University of Padova

Lecture Notes
of
Life Data Epidemiology

Collection of the lectures notes of professors Chiara Poletto and Sandro Meloni.

Authors:
Alice Pagano
Andrea Nicolai

Saturday 14th August, 2021

Contents

Part I

Meloni's Lectures

1

Basics Definitions and Compartmental Models

All models are wrong, but some of them are useful.

– Unknown author

Lecture 2. *Friday 2nd October, 2020. Compiled: Saturday 14th August, 2021.*

Models in science have two different roles: **understanding** what happens and **predict** will happen. Models can be of two types: simple and more complex ones. In the simplest ones we just consider the minimal number of parameters and events involved: this indeed allows to understand what are the main mechanisms of a phenomenon.

In this course, we are going to start with very simple models in which we assume that there is no structure behind in the population. Obviously this is not accurate, but allows us to understand at a first glance some underlying mechanisms. Then, we are going to consider social structures and introduce contact network models. We will also take into account interactions among different populations and exploit data to understand how members move from one population to another. Finally, we are going to introduce the so called "Agent Based" models, for a quick overview on them.

1.1 Compartmental models

We now introduce the **compartmental models**. These are fundamental since the most of epidemiological theories are based on them. In reality, however, there are different levels of understanding how diseases can diffuse: we can consider the disease only at a biological level, or at simpler one. Note as it is practically impossible to insert all the details of a process in a single model. We therefore need to summarize all the biological processes in few **parameters** which describe, on average, what we can see inside the population. This is the same principle behind the statistical mechanics in which we look for large scale (macroscopical) effects.

Let us consider a population of individuals and try to characterize it. Note as we have not made any assumption on the individuals and relationships between them. We now introduce three different **compartments**, denoted with **S** (that stands for *Susceptible*), **I** (*Infected*), **R** (*Recovered*), and want to label people according to the stage of their disease, as seen in Fig. 1.1. However, one should note that there can be also transitions from one state to another one, according to some rates that describe the **dynamic**. For instance, in Fig. 1.1 these are β and μ.

This approximation, on the other hand, is quite strong: by keeping the rates fixed we are assuming that the process underlying the spreading of the disease is **Markovian**. In reality, we do not see exponential distributions (i.e. decays), but some other distributions such as the *Gamma* one. This last point, however, will be

discussed during the course when we will deal with **"non-Markovian"** epidemics. The interpretation we may give to β is the *"per contact" infectious rate*, in this way we only need to count the number of contacts. Different models can be introduced according to the type of the disease: for instance **SI**, **SIR**, **SIS**, **SEIR** and so forth.

One should note that medical status is actually different from infectious status. In the latter we do not care about medical status of the person, but only about the disease and how the immune system reacts against it.

As an example, for the **SEIR** compartmental model, we have four main stages of the disease: starting from a healthy state (**S***usceptible*), the individual can contract the disease (**E***xposed*) and then, only after some time, becomes infectious (**I***nfectious*) until he recovers (**R***ecovered*) (Fig. 1.2). The most important thing to keep in mind is that these compartments are not the same ones of the medical status, since they keep into account different parameters despite the disease is the same one.

Now, let us introduce the **Basic Reproductive Number $\mathbf{R_0}$** (pr. *"R naught"*) which is a measure of the infection in the population. If we wanted to empirically determine it: we put one guy inside a group for an arbitrarily long time period and, at the end, we count the number of secondary cases we have. This is the main idea behind the computation of R_0. This parameter therefore determines whether a disease will spread or not:

$$\begin{cases} R_0 < 1 \\ R_0 = 1 \\ R_0 > 1 \end{cases} \tag{1.1}$$

Let us consider the plot of Fig. 1.3, we have a sort of **second order phase transition** at the point $R_0 = 1$. Note that R_0 for the SARS is higher than the one of COVID-19. However, since we did not observe any outbreak of that disease, it means that the Basic Reproductive number is not the only relevant parameter to be taken into account in the models. In order to compute R_0 we assume that the population is totally susceptible. This is however valid only at the very early stages, later on, we must consider both epidemiological and demographical aspects. The conclusion is the following: R_0 may vary from one population to another.

Since we are doing a **coarse-graining** of the dynamics, this number represents the average of all possible different distributions. A *wrong* argument is to think that similar R_0's lead to similar outbreaks. The distribution of infections can be quite heterogeneous: the mean could be quite representative only if we are dealing with homogeneous populations, that is not the case for real networks. For instance, let us consider the plot in Fig. 1.4. We see that SARS was heterogeneous, while Spanish Flu was a more homogeneous one. COVID-19 is most likely somewhere in the middle.

Let us now introduce the **Effective Reproductive Number R(t)**, which is the same of R_0 but varying wrt time. Hence, it is the average number of secondary cases

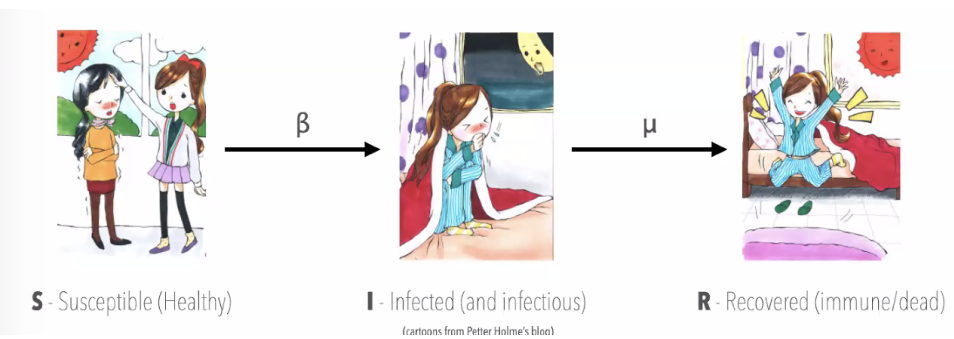


Figure 1.1: Classification of infected population in three different stages of the disease.

