3q_1p_1^2q_2$
$1 \to 1 \to 1 \to 1$	$6q_1^3p_1^3$
$1 \to 1 \to 2$	$3q_1^2p_1^3$
$1 \to 2 \to 1$	$3q_1p_1^2p_2$
$1 \to 3$	p_1^3

Table 3: Possible chains for a household of size four, following a single initial introduction

there are only four possible chains: 1, $1 \to 1$, $1 \to 1 \to 1$, $1 \to 2$. We can easily calculate the probabilities of each of these chains:

- 1. There are no secondary infections: both susceptibles escape infection, which occurs with probability q_1^2.

- $1 \to 1$. There is one infection in the first generation, and so one susceptible escapes: this occurs with probability $2p_1q_1$. Then there are no infections in the second generation: the one remaining susceptible escapes infection. This occurs with probability q_1. The probability of this chain occurring is, therefore, equal to $2p_1q_1^2$.

- $1 \to 1 \to 1$. There is one infection in the first generation, and so one susceptible escapes infection: this occurs with probability $2p_1q_1$. Then there is one infection in the second generation: the one remaining susceptible acquires infection. This occurs with probability p_1. The probability of this chain occurring is, therefore, equal to $2p_1^2q_1$.

- $1 \to 2$. There are two infections in the first generation, which occurs with probability p_1^2.

It is straightforward to check that these four probabilities sum to one.

We can repeat this exercise for the possible chains that result when one individual introduces infection into a household of size four (table 3). Notice the appearance of p_2 and q_2 in the probabilities: these appear in chains when there are 2 infectives present in some generation when one or more susceptibles remain.

The probabilities that we have calculated can be used to write down the outbreak size distribution that results from the introduction of a single infective into a household.

Outbreak size	Household of size 3	Household of size 4
1	q_1^2	q_1^3
2	$2q_1^2p_1$	$3q_1^4p_1$
3	$p_1^2(1+2q_1)$	$3q_1p_1^2(2q_1^3+q_2)$
4		$1-\{q_1^3+3q_1^4p_1+3q_1p_1^2(2q_1^3+q_2)\}$

Table 4: Outbreak size distributions that follow from the introduction of a single initial infective into households of sizes 3 or 4

Our description of the outbreak size description will be completed once we have decided how to model the p_i and q_i. Two assumptions are commonly made:

- The Reed-Frost assumption takes q_i to equal q_1^i. This corresponds to assuming that each infectious individual independently gives rise to secondary infections. The chance of escaping infection decreases as the number of infectives increases. This form of the probability is appropriate for infections that are spread by direct person to person contact. For notational simplicity, we write $q_1 = q$.

- The Greenwood assumption takes q_0 to equal one, and all the remaining q_i (i.e. for $i > 0$) to equal a common value, written as q. This means that the chance of escaping infection does not depend on the number of infectives, as long as one or more infective is present. This describes a saturation effect, which might be a good description of a highly infectious disease in a close-proximity household setting.

Notice that both models have only one parameter, q.

Using the Reed-Frost and Greenwood models, we can obtain the outbreak size distributions for families of size 4 in terms of the single parameter q. It should be kept in mind that the interpretation of q is different in the two models.

12.1 A Short Diversion: Deterministic Chain Binomial Models

Deterministic models simply look at the average numbers of susceptibles and infectives in each generation. In the chain binomial setting, if there are s susceptibles and i infectives, the average number of infections in the next generation will be $s(1-q_i)$ and the average number of remaining susceptibles will be sq_i.

	Reed-Frost Model	Greenwood Model
Outbreak size		
1	q^3	q^3
2	$3(1-q)q^4$	$3(1-q)q^4$
3	$3(1-q)^2q^3(1+2q)$	$3(1-q)^2q^2(1+2q^2)$
4	$(1-q)^3(1+3q+6q^2+6q^3)$	$(1-q)^3(1+3q+3q^2+6q^3)$

Table 5: Outbreak size distributions that follow from the introduction of a single initial infective into a household of sizes 4, based on the Reed-Frost and Greenwood models. These expressions are obtained by substituting the Reed-Frost and Greenwood formulae for q_i into the expressions in the final column of table 4.

We can solve for the average behavior of the Greenwood model. Since $q_i = q$ for all non-zero i, then if we have s susceptibles in one generation, the average numbers of susceptibles and infectives in the next will be sq and $s(1-q)$, respectively. Hence if we begin with s_0 susceptibles at time $t = 0$ then the average number of susceptibles at time $t > 0$ will equal s_0q^t and the average number of infectives will be $s_0q^{t-1}(1-q)$. The numbers of both susceptibles and infectives decrease geometrically for $t > 0$, and we see that both numbers approach 0 as t increases. In particular, the entire population becomes infected, unlike the deterministic models with recovery that we saw earlier. Furthermore, this situation holds whenever q is non-zero.

We cannot solve the deterministic Reed-Frost model in closed form, but we can easily simulate its behavior numerically. This model behaves in a similar way to the standard deterministic SIR model and we see an epidemic that has a familiar shape. We also notice that, in contrast to the Greenwood model, the number of susceptibles tends to a non-zero value as time gets larger.

An important observation is that the discrete-time equations do not take the following form that might be obtained by discretizing the mass-action term

$$S_{t+1} = S_t - \beta S_t I_t / N \tag{103}$$

$$I_{t+1} = \beta I_t S_t / N. \tag{104}$$

One obvious problem with this model is that if β were large enough then the number of susceptibles could become negative in a given time step. (Notice that the Fine and Clarkson model used to estimate seasonality is of this incorrect form.)

One has to be more careful than this when discretizing time. The general form of a discrete time infection process will take the number of new cases to equal

$$I_{t+1} = S_t \left\{ 1 - f\left(1 - \frac{I_t}{N-1}\right) \right\} \tag{105}$$

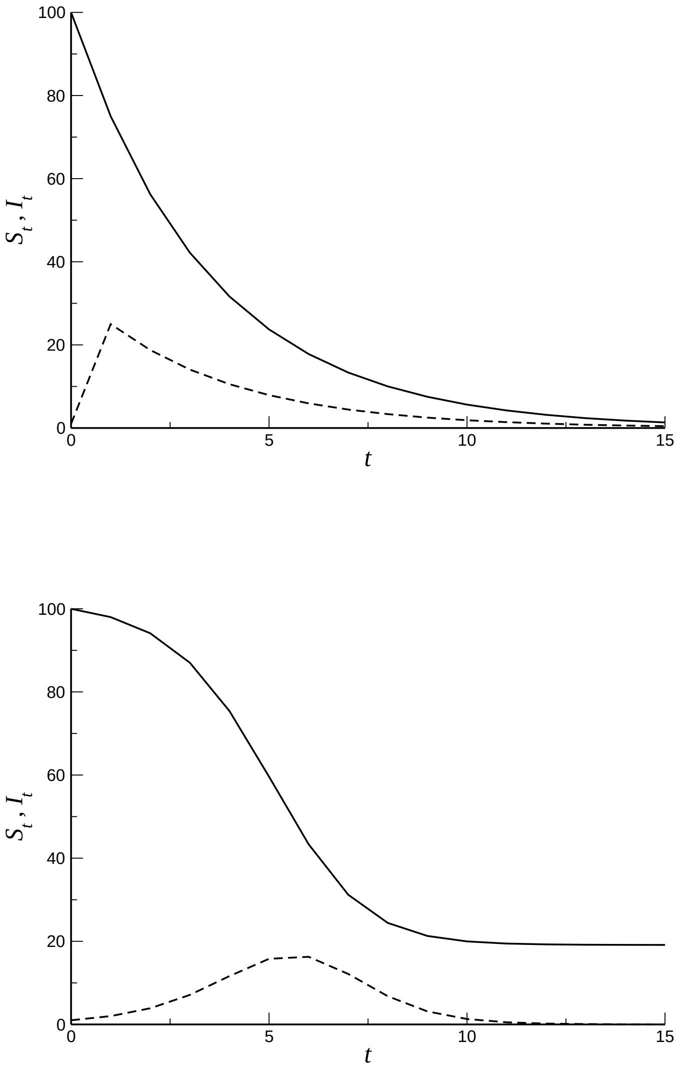

Figure 43: Dynamics of the deterministic Greenwood (upper panel) and Reed-Frost models (lower panel). The parameter q takes the value 0.75 in the Greenwood case and 0.98 in the Reed-Frost case. In both cases, we take the initial number of susceptibles to equal 100 and introduce a single infective.

as discussed by Dietz and Schenzle. Here, the value of the function f lies between 0 and 1. With this infection term the number of infectives can never exceed the number of susceptibles and so the numbers of individuals in all all groups remains non-negative.

Here, the function f describes how the infection probability depends on the number of infectives in the population. Dietz and Schenzle point out that the Reed-Frost model can be derived from this general description if one assumes a Poisson distribution for the number of contacts made between individuals.

12.2 Fitting Outbreak Size Data Using Chain Binomial Models

We can use these models to analyze data on outbreak sizes in households. A classic study of measles (and scarlet fever) outbreaks in Providence, Rhode Island, provides us with information on the sizes of outbreaks seen in families of different sizes. Restricting attention to households of particular sizes, we can fit the above chain binomial models by comparing the outbreak size distribution predicted by different values of q with those seen in reality. This fitting procedure will give us the most appropriate value of q and, we hope, some measure of our uncertainty in the true value of q.

Several different procedures could be employed to estimate the best fitting value of q. We could calculate measures of discrepancy between the model predictions and the observed data, such as $\Sigma_i(O_i - E_i)^2$ or $\Sigma_i(O_i - E_i)^2/E_i$, where the O_i denote the observed data and E_i the predicted values. The best fitting q could then be found by minimizing either measure of discrepancy.

An alternative approach, based on a sound statistical footing, involves the likelihood function.

Likelihood Functions

Introduction

We introduce the idea of likelihood by using a simple (non-infectious disease) example, namely coin tossing. We know that the distribution of the number of heads obtained in n independent coin tosses is Binomial(n, p) if the probability of obtaining a head is p each time. Suppose we carry out n such coin tosses and see n_1 heads (and hence $n - n_1$ tails). What is our best guess at the value of p?

If we knew the value of p, the probability of obtaining n_1 heads and $n - n_1$ tails would equal

$$\frac{n!}{n_1!(n-n_1)!}p^{n_1}(1-p)^{n-n_1}. \tag{106}$$

Suppose we don't know p but we have observed n_1 heads. We talk about the likelihood of p given the observed number of heads

$$L = \frac{n!}{n_1!(n-n_1)!} p^{n_1}(1-p)^{n-n_1}. \tag{107}$$

We remark that L isn't a probability. A set of probabilities would sum to one, but this isn't the case if we summed over p for a given n_1. (This would be the case if we summed over n_1 for a fixed p.) Notice that we could normalize L so that its integral over p between 0 and 1 equalled one, but we choose not to do this: it is easy to see that such a normalization constant would not play an important role in the following analysis.

The idea will be to maximize L, that is to say to pick the value of p that maximizes the probability of observing the data given p.

It is more convenient to work with the logged likelihood function

$$\begin{aligned} \log L &= \log n! - \log n_1! - \log(n-n_1)! + n_1 \log p + (n-n_1)\log(1-p) \\ &= c + n_1 \log p + (n-n_1)\log(1-p). \end{aligned} \tag{108}$$

Here c is a constant whose value does not depend on p.

Figure 44 shows a plot of $\log L$ as a function of p in two example cases: 200 coin tosses in which 120 heads were observed and 1000 coin tosses in which 600 heads were observed. In both cases, the frequency of heads is equal to 0.6, but because it involves a larger number of tosses, we imagine that the second experiment will give us a better estimate of p.

We can find the maximum of the logged likelihood (and hence of the likelihood) by differentiating with respect to p

$$\frac{\partial \log L}{\partial p} = \frac{n_1}{p} - \frac{n-n_1}{1-p}, \tag{109}$$

and setting the derivative equal to zero. It is easy to see that the maximum occurs when $p = n_1/n$. We sometimes denote this **maximum likelihood estimate** (**MLE**) of p by with a hat: $\hat{p} = n_1/n$.

The shape of the log-likelihood near the maximum is important. If the function relatively flat, then nearby values of p are fairly equally likely: the model fit is not so much worse as p is moved some distance away from $\hat{p}$. In such cases, we would have a fair amount of uncertainty in the true value of p. Put another way, the standard error for our estimate of p would be large. If the log-likelihood has a sharper maximum, we would have more confidence in our estimate of $\hat{p}$.

This notion can be captured by the second derivative of the log-likelihood

$$\frac{\partial^2 \log L}{\partial^2 p} = -I, \tag{110}$$

where I is called the information. Notice that the second derivative must be non-positive for the stationary point to be a maximum of L, so that I is a non-negative quantity.

Theoretical results show that as the sample size, n, gets larger and larger, then the product of the square root of the sample size and the difference between the maximum likelihood estimate of p and its true value, p_0, becomes Normally distributed. The variance of the distribution is given by the reciprocal of the information, calculated at p_0. In other words

$$\sqrt{n}(\hat{p}_n - p_0) \to N\left(0, 1/I(p_0)\right). \tag{111}$$

This can be seen as a “Central Limit Theorem” type result. It can be used to give a value for the standard error of our estimate $\hat{p}$. Notice that the variance (and hence standard error) will decrease as the information increases (see figure 44).

More Complex Settings and Practical Considerations

Unfortunately, use of the likelihood approach is typically more difficult than this simple example suggests. For instance, if our model has several parameters, p_1 through p_k, we would have to we have to solve k simultaneous equations of the form $\partial L/\partial p_i = 0$. (We would also have to be sure that the solution corresponds to a maximum point.)

In general, we may not be able to find the maximum point of the (log) likelihood function explicitly. Instead, we may have to adopt a numerical approach. This introduces a whole raft of additional problems: maximizing a general nonlinear function is a non-trivial process. In the above example, our likelihood function had a single maximum point, but in a general setting, the potential existence of multiple local maxima complicates matters considerably.

In a multi-parameter setting, the information matrix is defined as minus the matrix of second partial derivatives of the log-likelihood function, evaluated at the maximum point

$$I_{ij} = -\frac{\partial^2}{\partial p_i \partial p_j} \log L. \tag{112}$$

In an way analogous to the one parameter setting discussed above, the inverse of the information matrix can be used to find the variances and covariances of the parameter estimates. Notice that the appearance of covariances means that, in general, estimates of the different parameters are not independent.

The likelihood approach can be used to compare model fits in a systematic way. If we have a model that has two parameters, p_1 and p_2, we can compare how well this model fits to a second model in which p_2 is fixed at some value, but for which p_1 is still allowed to vary. It turns out that if L_2 denotes the likelihood for the 2 parameter model and L_1 denotes the likelihood for the

1 parameter model, then the quantity

$$2\left(\log L_2 - \log L_1\right)$$

is (for large sample sizes) distributed according to a chi-squared distribution with one degree of freedom. Hence it is possible to test whether the two parameter model provides a significantly better fit than the single parameter model.

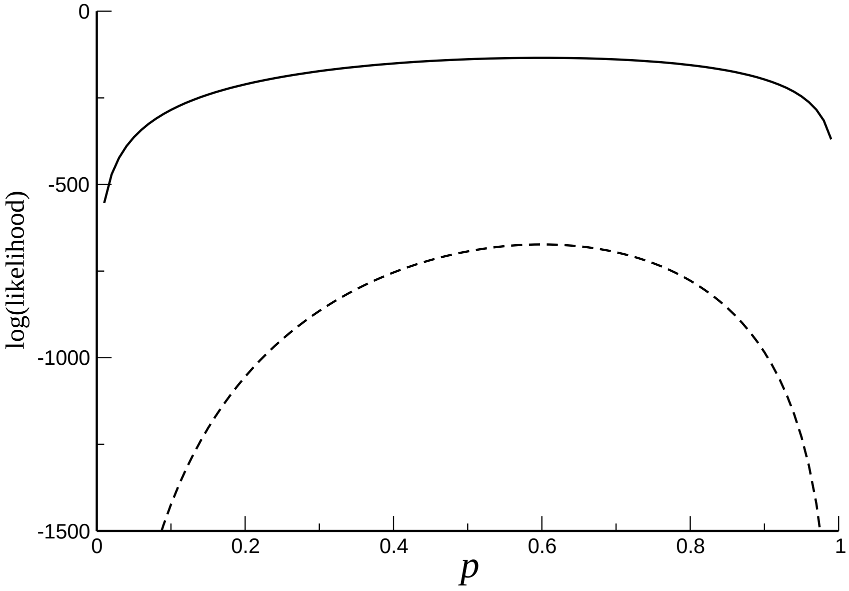

Figure 44: Maximum likelihood estimation of the parameter p of a binomial distribution (which might result, for example, from repeated independent coin tosses). Curves of $\log L$ as a function of p are shown for two sample datasets, one of size 1000 (lower curve) and one of size 200 (upper curve). In order to help make a comparison, the frequency of success (e.g. heads) is equal to 0.6 in the sample. Notice that the lower curve (larger sample size) is more sharply peaked about its maximum, with a larger second derivative (greater curvature) and hence corresponds to a smaller uncertainty in the true value of the parameter p.

12.3 Applying the Maximum Likelihood Approach to Outbreak Size Data

The likelihood approach can be used to fit the Providence outbreak size data. We shall restrict our attention to households of size four. Suppose we observe n_1 families in which only one case was observed, n_2 families with two cases, n_3 families with three cases and n_4 families in which all members became infected. Writing ϕ_i as a shorthand for the probability of having an outbreak of size i, we can write the likelihood function as follows

$$L = c\phi_1^{n_1}\phi_2^{n_2}\phi_3^{n_3}\phi_4^{n_4}. \tag{113}$$

Here c is a constant, whose value will be unimportant in what follows. Expressions for the ϕ_i, under the Reed-Frost and Greenwood models, have already been given in table (5) in terms of the single parameter q.

The log-likelihood is then given by

$$\log L = \log c + n_1 \log \phi_1 + n_2 \log \phi_2 + n_3 \log \phi_3 + n_4 \log \phi_4. \tag{114}$$

It is straightforward to find the maximum likelihood estimate, $\hat{q}$, of the parameter q. Substituting the expressions for the ϕ_i in terms of q into 114, differentiating with respect to q, and setting the derivative equal to zero gives the following equation for $\hat{q}$ for the Greenwood model

$$\frac{3n_1 + 4n_2 + 2n_3}{\hat{q}} - \frac{n_2 + 2n_3 + 3n_4}{1 - \hat{q}} + \frac{4n_3\hat{q}}{1 + 2\hat{q}^2} + \frac{3n_4(1 + 2\hat{q} + 6\hat{q}^2)}{1 + 3\hat{q} + 3\hat{q}^2 + 6\hat{q}^3} = 0. \tag{115}$$

For the Reed-Frost model we have the following

$$\frac{3n_1 + 4n_2 + 2n_3}{\hat{q}} - \frac{n_2 + 2n_3 + 3n_4}{1 - \hat{q}} + \frac{2n_3\hat{q}}{1 + 2\hat{q}} + \frac{3n_4(1 + 4\hat{q} + 6\hat{q}^2)}{1 + 3\hat{q} + 6\hat{q}^2 + 6\hat{q}^3} = 0. \tag{116}$$

Rearranging either of these expressions gives a polynomial for $\hat{q}$, but the high degree of these polynomials means that we cannot get a solution in closed form. It is, however, fairly straightforward to obtain a numerical solution. For the Providence data, the estimates for the values of q in the two models are found to equal $\hat{q}_{\text{G}} = 0.291$ and $\hat{q}_{\text{RF}} = 0.347$.

Outbreak size	Number of families	Greenwood fitted	Reed-Frost fitted
1	4	2.5	4.2
2	3	1.5	2.8
3	9	14.9	9.0
4	84	81.1	84.0

Table 6: Observed outbreak sizes in 100 households of size 4 during a measles outbreak in Providence, RI. (Wilson *et al.*, 1939), and outbreak sizes predicted by the Greenwood and Reed-Frost models. Fitted values of q for the two models: $\hat{q}_{\text{G}} = 0.291$, $\hat{q}_{\text{RF}} = 0.347$.

The goodness of fit of the two models can be assessed using the χ^2 test. Because of the small expected numbers of outbreaks of either size 1 or 2, we pool these two observations and use a

χ^2 test on one degree of freedom. For the Greenwood model, we get $\chi^2 = 4.69$, which gives $0.025 < P < 0.05$: we would reject this model at the 5% level. For the Reed-Frost model we get $\chi^2 = 0$ (keep in mind that two of the data points have been pooled for this test). The Reed-Frost model gives a better fit than the Greenwood model.

The Providence data gives more detailed information than just outbreak size: it also gives information on the infection chains themselves. For the case of households of size 4, there are eight possible infection chains (as listed in table 3). If the number of instances in which these chains are observed are labeled n_1 through n_8, it is straightforward to show that the log-likelihood function for the Greenwood model is given by

$$\begin{aligned} \log L = \text{const} &+ (3n_1 + 4n_2 + 4n_3 + 2n_4 + 3n_5 + 2n_6 + n_7) \log q \\ &+ (n_2 + 2n_3 + 2n_4 + 3n_5 + 3n_6 + 3n_7 + 3n_8) \log(1-q). \end{aligned} \tag{117}$$

This is a much simpler function of the parameter q than the corresponding expression in the outbreak size setting. In fact, it is easy to solve for the maximum of $\log L$ in closed form

$$\hat{q} = \frac{3n_1 + 4n_2 + 4n_3 + 2n_4 + 3n_5 + 2n_6 + n_7}{3n_1 + 4n_2 + 6n_3 + 4n_4 + 6n_5 + 5n_6 + 4n_7 + 3n_8}. \tag{118}$$

For the Reed-Frost model, the log-likelihood is a linear function of $\log q$, $\log(1-q)$ and $\log(1+q)$.