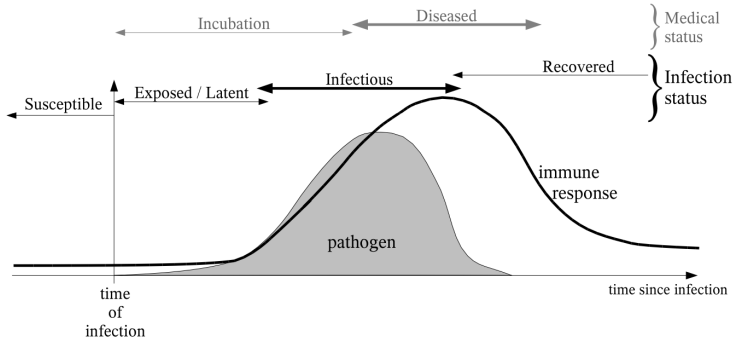


Figure 1.2: A sketch of the time-line of infection, showing the dynamics of the pathogen (grey area) and the host immune response (black line) with the labeling for the various infection classes: **S**usceptible, **E**xposed, **I**nfectious, and **R**ecovered. Note that the period when symptoms are experienced (medical status) is not necessarily correlated with any particular class of epidemiological models.

that a single case produces in a population at time t.

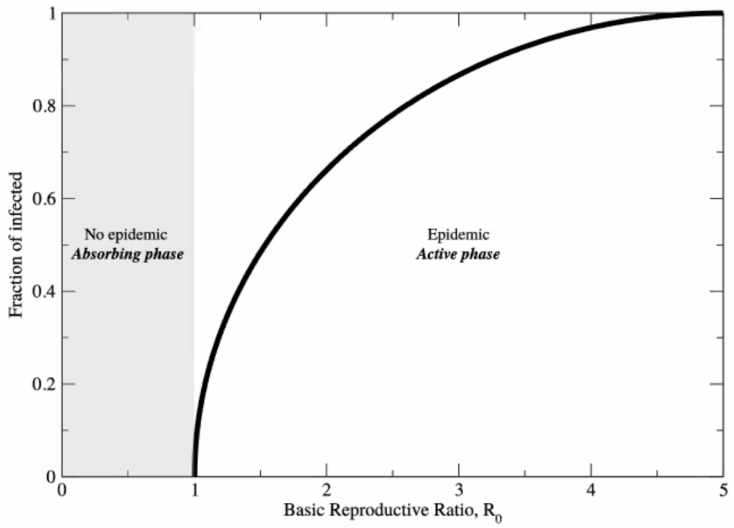


Figure 1.3: Fraction of infected vs basic Reproductive Ratio, R_0.

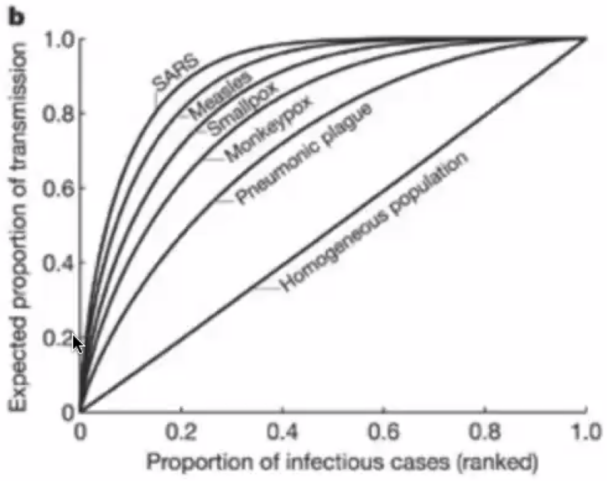


Figure 1.4: Figure from: Lloyd-Smith et al. Nature 438, 355–359 (2005).

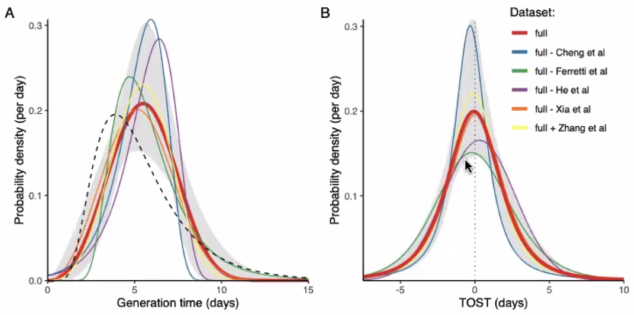


Figure 1.5: Figure from: Ferretti et al. `https://www.medrxiv.org/content/10.1101/2020.09.04.20188516v1`

Other important quantities we may want to introduce are:

- **Infectious period**: average period for a person to be infectious is computed either as $\tau = \frac{1}{\mu}$ or $\tau = \frac{1}{(\alpha+\mu)}$ where the presence of α depends on the model (α: average duration of "exposed time" stay).
- **Incubation period**: period of time between infection to occurrence of symptoms
- **Generation time**: time for an infected person to generate a second infection
- **Serial interval**: time between the onset of symptoms for a person and the onset of symptoms for another second infected person
- **TOST**: time between the onset of symptoms to an infection

A problem in predicting a possible outbreak of a disease is that TOST in many cases can be negative (see Fig. 1.5 for more details).

Lecture 3. *Thursday 8^{th} October, 2020. Compiled: Saturday 14^{th} August, 2021.*

1.2 Basic models

In this lecture we are going to introduce some of the basic models we will use for the entire course. The first assumption we make is that we are in **well-mixed populations**, or in other words *homogeneous mixing*. Mathematically, it is what is called **mean field approximation**.

In the well-mixed population assumptions, it holds that that:

- all individuals are **equivalent**, hence every one has the same probability of being infected;
- every individual has the **same number of contacts** $N-1$, or on average $\langle k \rangle$;
- we are in a **closed population**. That is to say that the sum of the density distribution of the individuals is equal to 1, hence we have no deaths or births. In practice, we are assuming that our time scale is so little that we can consider the population constant.

1.2.1 SI model

The simplest model one can think of is the **SI** (**S***usceptible* **I***nfected*). In this model one can get the infection and, once we have got, we cannot recover, that is to say we stay infected forever.

The **transition diagram** that describes this model is the following:

$$S + I \xrightarrow{\beta} I + I \tag{1.2}$$

where β is the *"per contact" infection rate* and dictates the speed of the spreading. We can write down the **equations** that can be solved exactly:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= -\beta \langle k \rangle si \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \beta \langle k \rangle si \end{aligned} \tag{1.3}$$

where $\langle k \rangle$ represents the average contacts, while i stands for the fraction of infected people in the entire population ($i = I/N$), and s is the fraction of susceptible people in the population ($s = S/N$). Note as prefactor $\langle k \rangle$ is constant, therefore sometimes it can be "absorbed" inside β. The product si is the probability of having a contact between an infected and a susceptible, and βsi is the probability of having a contact between an infected and a susceptible which in turns leads to an infection.

One of the most important quantity we may want to introduce in our lexicon is the so called **prevalence** $i = \frac{I}{N}$, that is another way to define the density of infected people wrt the entire population.

In order to solve it analytically, we recall that our population is closed. Therefore $s + i = 1$, and it follows that we only have one equation to be solved since $s = 1 - i$. We have that:

$$\frac{\mathrm{d}i}{\mathrm{d}t} = \beta i(1-i) \quad \rightarrow \frac{1}{\beta i(1-i)}\,\mathrm{d}i = \mathrm{d}t \quad \rightarrow \frac{1}{\beta(1-i)}\,\mathrm{d}i + \frac{1}{\beta i}\,\mathrm{d}i = \mathrm{d}t$$

Integrating both sides:

$$-\log|1-i| + \log|i| = \beta(t+C) \rightarrow \frac{i}{1-i} = e^{\beta(t+C)} = Ae^{\beta t}$$

with $A = i_0/(1-i_0)$. The result is:

$$i(t) = \frac{i_0 e^{\beta t}}{1 - i_0 + i_0 e^{\beta t}} \tag{1.4}$$

which is a sigmoid function (Fig. 1.6) that always saturates at 1. One should note that after the first part, where the growth is actually exponential [1], then at a certain point the slope starts to decrease. The reason for this is that the contribution given by the term si, namely the probability of funding new susceptible people, decrease. Finally, we saturates at 1 after some. As can be clearly seen from Fig. 1.6, it is the value of β that drives the spreading. By increasing it, we obtain a faster exponential growth. This actually was the simplest model one can think of.

Remark. In the course we are going to use capital letter for integer numbers, while small letters refer to densities.

1.2.2 SIS model

Now, let us introduce a slightly more complicated model, that is the **SIS** model, where compartments are **S***usceptible*, **I***nfected*, **S***usceptible*. Transitions now are two:

$$\begin{aligned} S + I &\xrightarrow{\beta} I + I \\ I &\xrightarrow{\mu} S \end{aligned} \tag{1.5}$$

[1] It is the one we have seen in the media for COVID-19.

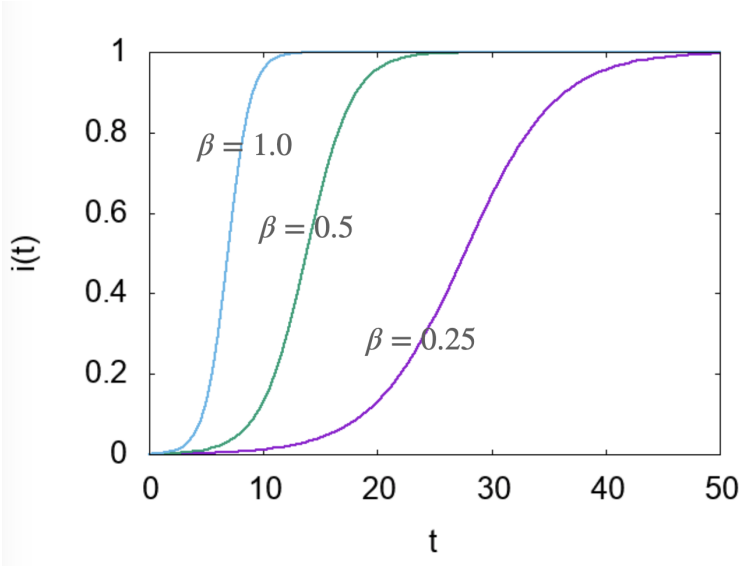


Figure 1.6: Plot of the solution of the SI model for different β.

where the first transition is mediated by I, that is to say we need to encounter another infected to contract the disease, while the second one occurs **spontaneously** according to the rate μ.

This model is used for diseases that do not confer immunity. When we use the expression **endemic state** it means that the disease keeps on circulating in the population for very large times.

The most important feature about this model is that it is the simplest one where **dynamical equilibrium** can be reached. Therefore an individual may recover from the disease, but he does not get immunity. Indeed there are always people infected that can propagate the disease. The μ is the **recovery rate** which determines the *time-scale of the infection*. Dividing β by μ you can **rescale** all the **dynamics**. The **equations** are exactly the same as before, except for a term:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= -\beta \langle k \rangle si + \mu i \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \beta \langle k \rangle si - \mu i \end{aligned} \tag{1.6}$$

and in addition can solved in the very same way we previously did.

Also, the shape of the **solution** is a sigmoid as before:

$$i(t) = i_0 \frac{(\beta - \mu)e^{(\beta-\mu)t}}{(\beta - \mu) + \beta i_0 (e^{(\beta-\mu)t} - 1)} \tag{1.7}$$

Note that we have included $\langle k \rangle$ inside the definition of β: it can be done since both of them are constant. By plotting it, one should note that despite the same form, we do not saturate at 1, but at $\frac{\beta-\mu}{\beta}$. Hence, as we said, we have some sort of **dynamical equilibrium**: the number of new infected is more or less the same of the new recovered people at each moment. The density $i(t)$ will therefore fluctuate around this value $\frac{\beta-\mu}{\beta}$ and, by enlarging μ, we can obtain larger fluctuations (Fig. 1.7).

It can be instructive to study what happens according to this model at the **transient**. At the beginning, one can assume that almost the entire population is composed by susceptible people ($s \sim 1$), while the number of infected is very small ($i \ll 1$). Hence, the differential equations can be rewritten as following:

$$\frac{\mathrm{d}i}{\mathrm{d}t} = \beta \langle k \rangle si - \mu i \sim \beta \langle k \rangle i - \mu i \to i(t) \sim i_0 e^{(\beta \langle k \rangle - \mu)t} \tag{1.8}$$

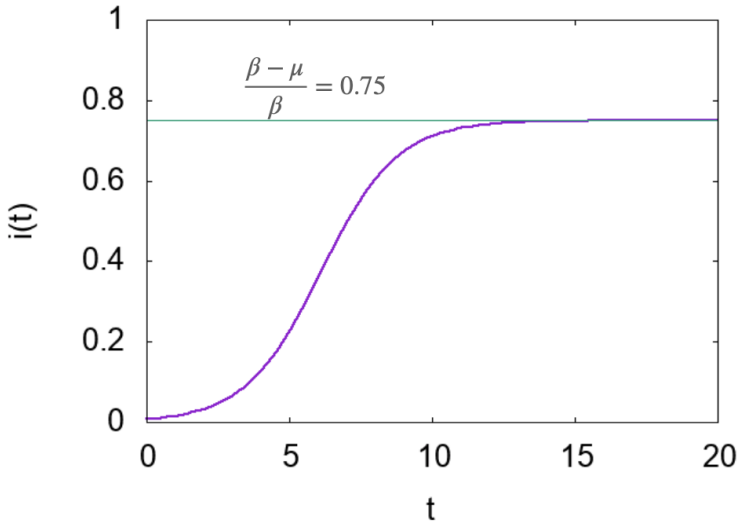


Figure 1.7: Plot of the solution of the SIS model.