Infection chain	Number of instances	Greenwood fitted	Reed-Frost fitted
1	$n_1 = 4$	0.9	1.2
$1 \to 1$	$n_2 = 3$	0.4	0.7
$1 \to 1 \to 1$	$n_3 = 1$	0.7	1.0
$1 \to 2$	$n_4 = 8$	8.2	2.2
$1 \to 1 \to 1 \to 1$	$n_5 = 4$	2.7	3.4
$1 \to 1 \to 2$	$n_6 = 3$	6.5	7.3
$1 \to 2 \to 1$	$n_7 = 10$	31.0	38.7
$1 \to 3$	$n_8 = 67$	49.6	45.5

Table 7: Observed chains of infection in 100 households of size 4 during a measles outbreak in Providence, RI. (Wilson *et al.*, 1939.) Fitted values of q for the two models: $\hat{q}_{\mathrm{G}} = 0.209$, $\hat{q}_{\mathrm{RF}} = 0.231$.

The fitted values of q are somewhat different when this more detailed chain data is used. We get $\hat{q}_{\mathrm{G}} = 0.209$ and $\hat{q}_{\mathrm{RF}} = 0.231$, compared to the previous values of 0.291 and 0.347. In order to do the χ^2 goodness of fit tests, several chains must be pooled together, leaving 3 degrees of freedom. This gives χ^2 values of 26.2 and 57.4, respectively, for the two models. These represent very poor fits indeeed.

This suggests that, even though the models give good fits to the outbreak size data, they are not capturing the details of the infection chains at all well.

There could be many reasons why the chain binomial models fail to capture the details of the transmission process. One possibility is that there may be heterogeneity between individuals or families. This could be accounted for by allowing the parameter q to vary across families in some way, albeit at the cost of making the model more complex. It might be that the description of the timecourse of infection employed by the chain binomials is deficient in an important way. Another possibility is that there may have been some instances in which there was more than one introduction of infection into a family. Many of these alternatives have been examined in the large literature on chain binomial models. The solid statistical footing on which these chain binomial models sit means that statistical tests can be used to assess whether the data support the use of more or less complex models in describing different data sets.

12.4 Other Uses of Related Models

Data on outbreak sizes in different households has also been used to estimate the duration of latency and infectiousness. A classic example is provided by Hope Simpson's data on measles outbreaks in a small British town, Cirencester.

The data of interest concern families in which there were two susceptible children, one or both of whom acquired infection during an outbreak. Of the 264 such households studied, there were 45 in which only one child had the infection (so there was no transmission between the children) and 219 in which both children had the infection. In these latter cases, either there was a transmission between the two children or they both acquired infection from outside the family.

Importantly, the data also gives information on the time interval between the two cases. It was found that the distribution of this time interval appeared to be bimodal. In some instances, the time between cases was short: this was interpreted as representing instances in which the two infection events must both have occurred outside the family. These were referred to as 'type A' households. Furthermore, it was assumed that these two infection events occurred simultaneously. In other instances, the time between cases was longer: the interpretation here is that the second case resulted from a within-family transmission event. These were referred to as 'type B' households.

A simple model that could be fitted to this data assumes that there is a latent period, with variance v, followed by a short infectious period and immediately followed by symptoms. In this case, the variance of the distribution of times between cases in type B households reflects the variance v because the time between successive cases will simply equal the latent period of the second individual. The times between cases in the distribution of type A instances, on the other hand, equal the difference between the latent periods of two individuals that are infected at the same timepoint. The variance of this distribution, according to this model, should equal $2v$.

Hope Simpson's data argue quite strongly against this model: the variance amongst type A households is much smaller than the variance amongst type B households. One weakness of this argument is that it is far from clear that all instances of within-family transmission are correctly identified: some of the households identified as type B might, in fact, represent instances in which both children acquired infection from outside the family. The model also assumed that both children in type A instances acquired infection simultaneously. Notice that if this were not the case then the predicted variance of the type A distribution would actually increase: this effect would appear not to help explain the poor fit of this model.

A more detailed model was developed in which it was assumed that the latent period is variable and is then followed by an infectious period of constant length. This infectious period is not necessarily asssumed to be short. In type B households, it is assumed that secondary infection could occur at any point over this infectious period. Consequently, the variance of the B distribution increases. The analysis of this model is quite involved, but it appears to give a good fit to the data. Based on the data, Bailey estimated that measles has a latent period of about 8 days followed by an infectious period of 7 days.

Number of days between cases	Total Number of Households	Number of Type A	Number of Type B
0	5	5	0
1	13	13	0
2	5	5	0
3	4	4	0
4	3	2	1
5	2	0	2
6	4	0	3
7	11	0	11
8	5	0	5
9	25	0	25
10	37	0	37
11	38	0	38
12	26	0	26
13	12	0	12
14	15	0	15
15	6	0	6
16	3	0	3
17	1	0	1
18	3	0	3
19	0	0	0
20	0	0	0
21	1	0	1
Total	219	29	190

Table 8: Hope Simpson's Data for the Time Intervals Between Measles Cases in Households of Size 2 in Cirencester, UK

13 General Descriptions of the Timecourse of Infection

The SIR models that we examined all employed a simple description of the timecourse of infection: individuals become infectious immediately upon infection and then remain infectious for some time. Over the population the distributions of infectious periods are exponentially distributed with average $\tau = 1/\gamma$. While they are infectious, it is assumed that infectivity is constant. We now look at ways in which we can relax both assumptions to allow for a more general description of the timecourse of infection. Notice that the SEIR model was a first step in this direction, allowing for a latent period between acquisition of infection and the start of infectiousness.

13.1 Describing Infectious Period (IP) Distributions (IPD)

Imagine following a collection of individuals who were infected at the same instant, and denote the fraction of these individuals who remain infectious at time t by $x(t)$. We write the recovery rate as $g(t)$, where this rate may vary over time. The recovery process is described by the following equation

$$\frac{dx}{dt} = -g(t)x(t), \tag{119}$$

where we have $x(0) = 1$ since $t = 0$ is the time at which infection occurred.

This model can also be used to describe other processes in which there is a time-dependent removal of individuals. As such, it is utilized in survival analysis, in which $g(t)$ is known as the hazard function. A specific example, is provided by human survivorship, in which case $g(t)$ is the age specific death rate.

There are several ways to solve this equation. Since it is linear in x, it can be solved using a so-called integrating factor. Alternatively, we see that the equation is separable and so can be solved by dividing through by $x(t)$ and then integrating over time:

$$\begin{aligned} \int \frac{1}{x(t)}\frac{dx}{dt}dt &= -\int g(t)dt \\ \Rightarrow \ln x(t) - \ln x(0) &= -\int_0^t g(t')dt'. \end{aligned}$$

Since $x(0) = 1$, we have that

$$x(t) = \exp\left(-\int_0^t g(t')dt'\right). \tag{120}$$

If we denote the duration of the infectious period by the random variable Y, we see that the function $x(t)$ gives the probability that Y is greater than or equal to t. In the language of survival analysis, this probability is known as the survivorship function, $F^s(t)$. This function is one minus the cumulative density function of Y, $F(t) = 1 - F^s(t)$. Differentiating $F(t)$ gives the probability density function, $f(t)$, for Y, with $f(t)dt$ giving the probability that an individual's infectious period lies in the interval $(t, t + dt)$.

The above discussion provides three ways of describing an infectious period distribution: the survivorship function, the hazard function and the probability density function. These descriptions are linked via the following formulae

$$\begin{aligned} \frac{d}{dt}F^s(t) &= -f(t) && (121) \\ \frac{d}{dt}F^s(t) &= -g(t)F^s(t) && (122) \\ f(t) &= -g(t)F^s(t). && (123) \end{aligned}$$

The last of these explicitly makes the point that the hazard function is the recovery rate conditional on having remained infectious until the time of interest.

13.2 Two Examples of Infectious Period Distributions

For simplicity, we ignore background mortality for now. We shall discuss the changes that this introduces later on.

- If one assumes that the hazard function is constant, $\lambda_f(y) = 1/\tau$, then one obtains the exponential distribution, $f(y) = (1/\tau)\exp(-t/\tau)$. Here, τ is the average duration of infectiousness. The assumption that the chance of recovery in a given time interval is independent of the time since infection is unlikely to be a realistic description of the recovery process. The exponential distribution has a fairly large variance: in reality, infectious period distributions are more concentrated around their mean, with few individuals being infectious for very short or very long times.

  The reason that we employ the exponential distribution is one of mathematical convenience. Since the chance of recovery in a given time interval does not depend on the time for which an individual has been infectious, we do not need to keep track of the times for which indivduals have been infectious. This translates into the $-\gamma I$ term that appears in the SIR model, which only depends on the **current** number of infectives.

- Alternatively, we might assume a **fixed** duration of infectiousness: individuals recover exactly τ time units after becoming infected. In other words, the rate at which individuals recover now is exactly equal to the rate at which individuals became infected τ time units ago. If the infection term at time t is of the form $\beta S(t)I(t)/N$, we get the recovery term as $-\beta S(t-\tau)I(t-\tau)/N$.

The model is now described as a **delay-differential equation**. Notice that the delay term means that we need to keep track of the S and I values at previous times: this is the price we pay for not having an exponential distribution. This makes numerical simulation (and analysis) more difficult.

13.3 A General Description of the Timecourse of Infection (I)

Taking our lead from the delay-differential formulation that described the fixed duration of infectiousness, we can describe the general recovery process using a recovery term of the form

$$-\int_0^\infty \frac{\beta}{N} S(t-t')I(t-t')f(t')dt'. \tag{124}$$

This can be understood by noting that the function $f(t')$ describes the probability of someone who was infected t' time units ago recovering now and the term $\beta S(t-t')I(t-t')/N$ describes the rate at which individuals became infected t' time units ago. The integral simply adds up all such contributions over all possible values of t'. (As a remark, such an integral is known as a **convolution**.)

Notice that this formulation assumes constant infectivity over the infectious period: the parameter β does not depend on the time since infection. Again, we need to keep track of the **history** of the system. This **integro differential equation** is again difficult to work with.

A closely related approach does allow for the consideration of variable infectiousness and general descriptions of recovery. The **infectivity kernel**, $A(t')$, describes the average infectivity of individuals at time t' after infection. Notice that recovery of all individuals at some time t^* would correspond to $A(t)$ being equal to zero for all $t > t^*$. Thinking about the SIR model in a closed population, we could write

$$\dot{S} = S(t)(c/N)\int_0^\infty A(t')\dot{S}(t-t')\,dt'. \tag{125}$$

To understand this, we note that in the SIR model, $\dot{S}$, the rate at which indiduals leave the S class, is just the infection rate. The rate at which new infections occur is the proportional to the number of susceptibles and the infectivity of the population. This second term is given by the integral, which adds up the product of the rate at which individuals became infected t' time units ago with the infectivity of such individuals. The constant of proportionality, c/N, describes the mixing process (e.g. depicting the average number of contacts made between individuals in a given time interval). If we were to write $cA(t') = \beta e^{-\gamma t'}$, we would recover the standard SIR model with constant infectivity and exponentially distributed infectious period.

13.4 A General Description of the Timecourse of Infection (II)

An alternative approach keeps track of the **distribution** of times since infection. We write $I(t;t')$ to denote this distribution, which depends on the time at which we are thinking about (t) and the time since infection (t'). Notice that we obtain the total number of infectives at time t by summing $I(t;t')$ over t': $I(t) = \int_0^\infty I(t;t')\,dt'$.

The following picture is helpful here:

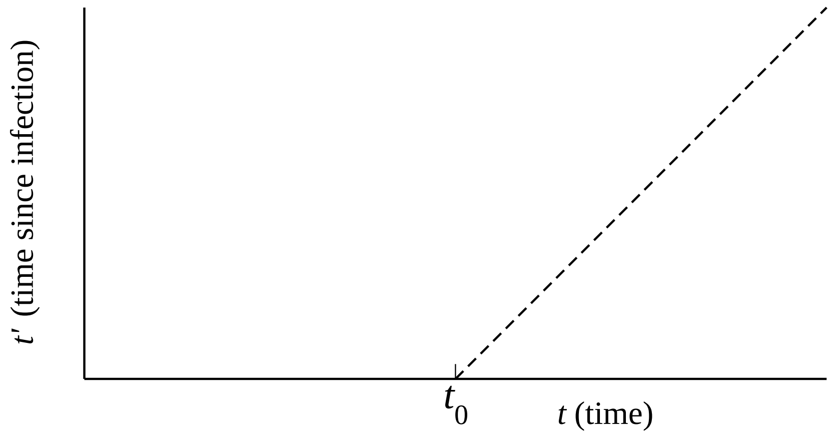

Figure 45: Graph showing t, t' and their relationship when following a cohort of individuals infected at time t_0.

The dashed line shows how we follow a group (or cohort) of individuals that were infected at time $t = t_0$: notice that for these individuals, $t' = t - t_0$. Following this cohort, we see that their number decreases according to the hazard function, $\lambda_f(t')$ via the following **partial differential equation** (PDE)

$$\frac{\partial}{\partial t} I(t;t') + \frac{\partial}{\partial t'} I(t;t') \quad = \quad -\lambda_f(t') I(t;t'). \tag{126}$$

The left hand side of the equation describes what is known as the **total derivative** of $I(t;t')$ with respect to time, and expresses the fact that as we follow any cohort (e.g. along the dashed line in the figure) both t and t' are increasing. The right hand side of the equation expresses the fact that the infectives recover, according to the hazard function.

This PDE formulation is completed by the boundary condition

$$I(t;0) = \frac{S(t)}{N}\int_0^\infty \beta(t')I(t;t')\,dt'. \tag{127}$$

The left hand side describes the newly infected individuals at any given point in time, i.e. the infection term. This is proportional to the number of susceptibles and the total number of infectives, weighted by their infectivity time t' after becoming infected (notice that β depends on t'). If infectivity is constant over the infectious period, this reduces to

$$\begin{aligned} I(t;0) &= \frac{\beta S(t)}{N}\int_0^\infty I(t;t')\,dt' && (128)\\ &= \frac{\beta S(t)I(t)}{N}. && (129)\end{aligned}$$

If we know the values of $I(t;0)$, we can integrate equation (127) along lines parallel to the dashed lines in the above figure. These lines are known as the **characteristics** of the system. The difficulty arises in the calculation of $I(t;0)$, since this involves the nonlinear interaction with the number of susceptibles. We can, as usual, make progress when considering the initial stages of an epidemic, when I is small and S is approximately equal to N. We shall defer this discussion until later.

13.5 A More Intuitive Approach: The Method of Stages

The problem with the exponential distribution is that too many individuals recover too soon or too late: the distribution has too large a variance. One way to build a less dispersed distribution splits the single I class into n **stages**, each of which is passed through in turn. We assume that the time spent in each stage is exponential, and, in the simplest case, we further assume that the average time spent in each class is the same. We remark that the stages are a mathematical device: they need not correspond to anything biological, although they might do.

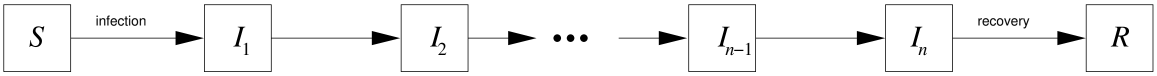

Figure 46: Arrangement of infective stages in the closed SIR model.

Keeping the average duration of infectiousness equal to $\tau = 1/\gamma$, we see that the average time spent in each stage must equal $1/(n\gamma)$. We can describe the stages by the following set of differential equations

$$\dot{I}_1 = \ldots - n\gamma I_1 \tag{130}$$

$$\dot{I}_2 = n\gamma I_1 - n\gamma I_2 \tag{131}$$

$$\vdots \tag{132}$$

$$\dot{I}_n = n\gamma I_{n-1} - n\gamma I_n. \tag{133}$$

We have that $I(t) = I_1 + I_2 + \ldots + I_n$, and so the infection term is of the form $(\beta S/N)(I_1 + I_2 + \ldots + I_n)$.

The infectious period is now the sum of n independent exponential distributions. This is known as a gamma distribution, and its p.d.f. is

$$f(t') = \frac{(\gamma n)^n}{\Gamma(n)} (t')^{n-1} e^{-\gamma n t'}. \tag{134}$$

Here, $\Gamma(n)$ is the gamma function, which equals $(n-1)!$ when n is an integer.

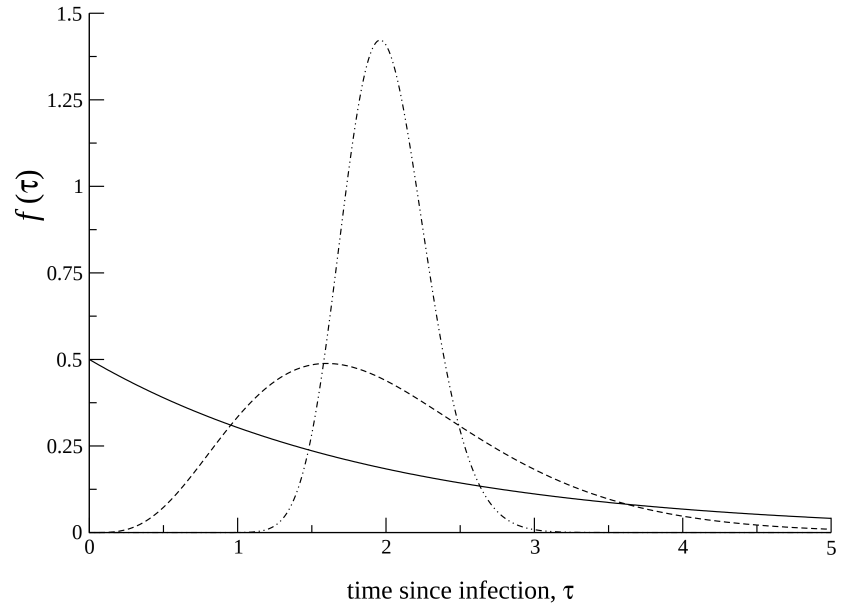

Figure 47: Probability density functions for the gamma distribution with $n = 1$ (exponential), $n = 5$ and $n = 50$ infectious stages. In each case, the mean infectious period is taken to equal two time units.

This distribution has variance $1/(n\gamma^2)$ and is much more concentrated around its mean than the exponential. As n gets larger, we approach the fixed duration of infectiousness case. Notice that there is an intuitive explanation as to why this distribution is more concentrated: if an individual was to recover much before the mean, it would need to move quickly through **each** of the n stages: this becomes less and less likely as n increases.

This stage approach is much more in keeping with our compartmental framework, but comes at the inconvenience of keeping track of n classes. This is still much easier than the infinite dimensional cases of the previous approaches.

The stage approach can describe a more general class of infection timecourses than the gamma distribution with constant infectivity. Variable infectivity can be included by having a different infectivity for each stage, making the infection term of the form

$$\frac{S}{N}\left(\beta_1 I_1 + \beta_2 I_2 + \ldots + \beta_n I_n\right). \tag{135}$$

The SEIR model can be seen in this light, with the I_1 to I_2 transition occurring at rate σ and β_1 equalling zero. More general distributions of infectious periods can be considered by allowing the transition rates between stages to differ. Stages need not be arranged as a simple chain: for instance parallel stages can describe situations in which distributions are more dispersed than the exponential (such as an infection for which some individuals have rapid recover and some individuals have slow recovery).

Detailed descriptions of infectiousness and recovery are important when we want to model infectious with long latent and infectious periods, or if there are marked changes in infectiousness over time. Such descriptions are often needed if we want to do **detailed** modeling, for instance if we want to make predictions of the timecourse of an epidemic. HIV is a particular example, and its study has motivated a lot of the development of the above theory. Notice that in HIV there's a related issue of importance: describing the survivorship function (time to death) for HIV infected individuals. This has also been an active research area, particularly statistical studies that attempt to find biological correlates of survival time (e.g. CD4+ cell count or virus load).

13.6 Consideration of Mortality

If we (as usual) assume a constant background mortality rate, the above formulations must be slightly altered. For instance, the integro-differential description of recovery (124) becomes

$$-\int_0^\infty \frac{\beta}{N} S(t-t')I(t-t')f(t')e^{-\mu t'}\,dt'. \tag{136}$$

The exponential factor discounts the past value of the infection term, reflecting the probability that an individual who was infected t' time units ago would or would not have died in that time.

In a similar way, the PDE that appears in the PDE formulation becomes

$$\frac{\partial}{\partial t}I(t;t') + \frac{\partial}{\partial t'}I(t;t') \;\;=\;\; -\lambda_f(t')I(t;t') - \mu I(t;t'). \tag{137}$$

As we follow our cohort of individuals, they not only recover but also die, as expressed by the second term on the right hand side.