One should note that if $\beta \langle k \rangle < \mu$ there is no spreading at this point anymore, while, if $\beta \langle k \rangle > \mu$ the exponent becomes positive and from this follows the exponential growth at the beginning (Fig. 1.8).

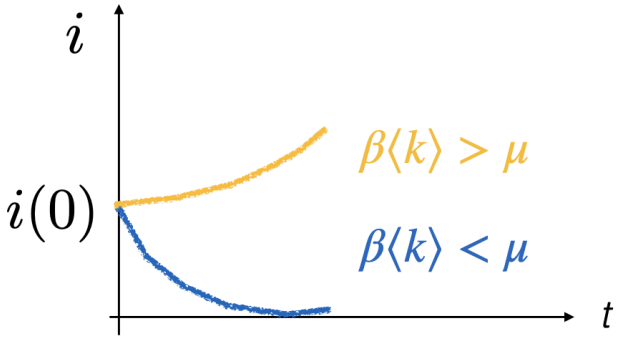


Figure 1.8: Initial transient for the SIS model.

One very important thing is that considering the **steady state** we can have two possible behaviors, according to the limiting values of the incidence:

$$\frac{\mathrm{d}i}{\mathrm{d}t} = 0 \rightarrow \begin{cases} i = 0 & \beta \langle k \rangle < \mu \\ i > 0 & \beta \langle k \rangle > \mu \end{cases}$$

and we have that, in the long run, the incidence is different from zero only when the epidemics has not become extinct in the early phase. More formally:

$$i > 0 \iff \beta > \beta_c = \frac{\mu}{\langle k \rangle} \tag{1.9}$$

where β_c is known as the **epidemic threshold**. This tells us whether the disease is going to spread.

In addition the epidemic threshold is the minimum value of the infection probability for which the disease survives. This is what in physics is called a **second order phase transition** (Fig. 1.9). In this case the **critical exponents** are the same of the Ising model, since they belong to the same class of universality. β_c is one of the most important quantities we are going to study.

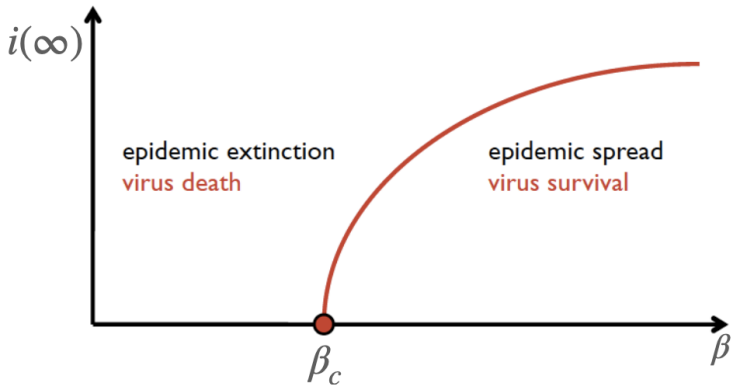


Figure 1.9: Epidemic diagram.

One may ask what is the relation between R_0 and the epidemic threshold. Obviously, they are strongly correlated. We actually say that given a **critical value**, below it we have no spreading, while above we have a fraction of infected people.

We refer to these two different cases as it follows: if our $\beta < \beta_c$, then we end up in the so called **epidemic extinction** and the virus, in the long run, will not be present any more. On the other hand, if $\beta > \beta_c$, the virus is going to be present in the population and therefore survives. This is the so called **endemic state**. Behavior around the critical point might be of our interest and can be studied using Statistical mechanics formalism and/or numerical simulations.

The epidemic threshold is given by the condition under which we observe the spreading. Mathematically, given a specific model, its critical version will return the values of the parameters for which $R_0 = 1$. If we are slightly above this threshold, we only need a minimum of infected people and the disease is going to spread. Considering for instance the case of the SIS model, the β_c critical is such that it makes R_0 being equal to 1:

$$\beta_c \; : \qquad R_0 = \frac{\beta_c \langle k \rangle}{\mu} = 1 \tag{1.10}$$

1.2.3 SIR model

We now discuss the so called SIR model, whose compartments are **S***usceptible*, **I***nfected* and **R***ecovered*. The idea behind is the same one of the SIS, but we are now adding a new state which accounts for long lasting immunity (**R**). Hence, once a person has got the disease and has recovered, he obtains a long **immunity**. Recall that, since we assumed that the population is closed, its density is still fixed to 1.

The transitions for this model are:

$$\begin{aligned} S + I &\xrightarrow{\beta} I + I \\ I &\xrightarrow{\mu} R \end{aligned} \tag{1.11}$$

and one should note that we cannot have any endemic state. For large times all individuals will have been infected, and recovered, so the disease will be spreading no more.

The differential equations that describe this model are:

$$\begin{aligned}\frac{\mathrm{d}s}{\mathrm{d}t} &= -\beta \langle k \rangle si \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \overbrace{\beta \langle k \rangle si}^{\text{New infections}} - \overbrace{\mu i}^{\text{Recovery}} \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= \mu i\end{aligned} \tag{1.12}$$

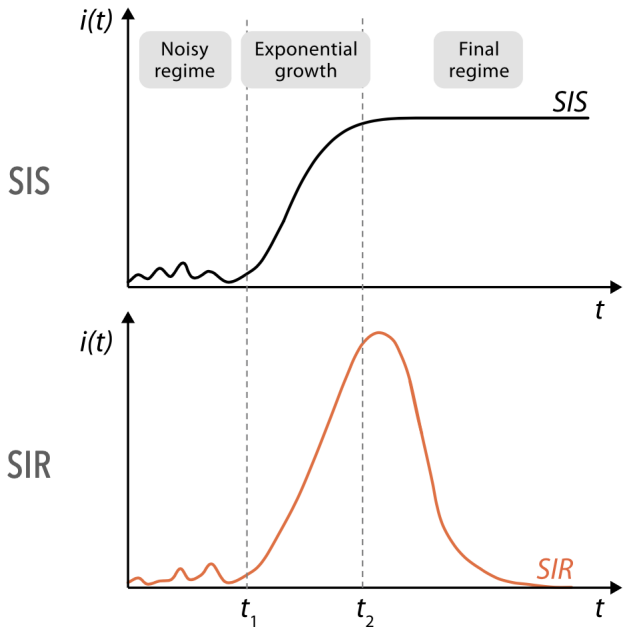


Figure 1.10: Epidemic regimes.