The stage description of infectiousness becomes

$$\begin{aligned}
\dot{I}_1 &= \ldots - (n\gamma + \mu)I_1 && (138)\\
\dot{I}_2 &= n\gamma I_1 - (n\gamma + \mu)I_2 && (139)\\
\vdots & && (140)\\
\dot{I}_n &= n\gamma I_{n-1} - (n\gamma + \mu)I_n, && (141)
\end{aligned}$$

as each stage experiences the background rate of mortality, μ.

13.7 Analysis of Models

We can do the sorts of analyses that we have become familiar with: obtaining expressions for R_0, the initial growth rate of an epidemic, equilibrium expressions and a description of behavior near the equilibria.

13.7.1 The Basic Reproductive Number

R_0 is the average number of secondary infections. With the general description of the IPD, but constant infectiousness, we have (again neglecting mortality)

$$\begin{aligned}
R_0 &= \beta \int_0^\infty t' f(t')\, dt' && (142)\\
&= \beta \int_0^\infty F^s(t')\, dt'. && (143)
\end{aligned}$$

We see that we can write expressions in terms of the pdf or survivorship function. The latter expression is obtained using a standard result for a non negative random variable, namely that its average value is the integral of one minus its cumulative density function. Since the cdf is just one minus the survivorship function, the result follows.

In the variable infectivity setting, we have

$$R_0 = c \int_0^\infty A(t')\, dt'. \qquad (144)$$

13.7.2 Initial Behavior

We can work with any of the formulations, but let's choose the variable infectivity model (125). As ever, we set $S = N$, giving

$$\dot{S} = c\int_0^\infty A(t')\dot{S}(t-t')\,dt'. \tag{145}$$

As mentioned above, we are thinking here about the SIR model in a closed population. If we were to write an equation for the I class, we wouldn't have to worry about describing individuals leaving the I class: this is already accounted for by $A(t')$, the average infectivity of individuals who were infected t' units ago. Therefore, the only transition that the model need describe is from S to I, and we have that $\dot{S} = -\dot{I}$. So we can rewrite our linearized model as

$$\dot{I} = c\int_0^\infty A(t')\dot{I}(t-t')\,dt', \tag{146}$$

or, writing $X(t)$ for dI/dt

$$X(t) = c\int_0^\infty A(t')X(t-t')\,dt'. \tag{147}$$

This equation is well known in mathematical demography, and has been studied by many people, including Lotka in the early part of the 20th Century. If one tries a solution of the form $X(t) = X_0e^{rt}$, one obtains the following equation satisfied by the growth rate r

$$1 = c\int_0^\infty A(t')e^{-rt'}\,dt'. \tag{148}$$

This has been well-studied, and the main results are that this equation has a unique real root (there may also be complex roots) and that the growth rate, r, being positive or negative depends on the basic reproductive number being greater than or less than one, respectively.

The growth rate equation can be solved explicitly in the exponential case. Unsurprisingly, one obtains the same relationship that we got some time ago. Inserting the gamma distribution gives

$$r = \beta\left(1 - \left\{\frac{r\tau}{n} + 1\right\}^{-n}\right). \tag{149}$$

If one multiplies both sides by τ, the resulting equation relates r, R_0, τ and n and so can again be used to estimate R_0, provided that we know n:

$$R_0 = \frac{r\tau}{1 - \{\frac{r\tau}{n} + 1\}^{-n}}. \tag{150}$$

Notice that this equation cannot, in general, be rearranged to give an explicit expression for r in terms of R_0.

As remarked earlier, letting $n \to \infty$ corresponds to a fixed duration of infectiousness, and one can use the result that $(1 + x/n)^n \to e^x$ to obtain

$$r = \beta \left(1 - e^{-r\tau}\right). \tag{151}$$

This result can be obtained directly by linearizing the delay differential equation:

$$\dot{I} = \frac{\beta S I}{N} - \frac{\beta S(t-\tau) I(t-\tau)}{N}. \tag{152}$$

Setting $S = N$ gives

$$\dot{I} = \beta I - \beta I(t-\tau), \tag{153}$$

and then substituting $I = I_0 e^{rt}$, and simplifying, gives the above expression. Notice the appearance of the $e^{-r\tau}$ term: the equation that relates r to R_0 is transcendental.

In all of this section, we ignored mortality. It is, however, straightforward to carry out the corresponding analyses in the slightly more complex settings in which mortality is accounted for.

13.7.3 Equilibrium Behavior

For the SIR model with demography, but generally distributed infectious period, it is straightforward to show that the endemic equilibrium is given by $S^* = N/R_0$ and $I^* = \mu(R_0 - 1)/\beta$. Here R_0 is as given above in equation (142), although mortality must be accounted for.

The approach to the equilibrium can be described by linearising the model, substituting $S(t) = S^* + s(t)$ and $I = I^* + k(t)$ and giving the small perturbations time dependence $\exp(\Lambda t)$. This leads to an equation which determines the values of Λ. The value of Λ which describes the approach to the equilibrium is obtained as the dominant root (i.e. the one with largest real part) of

$$\Lambda + \mu + \beta I^* - \beta S^* \left\{1 - G(\Lambda + \mu)\right\} = 0, \tag{154}$$

where

$$G(\Lambda + \mu) = \int_0^\infty f(\tau) e^{-(\Lambda+\mu)\tau} d\tau. \tag{155}$$

In the case of the gamma distribution, it is easy to see that

$$G(\Lambda + \mu) = \left\{1 + (\Lambda + \mu)/(n\gamma)\right\}^{-n}. \tag{156}$$

Special cases are the exponential case ($n = 1$), for which $G(\Lambda+\mu) = \gamma/(\Lambda+\mu+\gamma)$, and the delta case ($n \to \infty$), for which $G(\Lambda + \mu) = \exp\left\{-(\Lambda + \mu)/\gamma\right\}$. In the former case, equation (154) determining stability reduces to the familiar expression which describes the stability of the endemic equilibrium of the standard SIR model. In the delta case, the expression determining stability reduces to

$$(\Lambda+\mu)+\beta I^{*}-\beta S^{*}e^{-(\Lambda+\mu)/\gamma}=0. \tag{157}$$

A stability result of Hethcote & Tudor (1980) shows that Λ has negative real part, and numerically it is found that, at least for realistic parameter values, Λ is complex. The equilibrium, therefore, is approached via damped oscillations. Figure (48) shows the damping time and ratio of the damping time to the period of the oscillations. Both properties of the oscillations, in marked contrast to R_0, do vary considerably as the parameter n of the gamma distribution is varied. In particular, the damping time increases as n increases; the endemic equilibrium is less stable for less dispersed (i.e. more realistic) distributions of the infectious period.

More realistic distributions destabilize the endemic equilibrium, although not enough to lead to more complex dynamical behaviour.

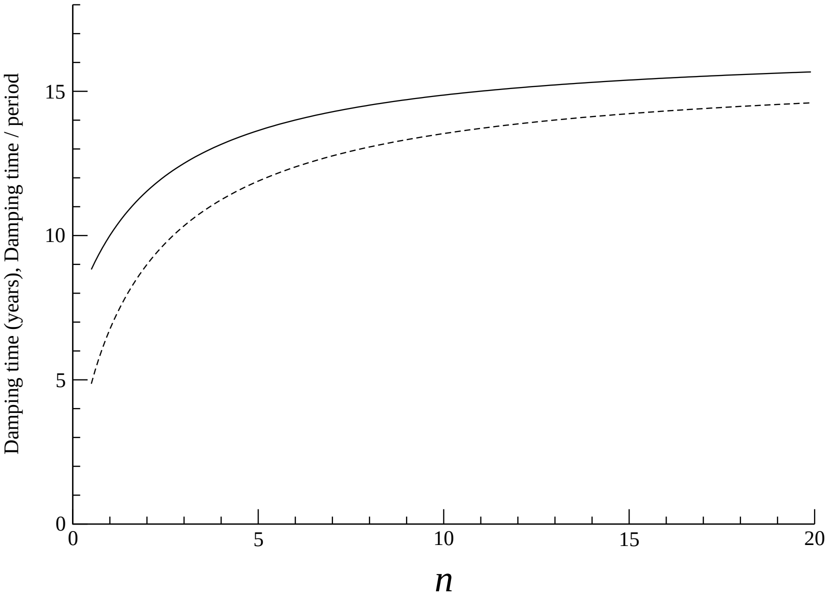

Figure 48: Damping time (solid curve) and ratio of damping time to period (dashed curve) for the damped oscillations in the SIR model with gamma distribution of infectious periods as functions of the parameter n. In each case, the average duration of infection equal to $1/\gamma$. Other parameter values in the model are $N = 10^7$ individuals, $\beta = 1000/N$ individual^{-1}year^{-1}, $\gamma = 100$ year^{-1} and $\mu = 1/50$ year^{-1}.

14 Multigroup Models

Up to this point, we have always assumed that transmission of infection follows the mass-action law. This means that the population is well-mixed and that the probability of transmission is equal between any infective and any susceptible.

In reality, populations are rarely well-mixed: contacts are not made at random in the population. Individuals spend more time in close contact with certain people, such as family members, work colleagues, classmates or people who live in their neighborhood. A striking example of non-random transmission is provided by a sexually transmitted infection in a heterosexual population. Here, transmission is only possible from males to females or from females to males.

People may differ in their susceptibility to infection, for instance one's genetic makeup can give rise to a heightened or lessened susceptibility. Similarly, some individuals may be more or less infectious once they acquire infection, or have longer or shorter average durations of infection.

The important point is that populations are often heterogeneous with respect to transmission. The **multigroup model** provides one way to depict heterogeneity. The population is subdivided into a number of subgroups and the model accounts for the numbers of susceptibles and infectives, S_j and I_j, in each group. If the number of groups is finite, then we will usually denote the number of groups by n. Transmission between groups is often depicted by a matrix of transmission parameters, β_{ij}, describing transmission from infectives in group j to susceptibles in group i. This matrix is often termed the 'who acquires infection from whom' (WAIFW) matrix.

For a simple SIR model with demography, a multigroup model could have the following form

$$\dot{S}_i = \mu N_i - S_i \sum_{j=1}^{n} \beta_{i\,j} I_j - \mu S_i \tag{158}$$

$$\dot{I}_i = S_i \sum_{j=1}^{n} \beta_{i\,j} I_j - (\gamma_i + \mu) I_i. \tag{159}$$

Here, N_i gives the number of individuals in group i. Notice that, in this formulation, we have incorporated the $1/N$-like quantities that arise from standard incidence terms into the $\beta_{i\,j}$.

14.1 The Basic Reproductive Number for Multigroup Models

Can we extend the definition of the basic reproductive number to cover the multigroup setting? It might seem that the verbal definition of the average number of secondary infections would be straightforward to apply. Indeed, some authors have tried to develop expressions for R_0 by taking arithmetic means. It turns out, however, that this naïve approach does **not** give the correct answer in general. The multigroup setting requires us to think more carefully about the appropriate notion of ‘average’.

We shall first consider the early stages of an epidemic, following some introduction of an infection. As usual, we assume that the population is entirely susceptible and imagine that any reduction in the number of susceptibles has negligible impact during the initial stages of an outbreak. During this initial period, the epidemic can be described by a linear model.

The **next generation** matrix K describes the average number of secondary infections in the ith group that are caused by a single infective individual in the jth group over the course of their infectious period. We usually denote the number of groups by n, and so K is an $n \times n$ matrix. (Remark: This definition can be extended to cover situations in which the types of individuals are described by a continuously varying quantity.)

The basic reproductive number, R_0, is defined to be the **dominant eigenvalue** (the one with greatest modulus– sometimes known as the maximal eigenvalue) of the matrix K.

If we have a vector ϕ_0 that gives the numbers of infectives in each group at the initial time (these entries could be fractional if we are thinking about an **average** taken over some initial **distribution** of infectives), then the product $K\phi_0 = \phi_1$ gives the numbers of infectives in each group after one generation. Similarly, the numbers of infectives in the mth generation are given by

$$\phi_m = K^m \phi_0. \tag{160}$$

Assume that the matrix K has a complete set of eigenvectors, which we denote by $\mathbf{\Psi}_j$, for $j = 1 \ldots n$. We can then write ϕ_0 as a linear combination of the eigenvectors

$$\phi_0 = \sum_{j=1}^{n} c_j \mathbf{\Psi}_j. \tag{161}$$

The numbers of infectives in the nth generation can then be written as

$$\phi_m = \sum_{j=1}^{n} c_j \lambda_j^m \mathbf{\Psi}_j. \tag{162}$$

Here, the λ_j are the eigenvalues that correspond to the eigenvectors $\mathbf{\Psi}_j$.

The simplest justification for the R_0 definition is that if K has a (strictly) maximal eigenvalue, λ_1, then for large m the sum will be dominated by the term with $j = 1$, and we have that

$$\phi_m \sim c_1 \lambda_1^m \mathbf{\Psi}_1. \tag{163}$$

We see that, with this linear description, the numbers of infectives increase by the factor λ_1 each generation: this is precisely what we mean by R_0. Also notice that the distribution of infectives amongst the groups, when n is large, is given by the entries of the eigenvector $\mathbf{\Psi}_1$.

Remark: Notice that this discussion has been in terms of **generations**, not **time**. The quantity R_0 describes how the epidemic grows in generational terms, while the initial growth rate, r, describes how the epidemic grows in time. The two quantities are related, though.

14.2 Calculating R_0 in Multigroup Settings

We have seen that the calculation of R_0 requires us to find the dominant eigenvalue of the next generation matrix. In some cases we might be able to calculate this eigenvalue explicitly. In general, though, this will not be possible. It would be nice to know some general theory about the dominant eigenvalues of matrices so that we gain some information in the general case.

Since the entries of this matrix are just numbers of secondary infections, all of its entires must be non-negative: we call this a non-negative matrix. It turns out that there is a large body of theory that looks at the eigenvalues and eigenvectors of non-negative matrices. Many of these results are associated with Frobenius and Perron. This theory is very useful in many areas of applied mathematics.

14.2.1 Types of Non-Negative Matrices and Their Properties

Definition: A non-negative square matrix A is **irreducible** if for each pair (i, j) there exists an integer $m > 0$ such that $a_{i\,j}^{(m)} > 0$.

In terms of our present context, K being irreducible means that it is possible for infection to pass from any group j to any group i, possibly via some chain of intermediate transmissions. If the matrix K was not irreducible, we would be able to decompose the multigroup system into two or more non-interacting subsystems.

Definition: A non-negative square matrix A is **primitive** if there exists an $m > 0$ such that $a_{i\,j}^{(m)} > 0$ for each pair (i, j).

If the next generation matrix K is primitive, then it is guaranteed that there will be some infectives in each group after m generations, regardless of how the infection was initially introduced.

Theorem: If A is a non-negative square matrix, then A has a non-negative eigenvalue, λ_1, that is at least as large as the absolute value of any eigenvalue of A, and has a non-negative eigenvector corresponding to λ_1.

Theorem: A non-negative eigenvector of a non-negative irreducible matrix must be strictly positive (i.e. all entries are positive).

Theorem: An irreducible non-negative matrix A has a real positive eigenvalue, λ_1, such that $\lambda_1 \geq |\lambda_i|$ for any eigenvalue λ_i of A. Furthermore, there is a positive eigenvector corresponding to λ_1 (i.e. each entry of the eigenvector is positive). The eigenvalue λ_1 is called a **maximal eigenvalue** of A and the positive eigenvector corresponding to λ_1 is called a **maximal eigenvector** of A.

Theorem: The maximal eigenvalue of an irreducible non-negative matrix is a simple root of its characteristic equation. Furthermore, if there are h eigenvalues with modulus λ_1, these eigenvalues are the distinct roots of λ_1^h.

Remark: This means that there can be other eigenvalues of A with modulus equal to λ_1, but that these lie on a circle of radius λ_1 in the complex plane.

Theorem An irreducible matrix has exactly one non-negative eigenvector.

Theorem If A is a non-negative primitive matrix, then $\lambda_1 > |\lambda_i|$ for all the remaining eigenvalues λ_i, $i > 1$.

Remark: If K is a primitive matrix, then we say that its maximal eigenvalue is strictly maximal.

14.2.2 Applications

Hopefully, the relevance of this theory is fairly clear. If the next generation matrix is primitive then it has a strictly maximal (dominant) eigenvalue λ_1. In this case, it's not too difficult to show that equation (163) holds. The maximal eigenvalue, which is how we define R_0, tells us how the number of infectives increases from generation to generation in the linearized picture.

Notice that if there are other non-zero eigenvalues of K, then the expression for $K^n\phi$ would also involve the nth powers of the remaining eigenvalues (equation (163) is just the leading term of this expression). The point is that the terms corresponding to the subdominant eigenvalues grow more slowly. If we wait long enough, then these terms become unimportant compared to the one involving the dominant eigenvalue. The subdominant eigenvalues describe **transient behavior**. The length of the transient period depends on the relative sizes of the sizes of the dominant and subdominant eigenvalues.

One point to wonder about this transient period is whether it is long compared to the time that the linearization of the model holds. In any case, R_0 is an upper bound on the number of secondary infections.

14.2.3 A Simple Example

Imagine a sexually transmitted disease that spreads by heterosexual contact. The two groups are males and females. If the average number of secondary infections that result in the female population from the introduction of one infective male is R_0^{MF}. The corresponding number of secondary infections in the male population resulting from the introduction of one infective female is R_0^{FM}.

The next generation matrix is

$$K = \begin{pmatrix} 0 & R_0^{\mathrm{FM}} \\ R_0^{\mathrm{MF}} & 0 \end{pmatrix}. \tag{164}$$

The eigenvalues of the next generation matrix are $\pm\sqrt{R_0^{\mathrm{MF}} R_0^{\mathrm{FM}}}$. The dominant eigenvalue has the plus sign. Notice that this is the **geometric mean** of the two between-group R_0 values.

Notice that the second eigenvalue has equal modulus. Why is this? The next generation matrix is irreducible, but not primitive. Because we only have male to female and female to male transmission, if we to start with infection only in one group, the infection would go back and forth between the two groups in alternate generations. If there was some homosexual transmission, then the next generation matrix would be primitive and the second eigenvalue would have smaller modulus.

A naive attempt to calculate R_0 for this system might involve us taking the arithmetic mean of R_0^{MF} and R_0^{FM}. This would not give us the correct answer (unless the two quantities are equal): in general, the geometric mean is smaller than the arithmetic mean.

14.3 Separable Matrices

We have a pretty wide variety of mixing patterns that we could choose. In the absence of more detailed information, we quite often choose mixing patterns so that we can calculate the eigenvalues of the next generation matrix. Since many mathematical problems involve eigenvalue calculations, having an experience for knowing what sorts of matrices are amenable to such analyses is an extremely useful skill.

One general class of next generation (or mixing) matrices whose eigenvalues we can calculate are known as **separable matrices**. These have the form $k_{ij} = a_i b_j$.

Take any vector ϕ and calculate $K\phi$. The ith component of $K\phi$ is equal to

$$\begin{aligned}(K\phi)_i &= \sum_j k_{ij}\phi_j \\ &= \sum_j a_i b_j \phi_j \\ &= a_i \sum_j b_j \phi_j. \end{aligned} \tag{165}$$

We therefore have that

$$K\phi = \left(\sum_{j=1}^{n} b_j \phi_j \right) \mathbf{a}. \tag{166}$$

Whatever vector ϕ we take, the product $K\phi$ is a multiple of the vector $\mathbf{a}$. In particular, the product $K\mathbf{a}$ is a multiple of $\mathbf{a}$: $\mathbf{a}$ is an eigenvector of K with eigenvalue

$$\lambda_1 = \sum_{j=1}^{n} b_j a_j. \tag{167}$$

All other eigenvalues of K must equal zero. We therefore have that

$$R_0 = \sum_{j=1}^{n} b_j a_j. \tag{168}$$

It is straightforward to show that this expression for R_0 can be written as follows

$$R_0 = n\langle a\rangle\langle b\rangle + n\rho\sigma(a)\sigma(b). \tag{169}$$

Here, $\langle a\rangle$ is the average of the a_i, $\sigma(a)$ is their standard deviation and the quantity ρ is the (product-moment) correlation coefficient of the a_i and b_i. These quantities are defined as

$$\langle a\rangle = \frac{1}{n}\sum_{j=1}^{n} a_j, \tag{170}$$

$$\sigma^2(a) = \frac{1}{n}\sum_{j=1}^{n} (a_j - \langle a\rangle)^2, \tag{171}$$

with corresponding expressions for $\langle b\rangle$ and $\sigma(b)$, and

$$\rho^2 = \frac{\left\{\sum_{j=1}^{n} (a_j - \langle a\rangle)(b_j - \langle b\rangle)\right\}^2}{\sigma(a)\sigma(b)}. \tag{172}$$

The derivation of equation (169) makes use of the identity

$$\sum_{j=1}^{n} \left(a_j - \langle a\rangle\right)\left(b_j - \langle b\rangle\right) = \sum_{j=1}^{n} a_j b_j - n\langle a\rangle\langle b\rangle. \tag{173}$$

We see that if there is no correlation between the a_i and b_i, the expression for R_0 simply involves the averages of the a_i and of the b_i. Otherwise, if the correlation is non-zero, the value of R_0 can be inflated (if the correlation is positive) or deflated (if the correlation is negative) by a term that involves the standard deviation of the a_i and the b_i.