This is actually a good point to introduce the **different regimes** we may encounter during a spreading, which are represented in Fig. 1.10 for the SIS and SIR models.

Initially, at the beginning of each spreading, we see the so called **noisy phase** where numbers are too small to cause a large spreading. Here we can observe only some sort of stochastic fluctuations. In many cases, we can end up without any spreading: this may happen if we assume that some nodes are much more linked than others (the so called *super spreaders*), and we are able to recognize and stop them before they can infect anyone[2]. If it is not the case, the disease starts spreading according to the characteristic **exponential growth**. Later, the slope slows down until we reach the **steady state**: for the SIS the disease keeps circulating among the individuals *(endemic state)*, while for the SIR it disappears *(absorbing state)*.

In order to compute the **epidemic threshold** for the SIR model, the path to follow is the same as before. In particular, we assume that at the starting point $s \sim 1$ and $r \ll 1$, hence:

$$\frac{\mathrm{d}i}{\mathrm{d}t} = \beta \langle k \rangle si - \mu i \sim \beta \langle k \rangle i - \mu i \rightarrow i(t) \sim i_0 e^{(\beta \langle k \rangle - \mu)t} = i_0 e^{\mu(R_0 - 1)t}$$

where we introduced the $R_0 = \frac{\beta \langle k \rangle}{\mu}$. It should be clear now why estimating R_0 is such important: it drives the exponential growth in early phase. Once again we see

[2]the assumption is one of the basis for **heterogeneous** mean field models. We will discuss them later in the course.

that we have no extinction when $\beta > \beta_c$:

$$\beta > \beta_c = \frac{\mu}{\langle k \rangle} \tag{1.13}$$

Since it is difficult to obtain an analytic expression for S and I in this SIR model unless we make some assumptions, we want to study what is the behavior for large times ($t = \infty$). Dividing $\mathrm{d\,s}_{\mathrm{d}t \mathrm{by} \frac{\mathrm{d}r}{\mathrm{d}t}, \mathrm{and consideringthat}} \mathrm{i}(\infty) \to 0$ one obtains that:

$$\frac{\mathrm{d}s}{\mathrm{d}r} = \frac{-\beta \langle k \rangle s}{\mu}$$

Assuming moreover that $r_0 = 0$ and integrating the above expression wrt r, we obtain:

$$s(t) = s_0 e^{-r(t)\frac{\beta \langle k \rangle}{\mu}}$$

As already said, we cannot find an analytical solution, but we can study the **behavior for large times** by making some approximations. At $t = \infty$, it holds that $i(\infty) = 0$, thus $s(\infty) = 1 - r(\infty)$ because of the closed population assumption:

$$1 - r(\infty) - s_0 e^{-r(\infty)\overbrace{\frac{\beta \langle k \rangle}{\mu}}^{R_0}} = 0$$

This is a transcendental equation that cannot be solved analytically, but still gives important hints on the behavior of the disease.
One may note that $R_0 = \beta \langle k \rangle / \mu$, and this should make us understand why it is R_0 that drives the exponential growth of the disease, being it proportional to $\beta \langle k \rangle$. Moreover, the initial fraction of susceptible people (s_0) plays a role in shaping the final fraction of recovered. In particular, if $s_0 \ll 1$, the disease cannot spread. This is how **herd immunity** can be obtained.

1.3 Extensions of the SIR model

We want now to modify the SIR to take into account some more features we want to implement our model with.

1.3.1 SIR with Demography

So far we have assumed that the population was totally closed, and so densities always sum up to 1. This is actually unrealistic, so our next step will be to **drop** the **closed population** assumptions: we will now introduce births and deaths. This reasoning is justified from what we observe in real world: considering the demography, we note as every year there are new children that are infected by diseases such as Measles and Chickenpox. Anyway, we do not expect that they will die out over weeks, but still it tells us that newborns increase the populations to the susceptible compartment.

The simplest assumption we can make is: similar to the infectious period, individuals can have a **lifespan**, denoted as $1/\alpha$ ($[\alpha] = \mathrm{year}^{-1}$). Note as in this approximations lifespan is much greater than the infectious period, so deaths are not due to the disease. In this way we assume that α is the death rate, common to all classes. Moreover, α is also the crude birth rate, and in addition we assume that births occur only for susceptible individuals and therefore increase its density.

In order to keep the population constant, we need to assume:

$$\frac{\mathrm{d}s}{\mathrm{d}t} + \frac{\mathrm{d}i}{\mathrm{d}t} + \frac{\mathrm{d}r}{\mathrm{d}t} = 0 \tag{1.14}$$

Our equations become then:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= \alpha - \beta si - \alpha s \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \beta si - \mu i - \alpha i \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= \mu i - \alpha r \end{aligned} \tag{1.15}$$

where the **infectious period** is:

$$\tau = \frac{1}{\alpha + \mu} \tag{1.16}$$

on average, individuals spend less time infected because some of them may die while infected. However, it is a small change compared to before, since lifespan is much greater than the infectious period.

Also, $\mathbf{R_0}$ is reduced due to mortality:

$$R_0 = \beta\tau = \frac{\beta}{\alpha + \mu} \tag{1.17}$$

We want now to study the **equilibrium points** of the dynamic for this model. Assuming:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{\mathrm{d}i}{\mathrm{d}t} = \frac{\mathrm{d}r}{\mathrm{d}t} = 0$$

we want to find the **equilibrium values** s^*, i^* and r^*. It holds that, at equilibrium:

$$\frac{\mathrm{d}i}{\mathrm{d}t} = 0 = \beta si - \mu i - \alpha i \quad \rightarrow \beta s^* i^* - (\mu + \alpha) i^* = 0$$

and, collecting i^*, we obtain the following equation:

$$i^*[\beta s^* - (\mu + \alpha)] = 0 \tag{1.18}$$

which is not differential anymore.

There are two different solutions for this equation: the one for which $i^* = 0$ (**disease free state**) and the one for $s^* = \frac{\alpha+\mu}{\beta} = \frac{1}{R_0}$, which is the **endemic state**. Here, the most important result is that the **SIR** model **with demography** can actually **show** an **endemic state**.

Replacing $s^* = \frac{1}{R_0}$ in $\frac{\mathrm{d}s}{\mathrm{d}t} = \alpha - \beta si - \alpha s$, we obtain:

$$i^* = \frac{\alpha R_0}{\beta}\left(1 - \frac{1}{R_0}\right) = \frac{\alpha}{\beta}(R_0 - 1)$$

Finally, the three **equilibrium values** (s^*, i^*, r^*) for the fraction of infected, susceptible and recovered in the endemic state are:

$$(s^*, i^*, r^*) = \left(\frac{1}{R_0}, \frac{\alpha}{\beta}(R_0 - 1), 1 - \frac{1}{R_0} - \frac{\alpha}{\beta}(R_0 - 1)\right) \tag{1.19}$$

Keep in mind that this solution exists only if $R_0 > 1$ and we obtained the equation for $r*$ by reverting the formula $s^* + i^* + r^* = 1$. Moreover, via linear stability analysis, it can be demonstrated that this equilibrium is stable and is reached through damped oscillations.

Lecture 5.
Thursday 15th October, 2020.
Compiled: Saturday 14th August, 2021.

1.3.2 SIRS Model

We now introduce another model, in which we take into account that during the years the **immune system may lose the ability to recognize a known pathogen**. This immunity could have been acquired via either a vaccine, or having recovered from that disease itself. Moreover, there could be the possibility that viruses mutate, as it occurs with the seasonal influenza, and so antibodies are not able to recognize it any more. Hence, let us build a model in which after an individual is recovered, can become again susceptible after a certain period of time.

The SIRS Model allows to interpolate between SIR ($w = 0$) and SIS ($w \to \infty$), where w is the **waning immunity rate**, namely the rate at which we lose our ability to defend ourselves from a certain pathogen. We can end up again into either an absorbing, with no more disease, or endemic state, where it keeps on circulating. The transitions for this model are:

$$\begin{aligned} S + I &\xrightarrow{\beta} I + I \\ I &\xrightarrow{\mu} R \\ R &\xrightarrow{w} S \end{aligned} \tag{1.20}$$

In particular, the differential equations that describe the model are:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= \alpha + wr - \beta si - \alpha s \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \beta si - \mu i - \alpha i \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= \mu i - wr - \alpha r \end{aligned} \tag{1.21}$$

In this case, the **endemic state** can be found by setting the derivatives equal to zero.

One may note that the transition $R \to S$ does not affect the I, so it holds that for the **infectious period**:

$$\tau = \frac{1}{\alpha + \mu} \tag{1.22}$$

while the $\mathbf{R_0}$ factor is:

$$R_0 == \beta\tau = \frac{\beta}{\alpha + \mu} \tag{1.23}$$

In addition, the equilibrium values s^*, i^* and r^* can be easily obtained using the same arguments as of the SIR model with demography.

1.3.3 SEIR Model

In reality people do not become instantaneously infectious, but there is a **latent period** which is the time between infection and becoming infectious. Indeed, the pathogen replication takes time, i.e. viral load is too low to be able to transmit the infection. This argument leads us to introduce the **S***usceptible*, **E***xposed*, **I***nfected*, **R***ecovered* model, where the class **E** takes into account that a person has already contracted the disease, hence is not susceptible anymore, but is not able to spread it yet.

Moreover, this period can be extremely heterogeneous depending on the disease: it can take from few hours to years, such as the case for HIV or, even longer, TBC. In the latter, latent periods might appear to be even longer than an individual's lifespan, with the result that he may have contracted the disease, but the death occurs for other causes before the onset of any symptom.

It is important to remind that the **latent period** is **not the same** of the **incubation period** (see Fig. 1.11). An individual can be infectious before symptoms. For instance, there might be a **pre-syntomatic infection period** as it occurs in the case of COVID-19! This explains, once again, why medical status is different from the infection status.

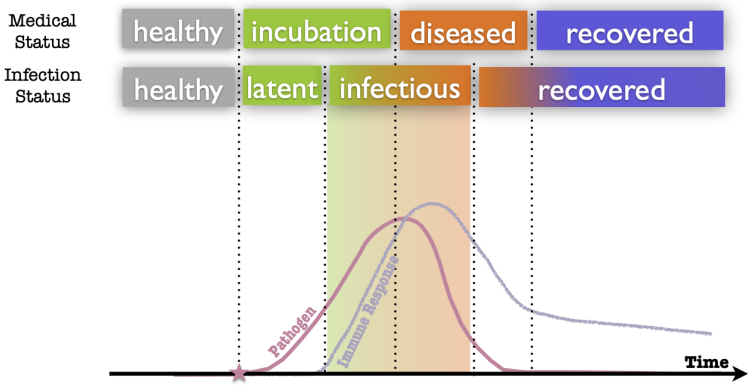


Figure 1.11: Difference between infection status and medical status.

The transition for the **SEIR** model are:

$$\begin{aligned} S + I &\xrightarrow{\beta} I + E \\ E &\xrightarrow{\sigma} I \\ I &\xrightarrow{\mu} R \end{aligned} \tag{1.24}$$

with the equations:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= \alpha - \beta si - \alpha s \\ \frac{\mathrm{d}e}{\mathrm{d}t} &= \beta si - (\alpha + \sigma)e \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \sigma e - (\alpha + \mu)i \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= \mu i - \alpha r \end{aligned} \tag{1.25}$$

Hence, the spreading is delayed due to the time spent in E class.

The **endemic state** is:

$$\begin{aligned} s^* &= \frac{(\alpha+\mu)(\alpha+\sigma)}{\beta\sigma} = \frac{1}{R_0} \\ e^* &= \frac{\alpha(\alpha+\mu)}{\beta\sigma}(R_0 - 1) \\ i^* &= \frac{\alpha}{\beta}(R_0 - 1) \end{aligned} \tag{1.26}$$

For very short latent time ($\sigma \to \infty$) we recover the endemic state of the SIR.

The $\mathbf{R_0}$ factor is:

$$R_0 = \frac{\beta\sigma}{(\alpha+\mu)(\alpha+\sigma)} \tag{1.27}$$

Since latent time is way shorter than demography one, usually $\frac{\sigma}{\sigma+\alpha} \simeq 1$, hence $R_0 = \frac{\beta}{\alpha+\mu}$ as in the SIR with demography.