Perhaps a more intuitive way of understanding the properties of a separable matrix involves the observation that the columns of K are just multiples of the vector $\mathbf{a}$. (Such matrices are known as rank one matrices.) In fact, the jth column of K is just b_j times $\mathbf{a}$. So the product of K and any vector is just some multiple of $\mathbf{a}$. The only possible non-zero eigenvector of K is $\mathbf{a}$. The remaining eigenvectors must have zero eigenvalues (if this were not the case, then the product of K and the eigenvector would be a multiple of $\mathbf{a}$). In fact, from (166), we see that these eigenvectors are defined by the condition that they are orthogonal to $\mathbf{b}$, so that their scalar product with $\mathbf{b}$, $\mathbf{x}.\mathbf{b}$, is zero. (We might wonder about the case when $\mathbf{a}$ and $\mathbf{b}$ are orthogonal. Since the entries of K must be non-negative, the only way that this can happen is if K is the zero matrix.)

14.3.1 What Does the Separable Assumption Mean?

The entries $k_{i\,j}$ in the next generation matrix are given by the products of the transmission parameters $\beta_{i\,j}$ and the appropriate average durations of infectiousness. Notice that if the average durations of infection are equal for individuals in each group, then finding the eigenvalues of K essentially corresponds to finding the eigenvalues of the beta matrix that describes the mixing between groups, since $k_{i\,j} = \beta_{i\,j}/\gamma$.

In general, the $k_{i\,j}$ depend on the the infectivity of individuals in group j (how infectious they are, and how long they remain infectious for), the susceptibility of individuals in group i, and the rate at which individuals in groups i and j meet each other. In the separable situation, for which $k_{i\,j} = a_i b_j$, the b_j depend on the infectiousness of individuals in group j and the a_i depend on the susceptibility of individuals in group i.

The separable assumption is a strong one to make: it essentially says that the probability of transmission between any pair of individuals depends independently on the types of the two individuals involved.

The separable assumption justifies the use of simple averages in the calculation of R_0 in some settings. For instance, imagine we have a population where there are two types of individual, one of which recovers more quickly than the other. Type I individuals have a recovery rate of γ_1 and type II individuals have a recovery rate of γ_2. We assume that the transmission parameter, β, is the same for both types (both within and between groups), so that the transmission between groups i and j is described by the term $\beta S_i I_j/N$ for any i and j.

Assume that a fraction p of the population is of type I. We can easily work out the next generation matrix. For instance, the term describing transmission from type I to type I is given by $\beta S_1 I_1/N$, which (in the linearized model) equals $\beta(pN)I_1/N$, or $\beta p I_1$. Therefore the corresponding entry of K, $k_{1\,1}$, equals $\beta p/\gamma_1$. Similarly arguments lead to

$$K = \begin{pmatrix} \beta p/\gamma_1 & \beta p/\gamma_2 \\ \beta(1-p)/\gamma_1 & \beta(1-p)/\gamma_2 \end{pmatrix}. \tag{174}$$

We see that this is a separable matrix, with $k_{i\,j} = a_i b_j$, and $a_1 = \beta p$, $a_2 = \beta(1-p)$ and $b_j = 1/\gamma_j$.

Therefore, $R_0 = p\beta/\gamma_1 + (1-p)\beta/\gamma_2$, which can be rewritten as $R_0 = pR_0^{(1)} + (1-p)R_0^{(2)}$. The overall R_0 is just the (weighted) average of the R_0 within the two groups, as might have been expected.

Keep in mind that, in general, R_0 cannot be calculated using a simple averaging procedure (recall the earlier example of the sexually transmitted infection).

14.4 Towards a General Description of Mixing

A commonly used multi-group model involves imagining that different groups have different **activity levels**: individuals in the various groups make different numbers of contacts per unit time.

As in most of the settings discussed in this course, all of the contacts are assumed to involve two individuals. This means that there are two possible ways of counting contacts: either by the number of contacts or the number of individuals involved in those contacts. These numbers differ by a factor of two.

One setting in which these activity level based models has been commonly used involves sexually transmitted infections. Here, the rate at which an individual acquires infection might depend on the rate at which they acquire new sexual partners. (Notice that models for homosexual transmission are simpler than those for heterosexual transmission since one only need keep track of one sex, rather than two!)

A general description of mixing, based on activity levels, is as follows:

- It is assumed that there are N_i individuals in group i. The fraction of individuals that are in group i is given by N_i/N.
- The activity level of group i, defined as the average number of contacts made by an individual from this group per unit time (i.e. the rate at which they make contacts), is denoted by a_i.

- The total rate at which contacts are made by all the individuals in group i is given by a_iN_i.
- The total rate at which contacts are made by all individuals in the population is given by $\sum_i a_iN_i$ and is written as D. (Notice that each contact is counted twice in this quantity.)
- It is useful to define the **fractional activity level** of a group, written as b_i. This is given by dividing the total rate at which all the individuals in the group make contacts by the total rate at which individuals make contacts in the population. So $b_i = a_iN_i/D$. By definition, the b_i sum to one.
- We define the mixing matrix, M, whose entries are written as $m_{i\,j}$. The entries of the mixing matrix give the fractions of those contacts made by an individual in group j that are with individuals in group i.

  Since the entries of this matrix are fractions, the sum of the entries in any column of M must equal one.
- The rate at which each individual in group j makes contacts with individuals in group i is therefore given by $a_jm_{i\,j}$.
- The total rate at which all individuals in group j make contacts with individuals in group i is given by $a_jN_jm_{i\,j}$.

The mixing matrix and activity levels are subject to a set of constraints, because the number of contacts made by individuals in group i with those in group j must equal the number of contacts made by individuals in group j with those in group i. This implies the following consistency conditions

$$a_jN_jm_{i\,j} = a_iN_im_{j\,i}. \tag{175}$$

14.4.1 A Concrete Example

The notational jungle of the above description of mixing is more easily navigated with the aid of a simple example. We consider a situation in which there are two groups, labeled 1 and 2.

Imagine that the first group has 100 individuals, each of whom make 6 contacts per day	$N_1 = 100$ $a_1 = 6$ per day
Imagine that the second group has 200 individuals, each of whom make 2 contacts per day	$N_2 = 200$ $a_2 = 2$ per day

Group one gives rise to 600 contacts per day and group two gives rise to 400 contacts per day.

$$a_1 N_1 = 600 \text{ per day}$$
$$a_2 N_2 = 400 \text{ per day}$$

The total contact rate for individuals, D, is 1000 contacts per day

$$D = \sum a_i N_i$$
$$= 1000 \text{ per day}$$

(Keep in mind that this means that a thousand individuals make contacts per day and so corresponds to there being 500 contacts made per day.)

Of the contacts made, individuals from group one are responsible for 60%, while individuals from group two are responsible for the remaining 40%.

The fractional activity levels are $b_1 = 0.6$ and $b_2 = 0.4$

$$b_i = a_i N_i / D$$
$$b_1 = 600/1000 = 0.6$$
$$b_2 = 400/1000 = 0.4$$

The coefficients m_{ij} specify the mixing within and between groups.

Of the 600 daily contacts made by all the individuals in group one:

$600\, m_{11}$ will be with individuals from group one $\qquad a_1 N_1 m_{11} = 600\, m_{11}$

$600\, m_{21}$ will be with individuals from group two $\qquad a_1 N_1 m_{21} = 600\, m_{21}$

Since all of these contacts must be with people from one of the groups, we must have that $600\, m_{11} + 600\, m_{21} = 600$

$$m_{11} + m_{21} = 1$$

Of the 400 daily contacts made by all the individuals in group two:

$400\, m_{12}$ will be with individuals from group one $\qquad a_2 N_2 m_{12} = 400\, m_{12}$

$400\, m_{22}$ will be with individuals from group two $\qquad a_2 N_2 m_{22} = 400\, m_{22}$

By the argument just given for the contacts of group one, $400\, m_{12} + 400\, m_{22} = 400$

$$m_{12} + m_{22} = 1$$

The between-group contact rates must balance: $400\, m_{12} = 600\, m_{21}$

$$a_2 N_2 m_{12} = a_1 N_1 m_{21}$$

14.4.2 Incorporating Mixing into Epidemic Models

A transmission model requires not only a description of mixing, but also some account of how infection can be transmitted as a consequence of the contacts that are made within the population. Once this is done, we can calculate the next generation matrix (and hence R_0) and write down equations for multi-group infection models.

- If the probability that an interaction between an infective in group j with a susceptible in group i leads to transmission of the infection is written as $p_{i\,j}$, then the transmission parameter between the groups is given by $a_j m_{i\,j} p_{i\,j}$. In many cases, we assume that all of the $p_{i\,j}$ are equal, and write the common value as p.
- We denote the average duration of infection for individuals in the ith group by τ_i (which equals $1/\gamma_i$).
- The entry $k_{i\,j}$ of the next generation matrix gives the average number of secondary infections in group i that result from there being a single infective in group j (when the population is otherwise entirely) susceptible. This is simply given by the product of the rate at which such infections occur and the average duration of infectiousness: $k_{i\,j} = a_j m_{i\,j} p_{i\,j} \tau_j$.
- We sometimes define the **effective contact number of a group**, written as k_j, to be the average number of effective contacts (i.e. those that lead to a secondary infection) made by an individual in the jth group over the course of their infectious period. We see that $k_j = \sum_i k_{i\,j}$.

If we employ an SIR-type description of the infection process, we can write down the following set of equations:

$$\begin{aligned}\dot{S}_i &= \mu N_i - \frac{S_i}{N_i}\sum_{j=1}^{n} a_j m_{i\,j} p_{i\,j} I_j - \mu S_i && (176)\\ \dot{I}_i &= \frac{S_i}{N_i}\sum_{j=1}^{n} a_j m_{i\,j} p_{i\,j} I_j - (\gamma_i + \mu) I_i. && (177)\end{aligned}$$

This can be written more compactly as

$$\begin{aligned}\dot{S}_i &= \mu N_i - S_i \sum_{j=1}^{n} \beta_{i\,j} I_j - \mu S_i && (178)\\ \dot{I}_i &= S_i \sum_{j=1}^{n} \beta_{i\,j} I_j - (\gamma_i + \mu) I_i, && (179)\end{aligned}$$

where $\beta_{i\,j} = a_j m_{i\,j} p_{i\,j} / N_i$. Here, we have incorporated the $1/N_i$ into the $\beta_{i\,j}$.

You might wonder why we have the N_i in the denominators of the transmission terms. Looking at the above description of the mixing process, the $a_j m_{i\,j}$ give the rates at which each individual in group j makes contacts with members of group i. With this description, the rates do not explicitly depend on the target population sizes. In order to achieve this, the transmission term must involve the fractions S_i/N_i, otherwise the transmission rates (and hence the entries of the next generation matrix) would scale with the N_i. Notice that this assumption corresponds to the standard incidence term $\beta SI/N$ that was employed when we looked at simple single population models.

Using the consistency relations (equation 175), we can rewrite the multigroup model as

$$\dot{S}_i \;=\; \mu N_i - S_i \sum_{j=1}^{n} a_i m_{j\,i} p_{i\,j} \frac{I_j}{N_j} - \mu S_i \tag{180}$$

$$\dot{I}_i \;=\; S_i \sum_{j=1}^{n} a_i m_{j\,i} p_{i\,j} \frac{I_j}{N_j} - (\gamma_i + \mu) I_i. \tag{181}$$

14.4.3 Mixing Patterns: Proportionate, Assortative and Disassortive Mixing

The **proportionate mixing model** assumes that individuals have no preference for those with whom they interact. In other words, the chance that a given contact will be with an individual of type i is determined by the relative activity level of individuals of type i. This mixing pattern has $m_{i\,j} = b_i$ and is sometimes known as **random mixing**. (The use of the term random mixing can cause confusion: contacts are not simply chosen at random from the population, but from the population according to the activity distribution.)

If each group has the same activity level, with a_i equal to a for all i, the proportionate mixing scenario has a particularly simple interpretation. D is then equal to $\sum_i a_i N_i = aN$, and so $b_i = a_i N_i/D = aN_i/(aN) = N_i/N$. Hence $m_{i\,j} = N_i/N$: the probability that a particular contact of an individual being with someone from group i is simply equal to the fraction of the population that are in group i.

Assortative mixing implies that groups have a higher within-group contact rate than would be expected under proportionate mixing. In this situation, individuals are more likely to interact with others who are similar to them. **Disassortative mixing** describes the situation in which within-group contact rates are lower than expected under proportionate mixing: individuals have a tendency to interact with others who are dissimilar to them.

A limiting case of assortative mixing is called **restricted mixing**: individuals only interact with their own kind. A commonly used model for assortative mixing is **preferred mixing**: a certain fraction of contacts are assumed to be within-group contacts, while the remaining contacts are subject to proportionate mixing.

14.4.4 A Return to Our Example Mixing Pattern

To complete the description of mixing, we need to specify the m_{ij}

For proportionate mixing, the chance of a given contact being with an individual from group one is 60% and the chance of a given contact being with an individual from group two is 40%. These probabilities follow the fractional activity levels of the groups.

$$\begin{aligned} m_{ij} &= b_i \\ m_{1j} &= b_1 = 0.6 \\ m_{2j} &= b_2 = 0.4 \end{aligned}$$

Notice that the m_{ij} do not depend on j: the probabilities that a given contact of person A is with an individual from group 1 or from group 2 do not depend on person A's group.

In our example, this means that $m_{11} = 0.6$, $m_{12} = 0.6$, $m_{21} = 0.4$ and $m_{22} = 0.4$.

Putting this all together:
Individuals in group 1 make 360 contacts with other individuals in group 1 and 240 contacts with individuals in group 2 per day

$$\begin{aligned} a_1 N_1 m_{11} &= 360 \\ a_1 N_1 m_{21} &= 240 \end{aligned}$$

Individuals in group 2 make 240 contacts with individuals in group 1 and 160 contacts with other individuals in group 2 per day

$$\begin{aligned} a_2 N_2 m_{12} &= 240 \\ a_2 N_2 m_{22} &= 160 \end{aligned}$$

We see that the balance condition holds: individuals in group 1 make the same number of contacts with individuals in group 2 as do individuals in group 2 with individuals in group 1.

Each day, 500 contacts are made: 180 between individuals of group 1, 80 between individuals of group 2 and 240 between individuals in the two groups.

If mixing was assortative, m_{11} and m_{22} would be larger (at the expense of the between-group mixing terms) than under proportionate mixing. In the disassortative setting, m_{11} and m_{22} would be lower.

14.5 Mixing Patterns and the Basic Reproductive Number

We can now calculate the values of R_0 under the different descriptions of mixing. To simplify matters, we assume that all of the per-contact transmission probabilities, $p_{i\,j}$, equal p, and that each group experiences the same average duration of infectiousness, τ. We have that $k_{i\,j} = a_j m_{i\,j} p\tau$.

For proportionate mixing we have that $m_{i\,j} = b_i$ and so $k_{i\,j} = a_j b_i p\tau$. We notice that proportionate mixing leads to the next generation matrix being separable. Consequently, the basic reproductive number is given by

$$R_0 = \sum_i a_i b_i p\tau. \tag{182}$$

Making use of the fact that the fractional activity levels are defined by $b_i = a_i/D$, we can write

$$\begin{aligned} R_0 &= p\tau \sum_i a_i^2 N_i/D \\ &= p\tau \frac{\sum_i a_i^2 N_i}{\sum_i a_i N_i}. \end{aligned} \tag{183}$$

If we divide the numerator and denominator of this fraction by N, we can rewrite in terms of the mean of the activity levels and the mean of the squares of the activity levels.

$$\begin{aligned} R_0 &= p\tau \frac{\sum_i a_i^2 (N_i/N)}{\sum_i a_i (N_i/N)} \\ &= p\tau \frac{\langle A^2 \rangle}{\langle A \rangle}. \end{aligned} \tag{184}$$

Here $\langle A \rangle$ denotes the average activity level, wieghted by the group sizes, $\langle A^2 \rangle$ denotes the (weighted) average of the squares of the activity levels.

The numerator $\langle A^2 \rangle$ can be re-expressed in terms of the (weighted) variance of the activity levels by manipulating the equation $\mathrm{Var}(A) = \left\langle (A - \langle A \rangle)^2 \right\rangle$. Multiplying out the squared term on the right gives the standard result $\mathrm{Var}(A) = \langle A^2 \rangle - \langle A \rangle^2$. Hence we have that

$$\begin{aligned} R_0 &= p\tau \frac{\langle A \rangle^2 + \mathrm{Var}(A)}{\langle A \rangle} \\ &= p\tau \left(\langle A \rangle + \frac{\mathrm{Var}(A)}{\langle A \rangle} \right) \\ &= p\tau \langle A \rangle \left(1 + \mathrm{CV}^2 \right). \end{aligned} \tag{185}$$

Here, CV denotes the (weighted) coefficient of variation (the standard deviation divided by the mean) of the activity levels.

If we had homogeneous mixing (i.e. a well-mixed population), the expression for R_0 would be $p\tau\langle A \rangle$: the product of the average number of contacts made over the duration of the infectious period ($\tau\langle A \rangle$) and the probability that a given contact leads to transmission of infection.

In the proportionate mixing case, there is an additional contribution to R_0 arising from the **variance** of the activity level distribution. Heterogeneity tends to increase R_0. We remark that this is a particular example of the general result (169) obtained for separable next generation matrices. In this case the correlation coefficient in equation (169) is equal to one.

Put another way, the expression for R_0 that would be obtained by assuming perfect mixing (corresponding to taking the arithmetic mean over the activity levels) **underestimates** the true value of R_0. If the variance of the activity level distribution is large compared to its mean, then we see that there can be a considerable difference between the simple naive expression for R_0 and its true value.

This difference can be very important for sexually transmitted diseases. Most individuals have a small number of sexual partners per unit time, but there is a small number of people who have a large number of partners (think of prostitutes or other sex-workers as an extreme example). This leads to the activity level distribution having such a large variance that the mean activity level is a very poor indicator of the true value of R_0. In many cases, the R_0 that would be predicted if one only looked at the mean lies well below one, but the high risk groups pull the overall value up above one.

The important observation is that high risk individuals are responsible for a disproportionate number of transmission events. We can see the source of this effect in equation (183), in which the a_i appear as squared terms. This reflects the fact that high risk individuals (i) have more contacts than the average person, and (ii) that a given contact is more likely to be with a high risk individual, under the proportionate mixing assumption. Put another way, you could think of taking two averages: the average activity level over the population, or the average activity level of people's partners. The latter average will be higher.

14.5.1 Assortative and Disassortative Mixing

The assortative and disassortative forms of mixing do not satisfy the separability condition. In general, it is not possible to get explicit expressions for R_0.

In the preferred mixing case, one can obtain a nonlinear equation that is satisfied by R_0. As an example, take $k_{ij} = A_i B_j + C_j \delta_{ij}$. Here, δ_{ij} is the Kronecker delta, which equals zero if $i \neq j$, but equals one if $i = j$. The proportionate mixing component is described by $A_i B_j$ and the additional within-group contacts by C_j. (Notice this employs a slightly simplified description of preferred mixing: we assume that the extra within-group contacts are made **in addition** to the proportionate mixing term. If we wanted to keep the total number of contacts fixed, we could just normalize by dividing through by an appropriate factor.)

One can show that R_0 satisfies the following equation:

$$\sum_{i=1}^{n} \frac{A_i B_i}{R_0 - C_i} = 1. \tag{186}$$

To see this, we need to find the eigenvalues of the next generation matrix. We have $K\mathbf{\Psi} = \lambda\mathbf{\Psi}$. Writing out the ith component of this, we have that

$$\sum_{j=1}^{n} \left(A_i B_j + C_j \delta_{i\,j}\right) \Psi_j \quad = \quad \lambda \Psi_i, \tag{187}$$

which implies

$$A_i \sum_{j=1}^{n} B_j \Psi_j + C_i \Psi_i \quad = \quad \lambda \Psi_i, \tag{188}$$

or that

$$A_i \sum_{j=1}^{n} B_j \Psi_j \quad = \quad (\lambda - C_i)\Psi_i. \tag{189}$$

The trick now is to multiply both sides by $B_i/(\lambda - C_i)$ and sum over i:

$$\sum_{i=1}^{n} \frac{A_i B_i}{\lambda - C_i} \sum_{j=1}^{n} B_j \Psi_j \quad = \quad \sum_{i=1}^{n} B_i \Psi_i. \tag{190}$$

This means that either $\Sigma B_j \Psi_j$ must equal zero or that

$$\sum_{i=1}^{n} \frac{A_i B_i}{\lambda - C_i} = 1. \tag{191}$$

If we think about the dominant eigenvalue, $\lambda = R_0$, we see that the sum $\Sigma B_j \Psi_j$ cannot equal zero: each of the B_j is non-negative and the Ψ_j are strictly positive (proportionate mixing implies that the next generation matrix is primitive). This gives the equation for R_0 as claimed.