One may object that, given that the infectious period and R_0 are similar between SEIR and SIR, adding the Exposed class may seem an unnecessary complication. However, if we look at the time evolution, at the **early stages** there is a huge difference between SEIR and SIR model:

$$\begin{aligned} i_{SEIR}(t) &\approx e^{\left(\sqrt{4(R_0-1)\sigma\mu+(\sigma+\mu)^2}-(\sigma+\mu)\right)t/2} \approx i_0 e^{\left(\sqrt{R_0}-1\right)\mu t} \\ i_{SIR}(t) &\approx i_0 e^{(R_0-1)\mu t} \end{aligned} \tag{1.28}$$

Even if the behavior at the steady state is similar, the temporal evolution of the prevalence of SEIR model is actually slower than the one using SIR. This has surely to be taken into account in policy making, given its important implications.

The SEIR can be the starting point for modeling realistic diseases: i.e. Covid-19 (see Fig. 1.12).

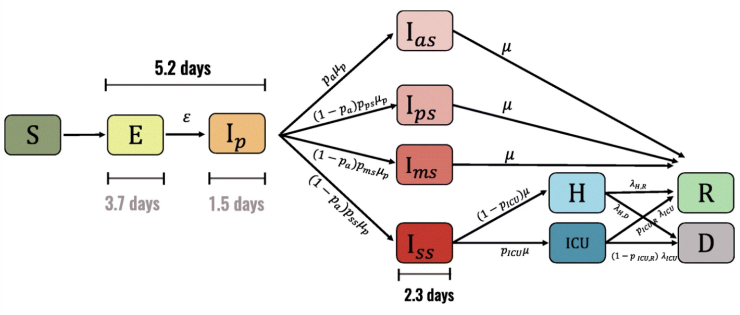


Figure 1.12: Model for Covid-19.

1.4 Summary of compartmental models in well-mixed populations

Let us summarize all the compartmental models in well-mixed populations we have tackled so far:

- we solved the **SI model** analytically, and observed that the growth is the one of a sigmoid:

  $$i(t) = \frac{i_0 e^{\beta t}}{1 - i_0 + i_0 e^{\beta t}}$$

  In the early stages we observe an exponential growth, governed by β, that always saturates at 1;

- in the **SIS model** things starts to change. We have an **endemic** (meta-stable) **state**:

  $$i(\infty) = \frac{\beta - \mu}{\beta}$$

  which is a sort of **dynamical equilibrium** we reach. We can define an **epidemic threshold** β_c according to which epidemics might spread or not:

  $$\beta > \beta_c = \frac{\mu}{\langle k \rangle}$$

  if $\beta > \beta_c$ then no exctinction for the epidemics occurs.

- for the **SIR model** equations cannot be solved analytically. However, we observe **no endemic state** and the **epidemic threshold** is once again:

  $$\beta > \beta_c = \frac{\mu}{\langle k \rangle}$$

- then, in the **SIRS model** we introduced **waning immunity**. This model interpolates between SIR and SIS model. We do observe **endemic state** and the **infectious period** is:
$$\tau = \frac{1}{\alpha + \mu}$$
and moreover:
$$R_0 = \frac{\beta}{\alpha + \mu}$$

- finally, we discussed **SEIR model** in which we included a **latent period**. We have that:
$$R_0 = \frac{\beta\sigma}{(\alpha + \mu)(\alpha + \sigma)}$$
and the Exposed class has the effect to slow down the spreading.

2

Network Science - Basics

2.1 Main definitions

When we talk about Network Science, as the name would suggest, we study **Networks** that, in math, are also known as graph. A **Graph** $G(V, E)$ is simply an object that is composed by a set of **nodes** (vertices) V and a set of **links** (edges) E:

- **nodes** represent the *entities* $V = [\dots, i, j, k, \dots]$ involved in some relationship. These might be entries, people belonging to a social network and so forth. The **number of nodes** is $N = |V|$;
- **links** represent the relationships between entities $E = [\dots, (i, j), (i, k), \dots]$. The **number of links** is $L = |E|$.

Links can be of different kinds and so networks: the basic distinction is between **undirected** and **directed** links. The former ones can be thought as directed edges, but with arrows pointing in both directions, i.e. to both node of the pair. While the second ones do have a direction according to which sense the relationship represented by the link holds.
Another important distinction is between **unweighted** and **weighted** links. The latter ones can be exploited to take into account the possibility that some nodes can be more connected than the others, therefore **weights** follow. In a certain sense, it describes the "strength" of the link between two nodes.

Another important quantity is the **network density** (connectance), that is the fraction of links present normalized to all the possible pairs, and for undirected networks is:

$$d = \frac{2L}{N(N-1)} \tag{2.1}$$

Real networks usually have a very low density, so are **sparse systems** ($L \ll N^2$).

A graph, mathematically, can be represented by the mean of a matrix. It is the so called **adjacency matrix** A of the network, where:

- $a_{ij} = 1$, if a link between nodes i and j exists;
- $a_{ij} = 0$ otherwise.

Many mathematical tools can be used to determine the properties of the system alongside with this matrix, as an example we may want to compute its spectrum in order to obtain the largest eigenvalue. Moreover, one should note that the matrix is symmetrical for undirected and unweighted graphs, i.e. $a_{ij} = a_{ji}$. However, as we already told, real networks are usually sparse, therefore the adjacency matrix will be

filled for large part by zeros. Hence in order to store graphs in a computer efficiently, it is better to use other tools such as adjacency lists, etc.

Two nodes that share a link are defined "connected", "adjacent", "neighbors". In particular, the **neighborhood** of node i is the set of nodes connected to i. The number of neighbors k_i of each node i is what is called the **degree** of the node i. This is the basic measure that we are going to encounter so many times. Once we have defined the degree, the next step is to define what is the **average degree** over the entire network (undirected case):

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i, \qquad \text{or} \qquad \langle k \rangle = \frac{2L}{N} = d(N-1) \tag{2.2}$$

The next definition is the one of **path**, which is a sequence of links which permits to go from node i to node j following edges. Another relevant quantity is the so called **shortest path** between i and j, it is important since it gives us the idea of how big the network is. In particular, the **distance** l_{ij} represents the length of the shortest path between i and j. There could be multiple shortest paths between i and j. The shortest path of maximum length in the network is defined as **diameter**:

$$l_{max} = \max_{ij} l_{ij}$$

Another measure we may want to introduce is the **average** (shortest) **path length**:

$$\langle I \rangle = \frac{\sum_{ij} l_{ij}}{N(N-1)}$$

The network is said to be **connected** if every possible couple of nodes is reachable through a path. Otherwise, each connected part is defined as a **connected component**.