14.6 Equilibrium Analysis

We can solve to find the endemic equilibrium values of S_i and I_i. As usual, we denote equilibrium quantities with an asterisk. We will use the following set of equations:

$$\dot{S}_i \quad = \quad \mu N_i - \lambda_i \frac{S_i}{N_i} - \mu S_i \tag{192}$$

$$\dot{I}_i \quad = \quad \lambda_i \frac{S_i}{N_i} - (\gamma_i + \mu) I_i, \tag{193}$$

where

$$\lambda_i = \sum_{j=1}^{n} \beta_{i\,j} I_j. \tag{194}$$

Setting the time derivative of S_i to zero gives

$$S_i^* = \frac{\mu}{\mu + \lambda_i^*}, \tag{195}$$

where λ_i^* is the force of infection that group i experiences at equilibrium.

Setting the time derivative of I_i equal to zero gives

$$\begin{aligned} I_i^* &= \frac{\lambda_i^* S_i^*}{\mu + \gamma_i} \\ &= \frac{\mu \lambda_i^*}{(\mu + \gamma_i)(\mu + \lambda_i^*)}. \end{aligned} \tag{196}$$

Substituting into equation (194) for the force of infection gives the following set of simultaneous equations

$$\begin{aligned} \lambda_i^* &= \sum_{j=1}^{n} \beta_{i\,j} I_j^* \\ &= \sum_{j=1}^{n} \beta_{i\,j} \frac{\mu \lambda_j^*}{(\mu + \gamma_j)(\mu + \lambda_j^*)}. \end{aligned} \tag{197}$$

Once these equilibrium forces of infection are found, we can find the equilibrium numbers of susceptibles and infectives by substitution.

14.6.1 Mass Vaccination and Disease Eradication

If we were to vaccinate a fraction p_i of new-borns in group i, we would replace the μN_i birth term by $\mu N_i(1 - p_i)$. This would propagate through the equilibrium analysis just discussed, and lead to the following equations for the equilibrium forces of infection:

$$\lambda_i^* = \sum_{j=1}^{n} \beta_{i\,j} \frac{\mu(1 - p_j)\lambda_j^*}{(\mu + \gamma_j)(\mu + \lambda_j^*)}. \tag{198}$$

Disease eradication would correspond to λ_i^* approaching zero. Taking each λ_i^* to be small, and expanding the above equations, we see that the small λ_i^* limit is

$$\lambda_i^* = \sum_{j=1}^{n} \beta_{i\,j} \frac{(1 - p_j)\lambda_j^*}{\mu + \gamma_j}. \tag{199}$$

This is a set of linear equations for the λ_i^*, and can be written in matrix form as

$$\lambda^* = B\lambda^*, \tag{200}$$

where λ^* is the vector of forces of infection and B is a matrix with entries

$$b_{i\,j} = \beta_{i\,j}(1 - p_j)/(\mu + \gamma_j). \tag{201}$$

In order for this to hold, we must have that $(B - I)\lambda^* = 0$. For there to be a non-trivial solution, the determinant of $B - I$ must equal zero. This gives us an equation that constrains the p_i.

14.6.2 A Special Case: Separable Mixing

If we assume that the beta matrix is separable, then we can derive a condition for this determinant to equal zero. In appendix E of Anderson & May, it is shown that if we have a matrix A, whose entries are $a_{i\,j} = a_i b_j - c_i \delta_{i\,j}$, then (assuming that $c_k/(a_k b_k) \neq 0$ for all k) the determinant of A vanishes if

$$1 = \sum_{j=1}^{n} \frac{a_j b_j}{c_j}. \tag{202}$$

(You may notice that there is a striking similarity between this expression and (186), although bear in mind that there are differences in the interpretations that underlie the equations. Notice that in order for the determinant of a matrix to equal zero, the matrix must have a zero eigenvalue. If we let λ equal zero in (186), we obtain (202).)

If we have that $\beta_{i\,j} = g_i h_j$, then we have $a_i = (1 - p_i)g_i$, $b_j = h_j/(\mu + \gamma_j)$ and $c_k = 1$. Substituting these into (202) gives

$$1 = \sum_{j=1}^{n} (1 - p_j) \frac{g_j h_j}{\mu + \gamma_j}. \tag{203}$$

14.6.3 Eradication Criteria

If we employ **uniform vaccination**, the factors $(1 - p_j)$ that appear in the eradication criterion (203) are all equal and do not depend on j. Hence the factor $(1 - p)$ can be written outside of the summation. We see that the sum that remains is, in fact, equal to R_0. Rearranging this equation shows that, under uniform vaccination (in which individuals are vaccinated at random, without reference to their activity level), the critical vaccination proportion is just $p_c = 1 - 1/R_0$. This is exactly what we saw in the simple models.

Remember that heterogeneity in the activity levels meant that R_0 is greater than it would have been in a corresponding homogeneous (well-mixed) population. So heterogeneity increases the critical vaccination proportion under uniform vaccination.

It turns out that the heterogeneity can be exploited in order to reduce the proportion of the population that need be vaccinated. Anderson and May discuss **optimal vaccination strategies** in which the vaccination fractions p_i differ between groups. For them, optimal means finding the minimum overall fraction of the population that need be vaccinated in order to achieve eradication. This overall fraction can be written as $\Sigma p_i N_i/N$.

Anderson and May find such optimal strategies using Lagrange multiplier techniques. (We don't worry about the details here.) The general message is that the optimal strategy is to concentrate effort on vaccinating the high risk individuals. This should not be at all surprising, given our observation that such individuals are responsible for a disproportionate fraction of transmission of the infection.

One particular example of an optimal vaccination strategy is particularly striking. Anderson and May's 'Bang-Bang' strategy says to vaccinate everyone in the highest risk groups and work downwards, vaccinating the next riskiest group, and so on, until one manages to reduce R_0 below one. No-one should be vaccinated in the lowest risk groups.

Given that there can be a large difference between R_0 in the well-mixed and heterogenous activity level models, it should not be surprising that there can be large differences in the effectiveness of uniform and targeted control measures. If there is a high degree of heterogeneity, uniform vaccination can be highly inefficient: most individuals are in low-risk groups and so little is gained by vaccinating them. On the other hand, targeted vaccination can be highly effective, and one can achieve control by vaccinating a smaller proportion of the population than would be the case in a well-mixed population.

A major caveat of targeted vaccination is that the high risk individuals have to be identified. The benefit of the smaller number of vaccinations needed must be weighed against the cost of having to track down the high risk groups. In some cases, these groups might be well defined (e.g. sex workers), but it may not always be so easy. The success of a targeted vaccination campaign depends crucially on the ability to reach high risk individuals.

14.7 More General Multigroup Models and the Calculation of R_0

The multigroup models that we have discussed to this point have assumed that individuals do not change their type (group). More general models allow individuals to move between groups: this might occur, for example, in a model depicting the spatial location of individuals if the individuals are allowed to move around.

The next generation matrix approach can still be used to calculate R_0, although there can be complications in the calculation of the entries of the matrix. As an example, in a spatial model, an individual could move between locations on several occasions during their infectious period. Situations in which an infection has a complex lifecycle can also present challenges to the deployment of the approach. We illustrate the complexities that can arise by means of a simple, two group, example, before discussing an algorithmic approach that allows the calculation of the next generation matrix in general settings.

Imagine an SIR-type infection for which their is some treatment that can be administered to infectious individuals. The number of treated individuals is denoted by T. We assume that infectious individuals move into the treated class at a per-capita rate α. The treatment is not 100% effective, so treated individuals can still give rise to new infections, although with a smaller transmission parameter than the untreated infectives. The infection rate in the model is given by $(\beta_1 I + \beta_2 T)S/N$. Infectives recover at per-capita rate γ_1 and treated individuals recover at per-capita rate γ_2. Finally, the treatment does not have a permanent effect, so treated individuals move back to the infective class at per-capita rate δ. The flowchart for this model is shown in figure 49.

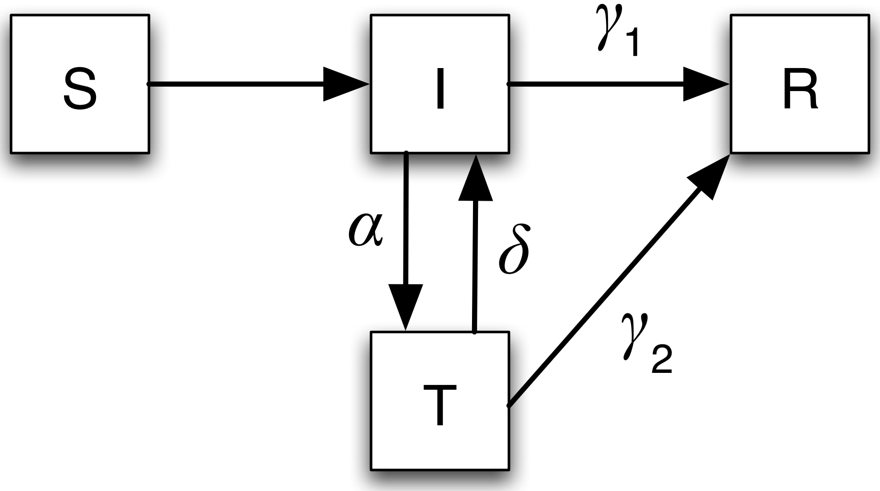

Figure 49: SIR model with treatment. Infective individuals are treated at rate α, but treatment loses its effectiveness at rate δ.

In this model, individuals can change their type, moving between the I and T classes. An individual might move back and forth several times before they recover, and so the average time spent in the I and T classes is not immediately obvious. We let T_I be the average time an individual spends in I and T_T be the average time they spend in T. The calculation revolves around determining how many times they visit each class before recovering.

On any visit to the I class, there are two ways that an individual can leave: either they get treated or they recover. The probability that they will leave because they get treated is $\alpha/(\alpha+\gamma_1)$, while the probability that they leave because they recover is $\gamma_1/(\alpha+\gamma_1)$. The average time spent in the I class during any one visit is $1/(\alpha+\gamma_1)$.

On any visit to the T class, there are two ways that an individual can leave: either their treatment wears off (and they move to the I class) or they recover. The probability that they will leave because the treatment wears off is $\delta/(\delta+\gamma_2)$, while the probability that they leave because they recover is $\gamma_2/(\delta+\gamma_2)$. The average time spent in the T class during any one visit is $1/(\delta+\gamma_2)$.

We write the probability that an individual who is in the I class will make another visit to the I class as ϕ. This is given by the product of the probability of them getting treated (i.e. moving from I to T) and the probability that the treatment wears off (i.e. moving from T to I). So we have

$$\phi = \frac{\alpha\delta}{(\alpha+\gamma_1)(\delta+\gamma_2)}. \tag{204}$$

It is easy to see that the probability of an individual who is in the T class making another visit to the T class is also equal to ϕ.

A newly infected individual finds themselves in the I class, so the probability that they make at least one visit to the class is 1. Each time they find themselves in the I class, the probability that they will return (and hence make one or more additional visits) is ϕ and the probability that they will not is $1-\phi$. Hence the probability that an individual who has become infected makes exactly k visits to the I class, for $k \geq 1$, is equal to $\phi^{k-1}(1-\phi)$. Thus the number of visits is geometrically distributed, and it is straightforward to show that the average number of visits is $1/(1-\phi)$. As a result, the average time spent in the I class is

$$\begin{aligned} T_I &= \frac{1}{\alpha+\gamma_1}\frac{1}{1-\phi} \\ &= \frac{1}{\alpha+\gamma_1}\frac{(\alpha+\gamma_1)(\delta+\gamma_2)}{(\alpha+\gamma_1)(\delta+\gamma_2)-\alpha\delta} \\ &= \frac{\delta+\gamma_2}{(\alpha+\gamma_1)(\delta+\gamma_2)-\alpha\delta}. \end{aligned} \tag{205}$$

A similar argument shows that the probability of making exactly k visits to the T class, for $k \geq 1$, is $\phi^{k-1}(1-\phi)\alpha/(\alpha+\gamma_1)$. The average number of visits is then $\alpha/\left\{(\alpha+\gamma_1)(1-\phi)\right\}$, leading to

$$T_T \;=\; \frac{\alpha}{(\alpha+\gamma_1)(\delta+\gamma_2)-\alpha\delta}. \tag{206}$$

We notice that if $\gamma_1 = \gamma_2 = \gamma$, then $T_I + T_T = 1/\gamma$: if the recovery rate is the same for both classes, then the total time spent in I and T is just the reciprocal of this rate. This is to be expected in this case, since if we only care about how long an individual spends in I and T in total, we need not worry about whether they are labeled I or T at any point in time. From the viewpoint of their recovery time, we can ignore the shuffling back and forth between classes. Of course, if β_1 and β_2 differ then this shuffling does have an effect on the number of secondary infections that result.

An individual's movements between classes depend only on their current state: they are independent of the previous history of the individual. (Recall that we discussed how having constant transition rates, which lead to exponentially distributed residence times, corresponds to the system being memoryless. As discussed in previous chapters, the biological adequacy of the constant rate assumption may, however, be open to question.) Thus the average time that an individual will spend in I in the future does not depend on how long they have been in the class or on how many occasions they have previously been in the class.

This observation provides a slightly simpler way to obtain T_I and T_T, using a technique that is commonly used in Markov chain theory, namely conditioning on what happens on the first departure from the class. With probability $1-\phi$ there will not be another visit to I and the average time spent in the class will be $1/(\alpha+\gamma)$. With probability ϕ another visit will occur, in which case $1/(\alpha+\gamma)$ will be spent in I on the current visit and an average of T_I will be spent in I during future visits. This gives

$$\begin{aligned} T_I \;&=\; (1-\phi)\frac{1}{\alpha+\gamma_1} + \phi\left(\frac{1}{\alpha+\gamma_1} + T_I\right) \\ &=\; \frac{1}{\alpha+\gamma_1} + \phi T_I. \end{aligned} \tag{207}$$

Solving this equation gives the expression for T_I that we derived above, but without having to sum an infinite series. A similar argument can be used to obtain T_T.

The next generation matrix is calculated at the infection free equilibrium, when the number of susceptibles is N and so the transmission term is of the form $\beta_1 I + \beta_2 T$. A newly infected individual will spend T_I time units in the I class, during which time they give rise to an average of $\beta_1 T_I$ secondary infections. They also spend T_T time units in the T class, giving rise to an average of $\beta_2 T_T$ secondary infections. All of these secondary infections are of type I, and so the next generation matrix is of the form

$$\begin{pmatrix} \beta_1 T_I + \beta_2 T_T & \cdot \\ 0 & 0 \end{pmatrix} \tag{208}$$

Notice that the entries of the second row of this matrix are both zero: all new infections are of the I type. The transitions between I and T do not represent the appearance of newly infected individuals. Since the matrix is upper triangular, its eigenvalues are simply the entries that appear on the principal diagonal: 0 and $\beta_1 T_I + \beta_2 T_T$. (Notice that we did not calculate the $(1, 2)$ entry of the matrix: its value does not affect the eigenvalues.) We have that

$$\begin{aligned} R_0 &= \beta_1 T_I + \beta_2 T_T \\ &= \frac{\beta_1(\delta + \gamma_2) + \beta_2 \alpha}{(\alpha + \gamma_1)(\delta + \gamma_2) - \alpha\delta}. \end{aligned} \tag{209}$$

14.8 A General Recipe for the Calculation of the Next Generation Matrix

The next generation matrix describes the linearization of the model about the infection free equilibrium, so only the classes that correspond to infected individuals need be considered. (This includes all individuals who have the infection, including all latent and infectious classes.) We denote the number of such classes by m, and label the numbers of individuals in these classes as x_i. We write these numbers in the vector $\mathbf{x}$.

For each class, we have a differential equation of the form

$$\frac{dx_i}{dt} = f_i(\mathbf{x}) \tag{210}$$

describing the flow into and out of the class. Here, the index i runs from 1 to m. We split up the function $f_i(\mathbf{x})$ as

$$f_i(\mathbf{x}) = \mathscr{F}_i(\mathbf{x}) - \mathscr{V}_i(\mathbf{x}), \tag{211}$$

where $\mathscr{F}_i(\mathbf{x})$ represents the rate at which newly infected individuals appear in class i and $\mathscr{V}_i(\mathbf{x})$ represents all other flows in and out of class i. Further, we split up $\mathscr{V}_i(\mathbf{x})$ as

$$\mathscr{V}_i(\mathbf{x}) = \mathscr{V}_i^-(\mathbf{x}) - \mathscr{V}_i^+(\mathbf{x}). \tag{212}$$

Here $\mathscr{V}_i^+(\mathbf{x})$ represents the flows that lead to individuals leaving the class and $\mathscr{V}_i^+(\mathbf{x})$ represents those that lead to individuals arriving in the class (except for the newly infected individuals).

The Jacobian matrices, F and V, of the vector functions $\mathscr{F}(\mathbf{x})$ and $\mathscr{V}(\mathbf{x})$ are calculated. These matrices have components

$$F_{ij} = \frac{\partial \mathscr{F}_i}{\partial x_j} \text{ and } V_{ij} = \frac{\partial \mathscr{V}_i}{\partial x_j}, \tag{213}$$

with all partial derivatives being evaluated at the infection free equilibrium.

The next generation matrix is then calculated as FV^{-1}, and hence the basic reproductive number can be found as its dominant eigenvalue.

14.8.1 Previous Example, Revisited

We illustrate the method using our two-group example. We have

$$\begin{aligned} \frac{dI}{dt} &= (\beta_1 I + \beta_2 T)\,\frac{S}{N} - (\alpha + \gamma_1)\, I + \delta T && (214) \\ \frac{dT}{dt} &= \alpha I - (\delta + \gamma_2)\, T. && (215) \end{aligned}$$

We write

$$\mathscr{F}(I,T) = \begin{pmatrix} (\beta_1 I + \beta_2 T)\, S/N \\ 0 \end{pmatrix}. \tag{216}$$

Notice that the second component is zero since newly infected individuals only appear in the I class. We also have

$$\begin{aligned} \mathscr{V}(I,T) &= \begin{pmatrix} (\alpha + \gamma_1)\, I - \delta T \\ (\delta + \gamma_2)\, V - \alpha I \end{pmatrix}, && (217) \\ \mathscr{V}^{+}(I,T) &= \begin{pmatrix} \delta T \\ \alpha I \end{pmatrix} && (218) \\ &\text{and} && (219) \\ \mathscr{V}^{-}(I,T) &= \begin{pmatrix} (\alpha + \gamma_1)\, I \\ (\delta + \gamma_2)\, T \end{pmatrix}. && (220) \end{aligned}$$

Evaluating the Jacobian matrices at the infection free equilibrium ($S = N$, $I = T = 0$) gives

$$F = \begin{pmatrix} \beta_1 & \beta_2 \\ 0 & 0 \end{pmatrix} \tag{221}$$

and

$$V = \begin{pmatrix} \alpha + \gamma_1 & -\delta \\ -\alpha & \delta + \gamma_2 \end{pmatrix}. \tag{222}$$

Inverting this second matrix gives

$$V^{-1} = \frac{1}{(\alpha + \gamma_1)(\delta + \gamma_2) - \alpha\delta} \begin{pmatrix} \delta + \gamma_2 & \delta \\ \alpha & \alpha + \gamma_1 \end{pmatrix} \tag{223}$$

and so

$$FV^{-1} = \frac{1}{(\alpha + \gamma_1)(\delta + \gamma_2) - \alpha\delta} \begin{pmatrix} \beta_1\,(\delta + \gamma_2) + \beta_2 \alpha & \beta_1 \delta + \beta_2\,(\alpha + \gamma_1) \\ 0 & 0 \end{pmatrix}. \tag{224}$$

This next-generation matrix agrees with the one calculated earlier, as does the resulting value of R_0.

We also now have the upper right entry of this matrix: this gives the average number of secondary infections that would result if a treated individual were directly introduced into an entirely susceptible population. (Of course, this event could never happen naturally over the course of an outbreak since newly infected individuals appear in the I class.)

15 Spatial Models

Transmission of infection often depends on spatial location. In most cases, people tend to interact more often with people who are nearby: transmission has a strong local component.

We shall look at three types of spatial model: (i) patch (metapopulation) models, (ii) spatially continuous models and (iii) network models.

15.1 Patch Models

Patch models are a particular type of the multi-group model where the groups represent different spatial locations. For instance, we might think of a collection of cities. In the ecological literature, this sort of model is often called a metapopulation model: the entire population is thought of as a collection of smaller populations.

We often assume that each patch is well-mixed and that there is some sort of "coupling" between patches. This coupling term describes how the infection can be transmitted between patches. Two different types of coupling are commonly employed:

- **Cross-coupling formulation**: as in the multi-group model, we take the force of infection in patch i to be a weighted sum over the numbers of infectives in the other patches. We have $\lambda_i = \sum \beta_{ij} I_j / N_i$, so that the transmission term is $S_i \sum \beta_{ij} I_j / N_i$. (We shall discuss the N_i denominators later.)

  We have a matrix of β_{ij} coefficients, often called the beta matrix, that describes transmission between and within patches.

  The interpretation is that individuals can acquire infection from infectives in other patches, but individuals do not move permanently between patches. (For instance, we might imagine that people make short visits to other patches, but always return to their home patch.) Notice that the coupling enters the model in a nonlinear way.

- **Diffusion formulation**: in this formulation, transmission only occurs within patches, but individuals can move between patches. For instance

$$\dot{S}_i = \mu N_i - \frac{\beta_i S_i I_i}{N_i} - \mu S_i + \sum r_{ij}^{(S)} S_j \qquad (225)$$

$$\dot{I}_i = \frac{\beta_i S_i I_i}{N_i} - (\mu + \gamma) I_i + \sum r_{ij}^{(I)} I_j. \qquad (226)$$

  Here $r_{ij}^{(S)}$ is the rate at which susceptibles in patch j move to patch i, and similarly for $r_{ij}^{(I)}$.

> Notice that $r_{ii}^{(S)}$ must be less than or equal to zero since this term depicts individuals that depart patch i. Notice that if individuals cannot leave the system by migration then we must have $\sum_i r_{ij} = 0$: individuals that leave patch j (r_{jj}) must appear elsewhere ($\sum_{i \neq j} r_{ij}$).
>
> Notice that in this diffusion-like description, the spatial coupling enters the model via linear terms.

It turns out that there are similarities between the two formulations. For instance there is a particular parameterization that allows us to move from one formulation to the other while preserving equilibrium values. The dynamics of the models about their equilibria may, however, differ. For instance, the damping time of the equilibria in the two models are unlikely to be the same.

15.1.1 Synchrony in Metapopulation Models

Disease fadeout is an important issue in the dynamics of measles: from our earlier discussion, in a given region there was a continual cycle of disease outbreaks followed by fadeouts. In order for this to continue, the infection must be reintroduced at some point after fadeout. This process can be modeled within the metapopulation framework, and questions can be asked about persistence of the infection on both the local and region scales. This is a well-studied question in ecology.

A key issue concerns the synchrony between outbreaks in different patches. If we have simultaneous fadeout in each patch, then we couldn't get reintroduction. If fadeouts are not synchronized then we can get reintroduction of infection from a patch that has not undergone fadeout. Asynchrony between outbreaks (and hence fadeouts) can enhance persistence of the infection at the regional level.

Spatial structure has been suggested as a way in which models can describe both persistence and temporal patterns in measles data. (Recall our earlier discussion: simple models can either reproduce the persistence pattern or the temporal pattern, but find it difficult to reproduce both simultaneously.)

An interesting question concerns the impact of vaccination on synchrony. Statistical studies of epidemic data suggest that measles outbreaks become less synchronized upon vaccination. This makes it more difficult to eradicate measles than we might expect if spatial structure were ignored.

What factors affect synchrony?

- **spatial coupling**: Stronger coupling enhances synchrony
- **seasonality**: This global forcing tends to enhance synchrony. (Unless, maybe, the seasonality is strong enough to give rise to chaotic dynamics?)

- **stochasticity**: Random effects will tend to desynchronize patches

It turns out that there is a complex interaction between these factors and the intrinsic predator-prey dynamics of the system.

Regional-level persistence of the infection is affected by the level of coupling and the level of synchrony. If the coupling between patches is weak then there is little chance that infection will be reintroduced following fadeout. If the level of coupling is high then the fact that patches are highly synchronized means that they will tend to undergo fadeout together. In this case there will again be little chance of reintroduction following fadeout. This argument suggests that regional-level persistence will be greatest at intermediate levels of coupling.

15.1.2 City to City Spread of Infection

One setting in which these metapopulation models have been employed is to describe the spread of influenza between cities in the USSR. The diffusion formulation was employed with between-city movement described using known transport patterns (such as rail links). It is claimed that this model can describe the general temporal trends of the epidemic as it moves across the country.

Stochastic effects are likely to be very important in spatial settings. If coupling between cities is weak, so that only a few cases in one city are directly due to individuals moving from another, a deterministic description would be deficient in important ways. For instance, imagine a newly appeared highly infectious disease, for which the entire population is initially susceptible. Imagine two cities, A and B, with infection present in A but not B and with the daily probability of transmission from place A to place B equalling 0.1 (ignore the fact that this probability would change over the course of an outbreak in place A). It would likely take several days before infection in place A would give rise to a case in place B, after which an outbreak could occur in place B. Since an ODE model quite happily talks about fractional individuals being infected, it would typically describe a continual low-level of transmission from A to B. As soon as any infection appeared in place B, it would grow exponentially, causing an outbreak. In a stochastic model setting, an outbreak in place B could only possibly occur after the introduction of one or more infections there. It should be clear that the outbreak will typically occur later in the stochastic setting than the deterministic.

The importance of stochastic effects makes it unlikely that we could predict details of transmission in a spatial setting.

15.1.3 Particular Forms for Coupling

A well-known weakness of spatial models arises because estimation of between-patch coupling is often difficult. This is partly because between-patch transmission is assumed to be weak compared to within-patch transmission.

In the absence of detailed data on spatial coupling, modelers will often pick simple forms for the coupling. This often involves assuming that each patch is of equal size ($N_i = N$) and that the coupling matrix is symmetric. In the cross-coupled formulation, the following choices are commonly made

- Equal coupling between patches. Here we have that the within-patch transmission parameters are equal to β and the between-patch transmission parameters are equal to $\epsilon\beta$ for each pair of patches, where ϵ lies between zero and one. Often, ϵ is assumed to be small. In a four patch setting, the beta matrix is given by

$$\beta \begin{pmatrix} 1 & \epsilon & \epsilon & \epsilon \\ \epsilon & 1 & \epsilon & \epsilon \\ \epsilon & \epsilon & 1 & \epsilon \\ \epsilon & \epsilon & \epsilon & 1 \end{pmatrix}. \tag{227}$$

- Nearest neighbor coupling. We again take the within-patch transmission parameter to equal β and the between-patch parameter to equal $\epsilon\beta$ but we only allow coupling between a patch and its nearest neighbors. We typically imagine that the patches are arranged in a linear array, so we have $\beta_{ii} = \beta$, $\beta_{i\,(i\pm1)} = \epsilon\beta$ and $\beta_{ij} = 0$ otherwise. (We could, of course, generalize to other situations such as where the patches are arranged on a two dimensional lattice.) Notice that in this one dimensional setting, the patches on the ends of the line only have one neighbor. To get around this asymmetry, we often imagine that the patches are arranged on a circle: this choice of periodic boundary conditions means that each patch is treated in the same way and (in a four patch setting) leads to the following beta matrix

$$\beta \begin{pmatrix} 1 & \epsilon & 0 & \epsilon \\ \epsilon & 1 & \epsilon & 0 \\ 0 & \epsilon & 1 & \epsilon \\ \epsilon & 0 & \epsilon & 1 \end{pmatrix}. \tag{228}$$

In order to calculate R_0 in these cross-coupled settings, we need to find the dominant eigenvalue of the next generation matrices. These are (provided that $\epsilon > 0$) non-negative primitive matrices: the dominant eigenvalue corresponds to their only positive eigenvector. The symmetry of the matrices implies that $(1111)^{\mathrm{T}}$ is an eigenvector, and so we have that

$$R_0 = \frac{\beta}{\mu + \gamma}(1 + 3\epsilon) \tag{229}$$

in the four patch equal coupling case, and

$$R_0 = \frac{\beta}{\mu+\gamma}(1+2\epsilon) \tag{230}$$

in the four patch nearest neighbor coupling case.

Notice that R_0 increases with ϵ. In the cross-coupled model, spatial coupling corresponds to individuals making additional (between-patch) contacts. In this formulation, infectives in a given patch can simultaneously infect individuals in more than one patch. We could normalize the transmission parameters so that R_0 is independent of ϵ, for instance:

$$\frac{\beta}{1+3\epsilon}\begin{pmatrix} 1 & \epsilon & \epsilon & \epsilon \\ \epsilon & 1 & \epsilon & \epsilon \\ \epsilon & \epsilon & 1 & \epsilon \\ \epsilon & \epsilon & \epsilon & 1 \end{pmatrix}. \tag{231}$$

Here, a fraction $1/(1+3\epsilon)$ of an individual's contacts are within their own patch and $3/(1+3\epsilon)$ of their contacts are with other patches.

15.1.4 Cities and Villages

One example where the different patches are of different sizes is Anderson and May's "cities and villages" model in which there is one large city that is weakly coupled to a number of smaller villages. Their model is phrased in terms of the fraction, f, of indivdiduals that are found in the city. They look at optimal vaccination strategies, asking the question of whether it is better to vaccinate individuals in the city or in the villages.

For Anderson and May, the optimal vaccination strategies tended to concentrate effort very strongly in the city. They predicted a considerable difference between the fraction needed to be vaccinated under uniform (which means that each individual has the same chance of being vaccinated, regardless of whether they live in the city or a village) and optimal strategies (which concentrate vaccinations in the cities).

Anderson and May's work has been criticized by other authors, who point out that their model does not include the N_i denomintors in the transmission terms. This means that, for them, R_0 is much greater in the city than in the villages. (Recall that our expressions for R_0 were typically of the form $\beta/(\mu+\gamma)$, independent of N, whereas Anderson and May have $\beta N/(\mu+\gamma)$, which is proportional to N.) Hethcote and van Ark argue that R_0 is not much smaller in the villages: the transmission terms should have the $1/N_i$ scaling.

If the transmission terms scale in the regular way, as suggested by Hethcote and van Ark, then there are still differences between uniform and optimal vaccination strategies. These differences, however, are not as great as Anderson and May claim. In any case, using the optimal vaccination strategy always requires fewer vaccination than the uniform strategy in order to eradicate infection.

The cities and villages example is interesting in the stochastic setting. It is often the case that there is a large urban center surrounded by a number of smaller population centers. The city may be sufficiently large that its size is above the critical community size for an infection, while the rural population centers are below the CCS. The infection will, therefore, remain endemic in the city but undergo frequent fadeout elsewhere. A graphic description of this situation is one of a fire (the large city) throwing out sparks that cause little fires elsewhere (introduction/outbreak/fadeout dynamics in smaller towns).

15.2 Spatially Continuous Models

We shall start off by looking at the simplest deterministic model

- The population is assumed to be continuously distributed across space: we refer to "population densities". The position in space is labeled by the variable $\mathbf{x}$, so we have the susceptible and infective densities $S(\mathbf{x},t)$ and $I(\mathbf{x},t)$.
- Infection is assumed to be a local event: at a given point in space, the transmission depends on the local densities of S and I according to $\beta S(\mathbf{x},t)I(\mathbf{x},t)$. (In general, the transmission parameter β might also be a function of space.)
- The infection spreads spatially because individuals move about at random, as described by a random walk or diffusion process.

For a simple SIR process, ignoring demography, we have the following equations

$$\frac{\partial S}{\partial t}(\mathbf{x},t) = -\beta S(\mathbf{x},t)I(\mathbf{x},t) + D_S\nabla^2 S(\mathbf{x},t) \tag{232}$$

$$\frac{\partial I}{\partial t}(\mathbf{x},t) = \beta S(\mathbf{x},t)I(\mathbf{x},t) - \gamma I(\mathbf{x},t) + D_I\nabla^2 I(\mathbf{x},t). \tag{233}$$

This partial differential equation describes the local processes of infection and recovery and the diffusion of individuals This is an example of a reaction-diffusion model, a class of models that has a long history of use in biomathematics.

The operator ∇^2 (the Laplacian operator) gives the second derivative with respect to the spatial variables. For example $\nabla^2 = \partial^2/\partial x^2$ in one dimension, and $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ in two spatial dimensions. The speed of diffusion is controlled by the diffusion coefficients D_S and D_I: these may differ if a person's movement depends on their disease status. This description of movement can be derived from the assumptions of a random walk process.

If the infection can be described as an SI process, and the densities can be scaled so that $S(\mathbf{x},t)+I(\mathbf{x},t)=1$, then the following model is obtained in the one dimensional setting

$$\frac{\partial I}{\partial t}(\mathbf{x},t) \quad = \quad \beta\left\{1-I(\mathbf{x},t)\right\}I(\mathbf{x},t)-\gamma I(\mathbf{x},t)+D_I\frac{\partial^2 I}{\partial x^2}(\mathbf{x},t). \tag{234}$$

This equation is of particular historical interest: it is the Kolmogorov-Fisher equation from population genetics, in which context it dates back to the 1930s.

15.2.1 Invasion and Traveling Waves

As ever, the invasion setting is of great interest and provides a setting in which detailed mathematical analysis is possible. It is interesting to note that there are many analogies between the spatial invasion of an infection and ecological invasion theory.

The system is easiest to study if we assume that the population is initially both entirely susceptible and that its density is uniform across space.

If the infection is able to spread, then the introduction of infection leads to a traveling wave of infection spreading out from the initial infection focus, with circular wavefront.

At a particular point in space, a one-time epidemic will occur (assuming that the susceptible pool is not replenished by births or loss of immunity). The number of infectives will rise as the wave approaches only to fall again as the local susceptible density is depleted (see figure 50). Behind the wavefront, the infective density will approach zero and the susceptible density will approach S_∞. This is all analogous to what happens in the non-spatial setting.

We look for solutions of this model that are traveling waves moving at speed c. The standard way to do this is to write $z=x-ct$, corresponding to a shift of co-ordinates to a frame of reference that moves at speed c. In the z co-ordinates, it is as if we are sitting at the wavefront, moving along with it in space and time (see figure 50).

Using the Chain Rule and $z=x-ct$, we see that

$$\frac{\partial}{\partial x} \quad = \quad \frac{\partial z}{\partial x}\frac{\partial}{\partial z}=\frac{\partial}{\partial z} \tag{235}$$

and

$$\frac{\partial}{\partial t} \quad = \quad \frac{\partial z}{\partial t}\frac{\partial}{\partial z}=-c\frac{\partial}{\partial z}. \tag{236}$$

As long as we are not too far behind the wave-front, we can linearize the model by setting S equal to its initial density, which we write as S_0. For simplicity, we consider the one dimensional setting,

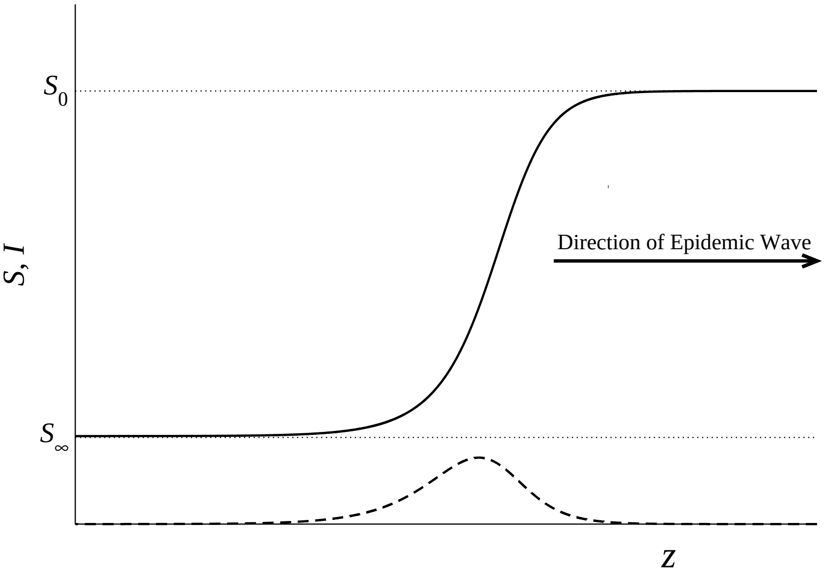

Figure 50: Spatial traveling wave in the reaction-diffusion version of the SIR model. We assume that the environment is spatially homogeneous. The solid curve denotes the susceptible density and the broken curve the infective density. The figure is illustrated in the co-ordinate z defined as $z = x - ct$. We can, therefore, view this figure in three ways: as a depiction of the susceptible and infective densities in the z frame that moves along with the wave, as a snapshot over space at a particular point in time (so that the horizontal axis is just a translated version of x) or as a time plot at a particular point in space (so that the horizontal axis depicts time, but rescaled by $-c$ and translated).

for which we have

$$\frac{\partial I}{\partial t} = \beta S_0 I - \gamma I + D_I \frac{\partial^2 I}{\partial x^2}. \tag{237}$$

In this linearized model, I will continue to grow exponentially as the wavefront passes: the depletion of susceptibles is ignored in the linear picture.

In terms of the variable z, the PDE reduces to the following ODE

$$-c\frac{dI}{dz} = (\beta S_0 - \gamma)I + D_I \frac{d^2 I}{dz^2}. \tag{238}$$

Keeping in mind the expected behavior (I is zero ahead of the wavefront and grows exponentially behind the wavefront), we try a solution of the form $I(z) = Ae^{rz}$. We need $I(z)$ to be positive and $I(z) \to 0$ as $z \to \infty$.

Substituting our trial solution gives the following quadratic for r

$$0 = D_I r^2 + cr + \beta S_0 - \gamma, \tag{239}$$

which has solution

$$r = -\frac{c}{2D_I} \pm \frac{c}{2D_I}\sqrt{c^2 - 4D_I(\beta S_0 - \gamma)}\,. \tag{240}$$

The number of infectives must be non-negative. This implies that r must be real (otherwise I would oscillate around zero, taking negative values at some points in time). Consequently, we have that the discriminant of the quadratic is non-negative, which implies that

$$c^2 \geq 4D_I(\beta S_0 - \gamma), \tag{241}$$

which means that

$$c \geq 2\sqrt{D_I(\beta S_0 - \gamma)}\,. \tag{242}$$

This says that there is a minimum wavespeed at which a traveling wave can move. We can rewrite this last expression as

$$c \geq 2\sqrt{D_I \gamma \left(R_0 - 1\right)}\,. \tag{243}$$

Here R_0 is the basic reproductive number, which equals $\beta S_0/\gamma$ (notice the appearance of the initial susceptible density here: unlike the non-spatial models, we did not have a denominator involving N in the transmission term).

It can be shown that the epidemic wave actually travels at this speed. Equation (243) provides a simple relationship between the epidemiological parameters (i.e. R_0 and γ), the diffusion coefficient (D_I) and the resulting wave speed (c).

This analysis has been used to describe the spread of the Black Death (bubonic plague) across Europe in medieval times. Ballpark estimates of the epidemiological and movement parameters were used to calculate the resulting wave speed. Its value was consistent with the observed spread.

The above analysis assumed that the susceptible density (and other environmental factors) were constant across space. If they vary then the local wave speed will vary. This leads to non-circular wavefronts, which may even "break up". A classic example of epidemic diffusion in a non-uniform environment is provided by Murray's study of rabies dynamics (see figure 51).

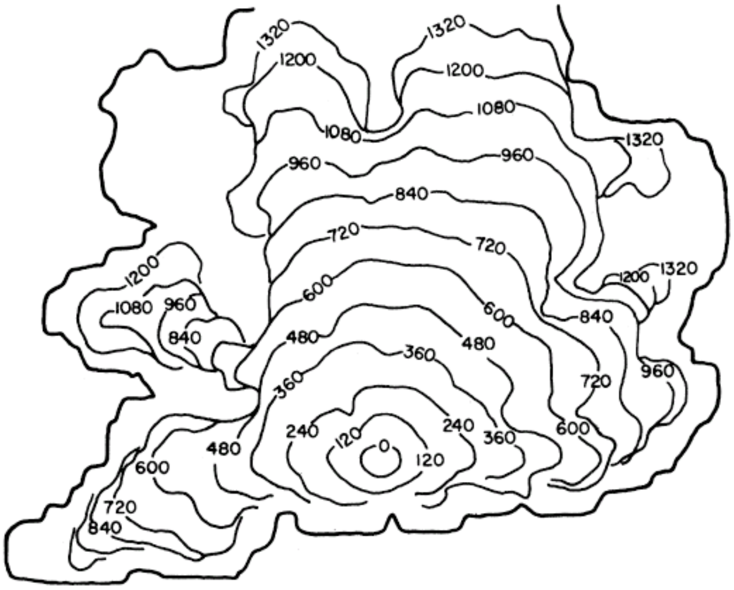

FIGURE 11. The position of the wavefront every 120 days as predicted by our model and the fox densities in figure 10. We assumed a diffusion coefficient of 200 km² per year, and took the other parameter values from table 1.

Figure 51: Spatial traveling wave in Murray's rabies model following introduction of rabies into the South of England. The fox density is taken to be non-uniform across the country. Notice how the epidemic wave-front is not circular, as the local wave speed depends on the fox density. This, and the following two figures are taken from Murray, Stanley and Brown (1986) Phil. Trans. R. Soc. Lond. B. **229** 111-150.

15.2.2 Secondary Traveling Waves

In the nonlinear model, the local depletion of susceptibles will cause the local epidemic to wane: I will decrease once the susceptible density falls below $1/R_0$. If we have births in the population, or individuals can recover to a susceptible state, then the local susceptible pool will be replenished over time. Eventually, the susceptible density will rise above threshold. If some infectives remain, a second epidemic can be triggered, leading to a secondary traveling wave of infection spreading out. Murray's rabies model exhibits this secondary traveling wave phenomenon (see figures 52 and 53).

In the non-spatial SIR model with demography we saw that the number of infectives often fell to very low levels (well below a fraction of an individual) following the initial epidemic. We saw that stochastic effects would, in many instances, lead to extinction of the infection in this period. This argument also applies to the spatial setting, perhaps even more so because we are thinking about the numbers of individuals in a local region.

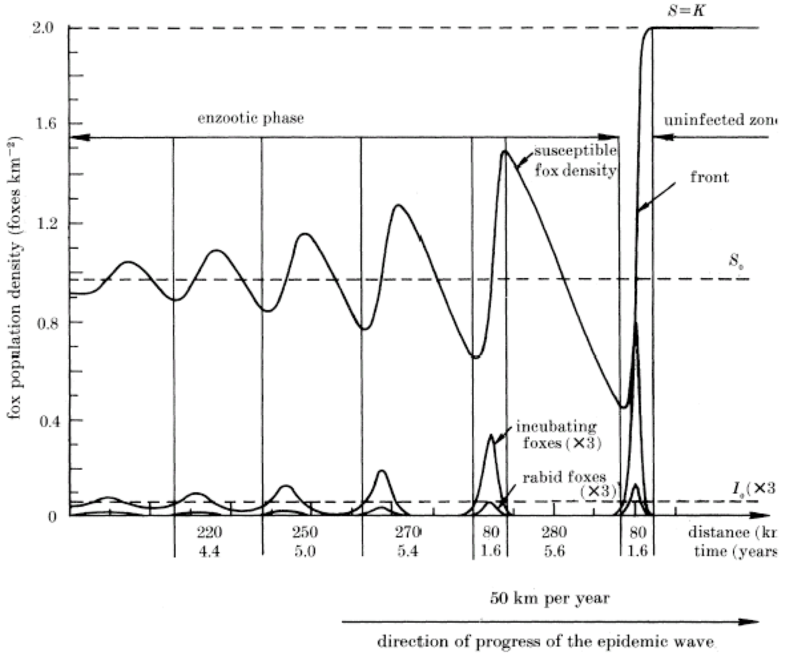

FIGURE. 1. Typical fluctuations in the fox populations due to the passage of a rabies epidemic

Figure 52: Epidemic and endemic behavior in Murray's rabies model. Notice that the figure takes a form that is familiar from similar non-spatial models.

Mollison made precisely this criticism of Murray's secondary epidemic waves by pointing out that persistence of the infection in his deterministic model was due to fractional levels of infected foxes ("atto-foxes") remaining in the wake of the primary wave. In a deterministic setting, infection can persist at such levels only to grow again once the susceptible threshold is exceeded. Mollison pointed out that the secondary wave would be extremely unlikely to occur in a stochastic setting where fadeouts could occur. This argument makes a good general point, namely that deterministic models are likely to be more troublesome in a spatial setting, partly because one is thinking about numbers of individuals in smaller locations.

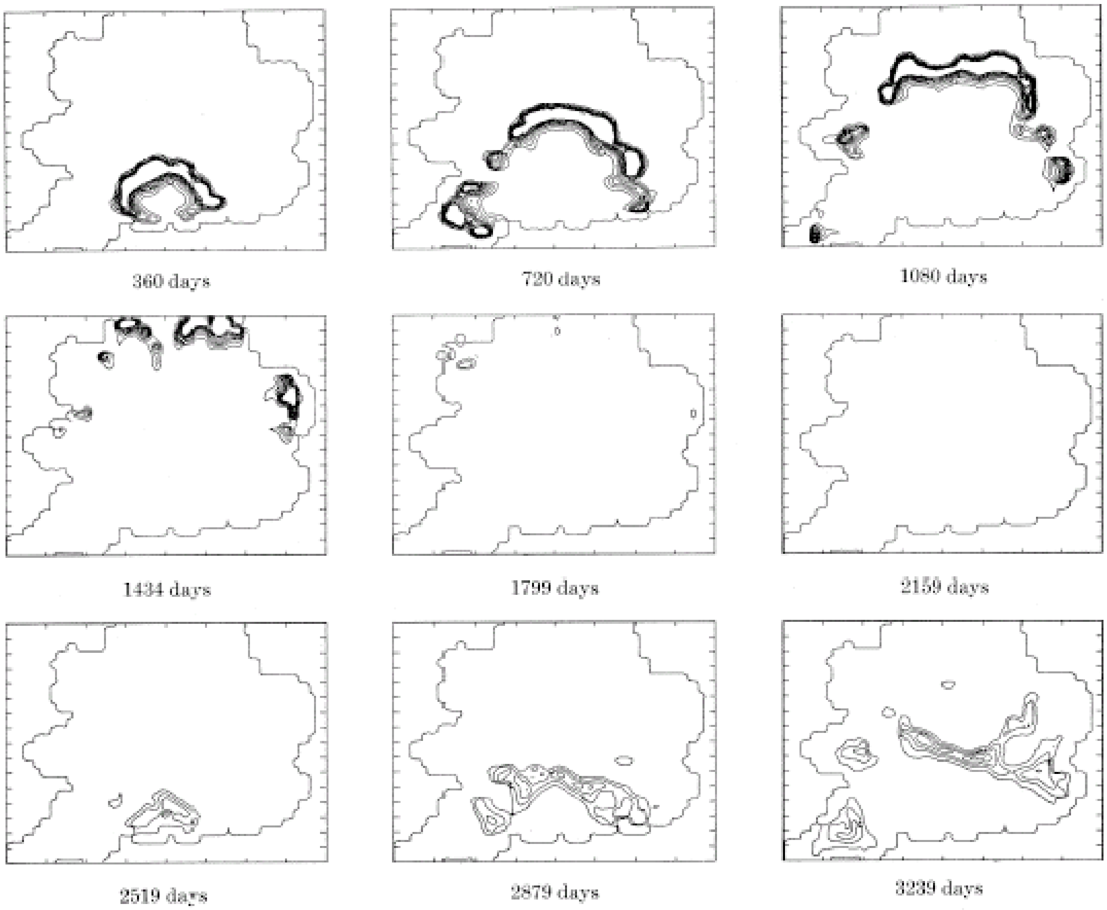

Figure 53: Primary and secondary traveling wave in Murray's rabies model following introduction of rabies into the South of England.

15.2.3 Control of Infection in Spatial Settings

Control of infection can be achieved by reducing the density of susceptibles so that the effective reproductive number falls below one. Better still, if the susceptible density can be reduced (and maintained) below $1/R_0$ then eradication will be possible.

If infection is localized in space then we can make use of this to facilitate control. For instance, we could concentrate a vaccination or treatment effort in the particular region where infection is present and perhaps neighboring regions. This is the idea behind "ring vaccination", which is the epidemiological analog of a forest fire-fighter's firebreak. Such strategies do not work if transmission has non-local aspects (e.g. if individuals travel outside this region).

In the rabies example, two possible control measures suggest themselves:the density of susceptible foxes could be reduced by vaccination or by killing foxes. The impact of these measures is modulated by the population dynamics of the foxes. Assuming that the vaccination affords long-lived protection, the former option leads to foxes being removed from the susceptible pool for at least a large part of the remainder of their lives. Culling foxes reduces the local population density of the species. Depending on the mechanisms that are regulating the local fox population density, the effect of culling could be very quickly overcome by increased survival and/or reproduction of the remaining individuals. For instance, with a reduction in the density, it is quite possible that the remaining foxes will have more food resources available to them. Another possibility is that foxes from neighboring non-culled regions could migrate into the now favorable habitat. Against this, culling has the benefit of directly reducing the local density of infectives. Analyses of the impact of the two alternative control measures have indicated that vaccination is the preferable option.

We remark that these population dynamic effects would not occur in the setting of farm animals: they are not free to reproduce and move about at will. In such settings, culling animals is often a more effective control strategy.

Murray's rabies model examined the spatial aspect of control: he looked at the impact of firebreaks. These zones in which the local density was reduced considerably slowed the spread of infection. Spatial spread could not, in their framework, be halted in this way. Again, this is a consequence of the use of a deterministic model. In a stochastic setting, effective firebreaks would lead to local extinction and hence halt spatial spread of an infection that was spreading via local diffusion.

15.2.4 More General Spatially Continuous Models

A more general version of the spatially continuous model, the reaction and dispersal kernel formulation, allows for non-local transmission. The local rate at which new infections arise, $\partial I(\mathbf{x}, t)/\partial t$, is allowed to depend on the local density of susceptibles and a weighted sum (integral) of the infective density elsewhere. These weights are described by a kernel function.

$$\frac{\partial I}{\partial t}(\mathbf{x},t) = S(\mathbf{x},t)\int\int K(\mathbf{x},\mathbf{y},\tau)\frac{\partial I}{\partial t}(\mathbf{y},t-\tau)\,d\tau d\mathbf{y}. \tag{244}$$

Here $K(\mathbf{x},\mathbf{y},\tau)$ describes the rate at which an individual located at $\mathbf{y}$ who was infected τ time units ago infects susceptible individuals located at $\mathbf{x}$. We remark that this formulation (via the dependence on the time since infection, τ) also allows for a general description of the timecourse of infection.

In many settings, the transmission process is assumed to only depend on the distance between individuals, so that the spatial dependence of the kernel function K enters only in terms of the distance between individuals, $|\mathbf{x}-\mathbf{y}|$.

In this setting, the speed of spread depends on the shape of the kernel function. In particular, substantial weight in the tails of the kernel enhance the spatial spread.

A general theory derives the wave speed in such settings. In a one dimensional spatial setting, we define

$$L(c,\lambda) = \int_{t=0}^{\infty}\int_{x=-\infty}^{\infty} e^{\lambda(x-ct)}K(x,t)\,dx\,dt. \tag{245}$$

The minimal wave speed, c_0, can then be found by solving

$$\begin{aligned} L(c_0,\lambda_0) &= 1 \qquad (246)\\ \frac{\partial L}{\partial\lambda}(c_0,\lambda_0) &= 0. \qquad (247)\end{aligned}$$

Convergence of the integral needs K to decay to zero sufficiently quickly as $x \to \pm\infty$. In fact, we need the tails of K to decay exponentially.

If K does not decay exponentially, then there can be jumps in the position of the wave-front, corresponding to an “infinite” speed of spread. The analysis of this model makes the importance of long-range transmission events clear. If these events are relatively infrequent, one may need to employ a stochastic formulation in order to satisfactorily capture the dynamics.

This formulation can be extended to cover non-isotropic spread. For instance, for an airborne infection, there may be a preferred direction of transmission due to the prevailing wind. In terms of the analysis, one simply transforms to a co-ordinate frame that moves in the preferred direction.

15.3 Network Models

Network models provide the most detailed description of the structure of the population. Individuals and their interactions are depicted as the nodes and edges of a graph. Associated with each node will be the infection status of the individual. The edges depict opportunities for transmission of the infection. It should be noted that whether transmission will occur depends both on whether there is an edge between two nodes and on the infection statuses of the nodes.

In many cases, infection is assumed to be equally likely along any edge of the graph (assuming that one node is susceptible and the other infectious). More generally, different edges could be assigned different weights, for instance representing stronger or weaker connections between pairs of individuals. Furthermore, transmission need not be symmetric: one instance might be if the male to female and female to male transmission probabilities differed in an STI setting. A directed graph could be employed in such situations.

In the simplest settings, the contact network is taken to be fixed: over time, population interactions occur between the same pairs of individuals. This assumption is often employed in models for infections that spread rapidly through a population.

The complexity of network models typically makes analysis difficult. We usually fall back on simulation approaches. Even this can be far from straightforward: if the population is large then simulation can be very slow and memory intensive. The code must account for every individual in the population, each of the transitions they make between infection classes, and keep track of all of the contacts of each member of the population.

The use of network approaches is also difficult in practical settings. The formulation of these models requires a large amount of detailed population data. Not only does the model need to know about each member of the population, it also needs to have information on who they contact within the population. Such detailed data is rarely available.

A useful way to gain general understanding of the behavior of network models is to examine behavior on particular types of networks. These canonical network types attempt to capture and depict particular network properties that are likely to be important for the spread of infection. Before we look at these canonical network types, we shall discuss the measures used to describe networks.

15.3.1 How Do We Describe Networks?

- **Connectedness:** a network is connected if it is possible to travel from any node to any other node by moving along edges of the network. Typically this will involve passing through a number of intermediate nodes.

- **Distance between nodes.** We denote the shortest distance between nodes u and v by $d(u, v)$. The diameter of a graph is the maximum value of $d(u, v)$ taken over all pairs (u, v). We also talk about the average shortest path between pairs of nodes, which is the value of $d(u, v)$ averaged over all pairs (u, v).

- **Degree of a node:** this is the number of neighbors of a node. We often call the degree the connectivity of a node, and write its value as k. We talk about the degree distribution or connectivity distribution, $P(k)$, which summarizes the degrees of all the nodes of the network. From this, we can calculate the average degree of the nodes of the network as $\langle k \rangle = \sum_k kP(k)$. Higher order moments such as the variance can also be calculated: $\mathrm{Var}(k) = \sum_k (k - \langle k \rangle)^2 P(k)$. This variance is a measure of the heterogeneity in the connectivity distribution.

- **Cliquishness or Clusteredness.** An important property of networks is the extent to which a pair of connected nodes share common neighbors. In figure (54), we see that nodes A and B share node C as a common neighbor. Cliques give rise to triangles in the network. One measure of cliquishness looks at the fractions of triples (node X connected to node Y connected to node Z) that give rise to triangles.

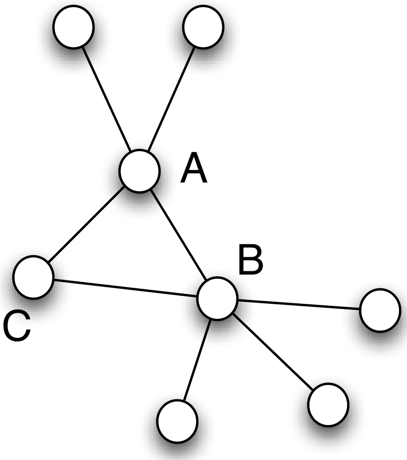

Figure 54: A clique in a network: nodes A and B share a common neighbor, node C. Notice that cliques lead to the presence of triangles in the network.

- **The mixing pattern.** As in the multi-group model, we talk about random, assortative and disassortative mixing. In many cases, we assume that the connectivity of nodes specifies their type, so the mixing pattern would be described in terms of how nodes of a given degree are connected to nodes of other degrees.

15.3.2 Mixing Types, Revisited

In this section we are thinking about the setting in which the nodes are distinguished by their degrees and so the mixing pattern is described in terms of the connectivities of the nodes.

We have a network with N nodes and degree distribution $P(k)$. Here, the values of k will play the role of the activity levels, a_k, in the earlier multi-group formulation.

The number of nodes of degree k is equal to $NP(k)$. Each node of degree k makes k connections, so the total number of connections made by individuals in the network is $\sum_k kNP(k)$. This quantity is the analog of the quantity D in the multigroup formulation. Since each edge involves two nodes, we have that

$$\sum_k kNP(k) = N\langle k\rangle = 2 \times (\text{number of edges in the graph}). \tag{248}$$

The fraction of connections made by individuals that are made by nodes of degree k is given by $kP(k)N/\sum_j jNP(j)$. These quantities are the analogs of the fractional activity levels in the multi-group formulation.

Random (proportionate) mixing assumes that a particular connection of a given node is made at random from the pool of available connections. Consequently, the probability that a given connection of an individual of connectivity k' is with an individual of degree k is given by

$$\frac{kNP(k)}{\sum_j jNP(j)}. \tag{249}$$

Notice that this quantity is not simply $P(k)$ since we select at random from the population of connections, not the population of individuals. More highly connected individuals make more connections than poorly connected individuals, so are more highly represented in the population of connections.

If we choose an individual at random from the population and then choose one of their neighbors at random, the probability that this second individual has connectivity k is given by $kP(k)N/\sum_j jNP(j)$. So we can calculate the average connectivity of neighbors of individuals:

$$\sum_k k\left\{\frac{kP(k)N}{\sum_j jNP(j)}\right\}. \tag{250}$$

This sum can be rewritten as

$$\frac{\sum_k k^2 P(k)}{\sum_j jP(j)}. \tag{251}$$

This sum should look familiar from the multi-group section, and is easily shown to equal $\langle k \rangle + \mathrm{Var}(k)/\langle k \rangle$. This is precisely the mean-variance formula from before.

In the proportionate (random) mixing setting, we see that heterogeneity in the degree distribution leads to the connectivity of randomly chosen neighbors of randomly chosen individuals being greater than that of randomly chosen individuals. In the sociological literature, friendship networks are commonly discussed. In that setting, this phenomenon has been talked about in terms of "Why your friends have more friends than you". As mentioned above, the origin of this effect is that highly connected individuals are more highly represented in the pool of connections than they are in the population (i.e. in the pool of individuals).

15.3.3 Canonical Network Types

- **Random Graph (Erdös-Renyi)**

  For a graph of N nodes, connections are made randomly and independently in the following sense: for each pair of nodes, the probability of there being an edge between the two nodes is equal to p. For each node, their connectivity is binomially distributed with mean $(N-1)p$, which is well approximated by Np if N is large. Connections are made without any regard to the nodes' spatial location, so there is essentially no spatial structure in the network. In some sense, the graph is "well-mixed", but notice that it is not true that every node is connected to every other node.

  An interesting property of random graphs concerns their connectedness. If p is "small", then the graph consists of a large number of small disconnected parts. If p is "large", then it consists of a single connected piece that contains most of the nodes (this is called the giant connected component), perhaps together with a few isolated small components and nodes. There is a critical value of p, written as p_c, at which this transition occurs. This critical value is somewhat analogous to the $R_0 = 1$ threshold.

  This notion is made concrete in the following: Let $\phi = Np$, the transition occurs when $\phi > 1$, after which the giant component contains a proportion z of the nodes, where z is the largest root of $z = 1 - \exp(-\phi z)$.

  The average shortest path of a random graph increases with the logarithm of the number of nodes in the random graph. This fairly slow increase with N reflects the fact that connections are made without regard for the location of nodes.

  Cliques are rare in random graphs and the connectivity distribution has relatively small variance.

- **Regular Lattice**

  Individuals are situated at the sites of a regular lattice, and connections are made to some neighborhood of each node. In a one dimensional setting, the nodes are situated on a line and connections are made to some number of the nearest neighbors. In a two dimensional setting, nodes could be connected to their 4 nearest neighbors (up, down, left and right): the von Neumann neighborhood of range one. An alternative is the Moore neighborhood of range one, which involves connections to the eight nearest neighbors (von Neumann neighborhood plus the four diagonal neighbors). Neighborhoods can have a longer range, with connections made to second, third, or further away nearest neighbors. Sites on the edge of the lattice may have fewer neighbors: such edge effects can be avoided by imposing periodic boundary conditions.

  Connections in the lattice are local in nature. Paths between arbitrary pairs of nodes tend to be long, with a large number of intermediate nodes in between. All nodes have the same number of neighbors (possible edge effects excepted): the connectivity distribution is not heterogeneous.

  Regular lattices tend to exhibit cliquish behavior, although it should be pointed out that the two dimensional lattice with the von Neumann neighborhood contains no triangles. (This should be thought of as a deficiency in the triangle definition as a way of capturing cliquishness.)

- **Small World Network**

  The random graph and regular lattice have a long history of use; the small world network is a more recent addition, introduced in 1998 by Watts and Strogatz in a paper that has stimulated a large body of work on complex networks.

  Watts and Strogatz small world networks are generated from a regular one dimensional lattice in which each node is connected to some collection of their nearest neighbors. In their original recipe, each edge is randomly and independently rewired with probability p. If a given edge is to be rewired, one end of the edge is left unaffected and the other end is chosen at random from the set of nodes. Notice that these rewired connections are made without regard to spatial location: we call them "long range" links (see figure 56).

  When p equals zero, the small world network is simply the regular lattice. Connections are entirely local, so path lengths in the network are long, and the network exhibits a high degree of cliquishness. When p approaches one, the small world network becomes a random network. In this case, the network is well-mixed, so the path lengths in the network are short, and cliquishness is low.

The surprising observation of Watts and Strogatz is that the small world network rapidly acquires the well-mixed, short path length, nature of the random graph as p increases. Just a small number of long range connections give a small world network a global feeling, while retaining the clique property of the regular lattice (figure 56).

The small world property has been observed in many real world networks. One of its origins is in social network theory, with the work of Milgram. The idea is that one can link any two people in a population via a surprisingly small number of intermediate friends or acquaintances. This is the basis for the "six degrees of Kevin Bacon" game.

The degree distribution of the small world network is still fairly closely centered on its mean, just like those of the random graph and regular lattice.

There are other algorithms for generating small world networks. Perhaps the simplest modification is for the long range links to be additions to the network (so that, for instance, there is some probability that each pair of nodes has an edge added between them) rather than being obtained by rewiring existing links. In many cases, this change makes little difference, although it turns out to facilitate mathematical analysis of the network and the dynamics.

- **Scale Free Networks**

  The three network types exhibit limited heterogeneity. The scale free network is, in contrast, highly heterogeneous. Barabasi and Albert introduced this type of network, generating it via a dynamic algorithm one node at a time. The key feature of their algorithm is preferential attachment of edges: when a new node is introduced, its contacts are made to existing nodes, but with a preference for attachment to already well connected nodes.

  The network is started with a small number of nodes. These have some initial connectivity distribution. At each step, a single new node is added and is linked to a certain number, m, pre-existing nodes. For each of these m links, the connectivities of the n pre-existing nodes, $k_1, k_2, \ldots, k_n$, are examined. The probability that the link is made to node i is given by $k_i / \sum_j k_j$. This process is repeated until the network has the desired number of nodes.

  This preferential attachment of links leads to a situation in which the rich get richer: the highly connected nodes attract many more links. This leads to a network whose degree distribution is highly heterogeneous. Most nodes have few neighbors, but a few nodes have many neighbors. In fact it can be shown that the degree distribution follows the power law $P(k) = Ak^{-3}$. This distribution has a large variance: in fact, as the network size tends to infinity, this variance diverges.

  We remark that there are many alternative types of scale free network. In what follows, unless we say otherwise, we use the term "scale free network" to mean the Barabasi and Albert scale free network with $P(k) = Ak^{-3}$.

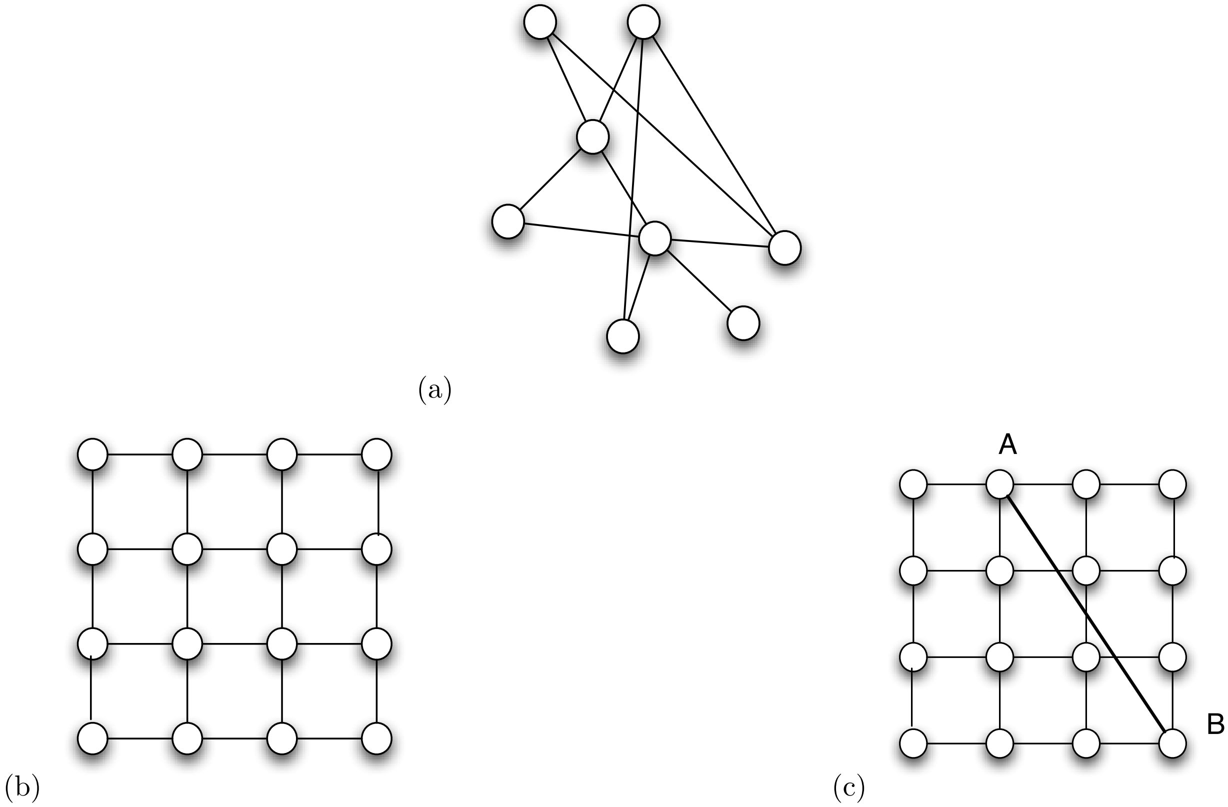

Figure 55: Random graph (a), regular lattice (b) and small-world network (c). Notice how the addition of one extra link in the small-world network reduces the distance between nodes A and B to one.

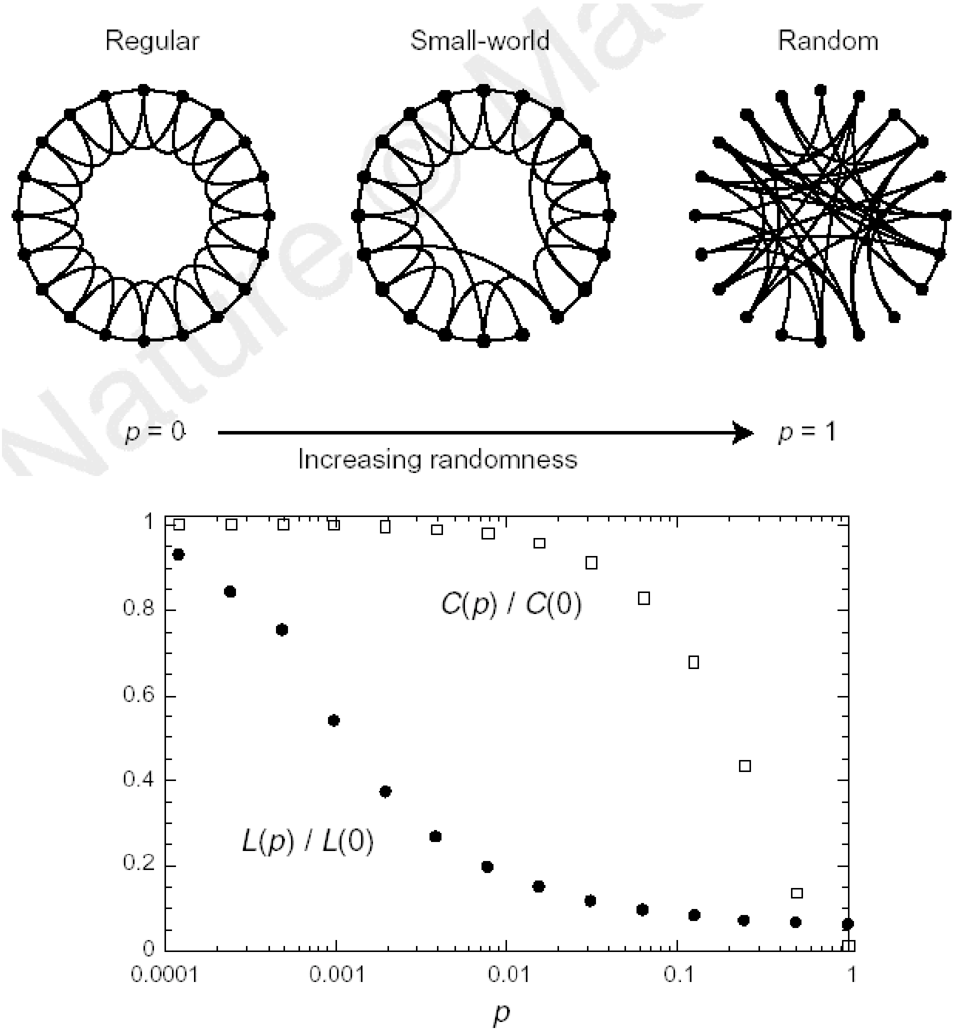

Figure 2 Characteristic path length $L(p)$ and clustering coefficient $C(p)$ for the

Figure 56: Watts and Strogatz small world networks. Upper panel depicts the construction of the network by random rewiring of a regular 1D lattice. Second panel depicts the dependence of the average shortest path length and clustering coefficient on the rewiring probability, p. Notice that the path lengths rapidly fall as p increases but that the clustering coefficient remains large until p becomes comparable in size to one. Figure taken from Watts and Stogatz (1998) Nature **393**, 440-442.

Path lengths tend to be short in this scale free network, its mixing is global in nature and tends to exhibit low levels of cliquishness.

Many real world networks exhibit high degrees of heterogeneity, and so perhaps share this scale free property. Claimed examples include computer networks and sexual partnership networks.

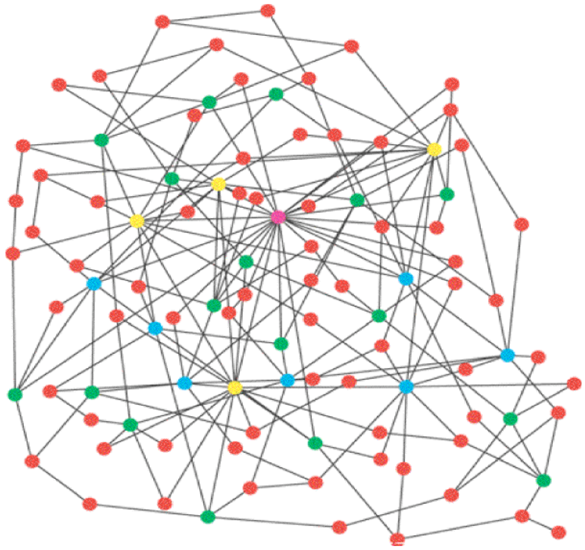

Figure 57: Example scale free network. Nodes are colored according to their connectivity.

15.3.4 Epidemic Dynamics on Networks

The most commonly thought about settings involve the spread of an infection on a fixed network. Typically, the timescale thought about is sufficiently short that births and deaths can be ignored. If we have an SIR infection then this will lead to a single epidemic. If recovery is possible (e.g. SIS or SIRS infection) then an endemic can be established. As an example, we might assume that the probability of transmission along a given edge between an infective and a susceptible in a short interval of length dt is equal to βdt and that the probability of recovery in a similar time interval equals γdt.

Analysis is typically difficult for network models. Most analyses involve making considerable simplifying assumptions, such as ignoring any local structure (such as cliques) or assuming that there are no loops in the network. This second assumption corresponds to treating the network as if it were a tree. One approach has been to employ the multi-group formulation, discarding almost all of the network structure. In the statistical physics literature, such descriptions are referred to as "mean-field" models.

15.3.5 The Basic Reproductive Number

Expressions for the basic reproductive number can be derived using a number of approaches. Using the mean-field approach in a homogeneous setting, the following result is obtained

$$R_0 = \frac{\beta}{\gamma} \langle k \rangle. \tag{252}$$

In a random mixing hetereogeneous setting, we have that

$$R_0 = \frac{\beta}{\gamma} \langle k \rangle \left(1 + \mathrm{CV}^2 \right). \tag{253}$$

This looks quite familiar from our earlier discussion of proportionate mixing: this should not be surprising since the mean-field model is just the multi-group model. As before, heterogeneity increases the value of R_0.

Scale free networks exhibit high levels of heterogeneity, so their value of R_0 is considerably inflated compared to the level that would be predicted using just the average degree. In fact, as the number of nodes increases, the variance diverges and so R_0 will become infinite. This leads to the strange result that infections can always spread on a true scale free network as long as the transmission parameter β is non-zero. There is no $R_0 = 1$ epidemic threshold. Of course, this result only applies to scale free networks of infinite size. Finite size effects mean that the variance is finite, but large: the epidemic threshold is regained, but at very small levels of β.

More detailed analyses borrow ideas from percolation theory and graph theory. It is possible to show that the transmission of infection on a fixed network can be mapped onto a set of problems that has long been studied in percolation theory. Generating functions, which summarize the degree distribution as a single function of a dummy variable, can then be used to derive threshold conditions and equations for the expected size of an outbreak.

The percolation approach describes the probability of transmission along an edge in more detail. Taking the per-edge rate of transmission to be given by β and assuming then the probability of transmission along an edge between an infective and a susceptible over a time interval of length τ is given by $1 - \exp(-\beta\tau)$. This quantity forms the basis of what Newman calls the transmissibility of the infection.

In the setting of a general distribution of infectious periods, described by the pdf $f(\tau)$, and assuming that the rate of transmission is identical between any connected pair of susceptible and infective individuals, Newman defines the transmissibility as

$$T = 1 - \int_{\tau=0}^{\infty} f(\tau) e^{-\beta\tau} \, d\tau. \tag{254}$$

In the case of a fixed duration of infection, this expression simplifies to the one given earlier.

Using generating function methodology, Newman then shows that the basic reproductive number in the proportionate mixing setting is given by

$$R_0 = T\left(\langle k\rangle\left(1+\mathrm{CV}^2\right)-1\right). \tag{255}$$

In the case of a homogeneous degree distribution, this simplifies to $R_0 = T(\langle k\rangle - 1)$. Keeling also obtained a somewhat similar expression, at least for the transmissibility of infection.

The expression for R_0 involves one less than the average number of neighbors. The origin of the minus one is clear: each individual, apart from the initial case, must have acquired infection from another individual in the population. The average number of susceptible neighbors is equal to $\langle k\rangle - 1$.

15.3.6 Impact of Network Structure on the Spread and Control of Infection

Many of the consequences of network structure are clear. Infection can spread rapidly across a network if the distances between individuals are short, otherwise infection must pass through a large number of intermediates in order to traverse the population. Infections spread much more quickly on random graphs than they do on lattices. The small world effect can have a major impact in this regard as it dramatically shortens distances in the network.

The spread of infection is further slowed on the lattice by the clique property: the overlapping neighborhoods of connected pairs of nodes dramatically reduces the potentials for secondary infections.

Heterogeneity in the connectivity of individuals can enhance the spread of infection. The highly connected nodes serve as foci of transmission. Most individuals are poorly connected, so spread may be slow at first but this is overturned once the infection reaches a highly connected person.

The local spread of infection on the lattice means that local control measures are likely to be effective. Unfortunately, the small world property quickly overcomes this: the small number of long range links thwart such local control measures. In the heterogeneous networks, control measures must target highly connected individuals. Random vaccination is futile in the true scale free network, but targeted vaccination is highly effective.

These results have interesting analogs in terms of the robustness and resilience of networks such as power grids or computer networks. For highly heterogeneous networks, the random removal of nodes or edges has little effect compared to what it would have on a random graph. Removal of appropriately targeted nodes, however, quickly degrades a heterogeneous network.

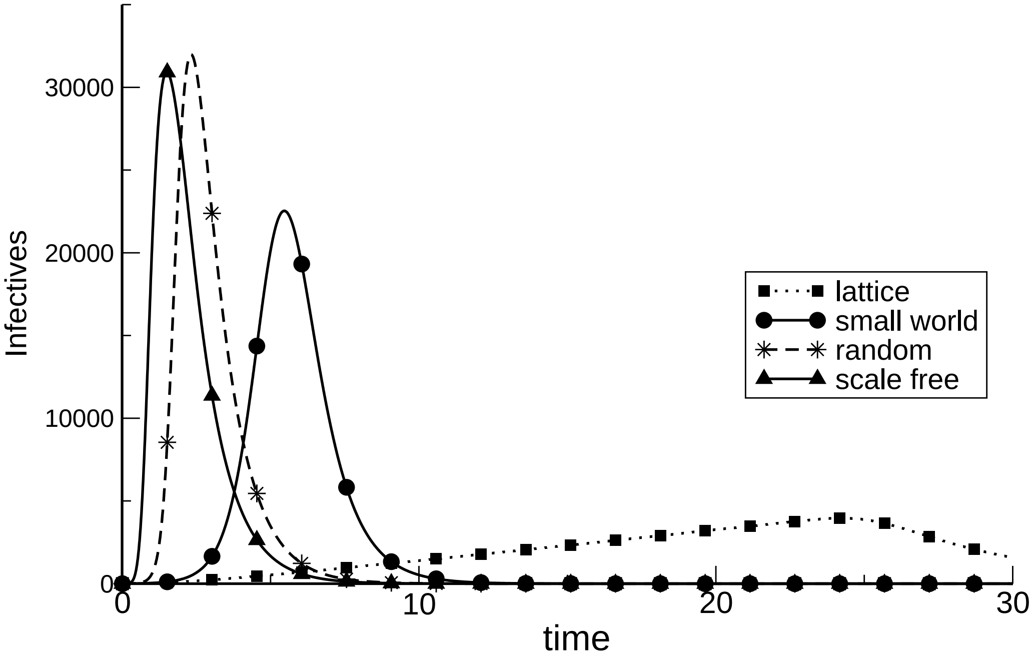

Figure 58: Invasion dynamics on a random graph, a regular lattice, a small world network and a scale-free network. In each case, the parameters β and γ were taken to have the same values and the average number of neighbors was fixed at 8. For the small world network, just 1% of nodes were rewired.

15.3.7 Pair Approximation Models

The complexity of network models often makes their simulation a slow task. It would be nice to have some simple way of incorporating some aspect of the effect of network structure within a simpler modeling framework. This would allow us to reduce our dependence on the simulation approach. Pair models are an attempt in this direction.

In the standard deterministic SIR model, the transmission of infection depends on the numbers of susceptibles and infectives. In the network setting, transmission not only depends on these numbers but also on the configuration of the susceptibles and infectives: we need a pair in which a susceptible and infective are connected. Pair models depict equations for how the numbers of S, I, S-I pairs, S-S pairs, I-I pairs (and so on) change over time.

It turns out that the equations for the changes in the numbers of pairs involve the numbers of triples. For instance, Keeling derived the following equations for the SIR model

$$\begin{aligned}
\dot{[SS]} &= -2\beta[SSI] && (256)\\
\dot{[SI]} &= \beta\left([SSI]-[ISI]-[SI]\right)-\gamma[SI] && (257)\\
\dot{[SR]} &= -\beta[RSI]+\gamma[SI] && (258)\\
\dot{[II]} &= 2\beta\left([ISI]+[SI]\right)-2\gamma[II] && (259)\\
\dot{[IR]} &= \beta[RSI]+\gamma\left([II]-[IR]\right). && (260)
\end{aligned}$$

Also see figure 59. The problem is that this goes on and on (for instance, the equations for triples involve quartets).

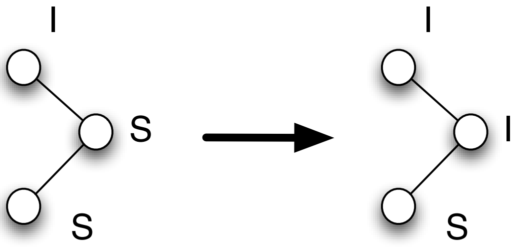

Figure 59: The configuration of triples can impact infection dynamics and be altered by infections. In this case, the infection event changes an S-I pair and an S-S pair into an I-I pair and S-I pair. Notice that the central susceptible would be more likely to acquire infection if they had two infective neighbors rather than one.

The idea is to truncate this set of equations at some point. The pair approximation makes some assumptions that can be used to approximately relate the numbers of triples to the numbers of pairs. One example involves the frequency of triangles in the network:

$$[ABC] \approx \eta \frac{[AB][BC]}{[B]} \left((1-\phi) + \phi \frac{N}{n} \frac{[AC]}{[A][C]} \right). \tag{261}$$

This then gives a closed set of equations involving pairs. The end result is typically a set of ODEs that captures some aspects of the network structure. The number of ODEs is greater than in the simple deterministic SIR framework, but the set is much easier to work with than the full network model.

This approach is reminiscent of the one we used for studying the stochastic model, when we obtained a small closed set of moment equations.

Keeling showed that this approach works well. The important aspect is that the locations of susceptibles and infectives are correlated in the early part of the epidemic, retarding the spread of infection. Keeling also derived an expression for R_0 that accounts for local spatial structure in terms of the frequency of triangles in the network.

15.3.8 Dynamic Networks

If the network is changing over time, then there is often little we can do other than simulate the model.

One important setting in which dynamic networks are employed is for sexually transmitted infections. The network is dynamic as partnerships are formed and broken. The network often looks quite different to those considered above. Most individuals are monogamous, so a snapshot of the network consists mainly of disconnected pairs and singles (figure 60). Some individuals may be involved in more than one simultaneous partnership: we call this phenomenon concurrency.

Transmission can occur within a pair. If there is no concurrency in the network, further transmission requires the pair to break up and form new partnerships. This may be a slow process. Concurrency clearly facilitates the spread of infection by providing links between pairs.

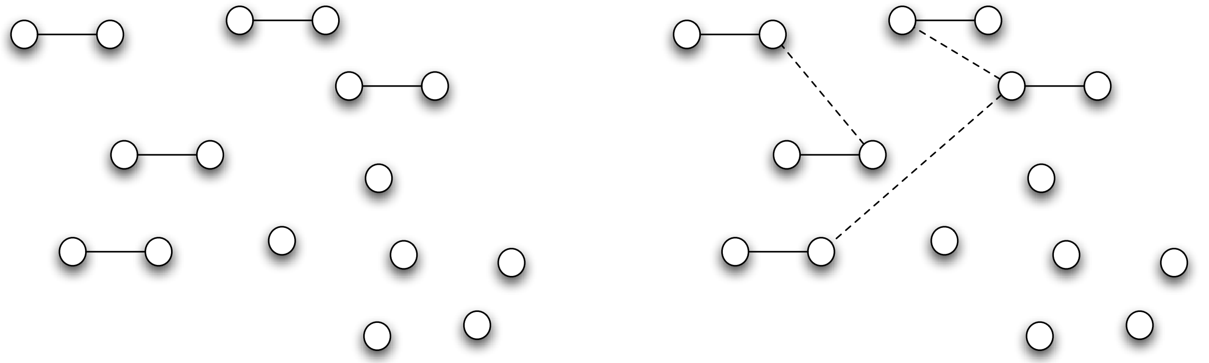

Figure 60: Sexual networks and concurrency. In the left hand panel, the network consists solely of disconnected pairs and singles. In the right hand panel, some individuals are involved in concurrent partnerships.