Now, let us see some examples of networks, such as “The Oracle of Bacon”, or the so called “Erdos Number”. The first one is a site that, given the name of an actor, returns the distance between this actor and Kevin Bacon, in unit of costarring movies. This quantity is indeed computed by taking into account the network of actors, linked by common movies in which they starred. The **Erdos Number** instead is the "academical version" for the "Oracle of Bacon": we compute the distance, in terms of collaborations in publications, between a given researcher and the mathematician Paul Erdos through the publications network. The most surprising fact is that, for both examples, the distance is very low! Therefore a question arises: why such short distances in such large networks? In particular, real networks are smaller (i.e. shorter) than one would expect. This is pointed out by the idea of the “Six degrees of separation”. It refers to an experiment that was run in the ’60s by Stanley Milgram: he gave a postcard to a person on the West Coast, with the instructions that it had to be delivered to a place situated in the East Coast. The main goal was to count how many people would receive that postcard, given the rule that it was allowed to give the postcard to acquaintances of the actual possessor. It was discovered that this postcard actually was delivered to 6 people before reaching the destination. This is what is called the **small world phenomena**. When we study the average path length, for some networks we may find that $\langle I \rangle \sim ln(N)$ or, in some cases even $\langle I \rangle \sim ln(ln(N))$. This is extremely important in the spreading of diseases, since we are able to cover the whole system in few steps.

Lecture 6.
Friday 16th October, 2020.
Compiled: Saturday 14th August, 2021.

To summarize what we have seen last lecture: it holds for most real networks that the average path length scales as:

$$\langle l \rangle \approx \ln N$$

the logarithm of the number of nodes in the network, not just with the number of nodes. Or in some cases as $\langle l \rangle \approx \ln(\ln(N))$. But how is it possible? A paper which explains it is "Collective dynamics of small world networks" by Watts and Strogatz. Their idea is what is called the **Watts and Strogatz model**.

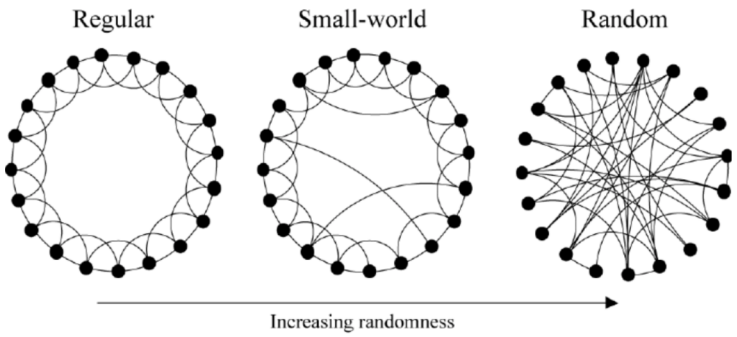


Figure 2.1: Idea of Watts and Strogatz model.

Let us focus on the first regular ring in Fig. 2.1, in which we have that each node is connected with its two nearest neighbours in both sides. The structure as we can see is totally regular. If we want to measure the **longest distance** that we can find in the network:

$$\left\langle l^{circle} \right\rangle \sim \frac{N}{4m}$$

But what actually happens if we rewire only a single link? We therefore want to connect it with another random node in the network as in the picture in the middle "small-world" in Fig. 2.1. It can be seen that, by doing a single rewiring, the size of the system reduces in an incredible way. On the other hand, if we extend this argument and choose a probability p for rewiring (i.e. we increase randomness), what happens is that every time we rewire a connection, the average distance is reduced by a factor 2. Repeating this process several times, we observe a **logarithmic scaling**. Finally, the random network we obtain scales as:

$$\langle l \rangle \sim \log N$$

And it is represented by the random circle in Fig. 2.1.

2.2 Degree distribution over networks

Now the question is how degrees are distributed for different type of networks. Let us consider a **small network**, its degree distribution will be really resembling to the plot on the left of Fig. 2.2. However, now we want to understand how this quantity distributes in **real networks**. In order to build a real network, the first assumption that we can make is building the connections *at random*, so with a probability p. Consequently the degree distribution is one of the kind as in the right of Fig. 2.2.

2.2.1 Erdös and Rényi Model: random graphs

Let us consider the Erdös and Rényi model which represents the evolution of a graph where links between nodes are drawn at random, according to a predefined probability p. Before 1959 (the year of the publication of Erdös and Rényi's paper) people were actually assuming that connections were regular, so no randomness at all. However, since randomness in real world is a deal, thanks to E.R. random connections were taken into account for the first time. In particular, the algorithm for creating such a network is:

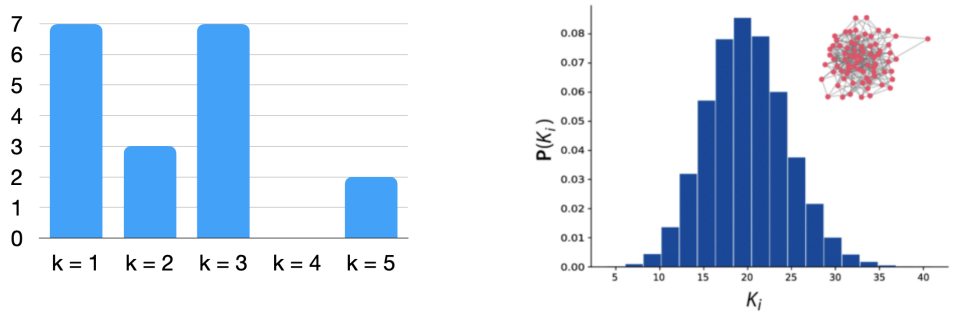


Figure 2.2: **Left:** degree distribution in a small network. **Right:** degree distribution in a network with random connections.

- create an empty graph with N nodes;
- connect each possible couple of nodes with probability p;
- avoid self-loops and multiple edges.

What are the **properties** of this graph? Let us consider a graph $G(N, p)$, where N are the **number of nodes** and p is the **probability of link**. If links are drawn at random with probability p, the probability p_k that a node has k neighbors is given by a binomial distribution:

$$p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{2.3}$$

The **average** and **variance** of such a distribution are:

$$\langle k \rangle = p(N-1), \qquad \sigma_k^2 = p(1-p)(N-1) \tag{2.4}$$

As we can see, the average and the variance scales in the same way with the size of the network (i.e. linearly!).
The problem of this distribution is that it is difficult to be dealt with analytically, specially as N increases, indeed:

$$\frac{\sigma_k}{\langle k \rangle} = \sqrt{\frac{1-p}{p(N-1)}} \overset{N\to\infty}{\longrightarrow} 0$$

which becomes narrower as N becomes larger, therefore some sort of **approximation** needs to be introduced.

Fortunately, since for sparse networks we have $k \ll N$, the binomial distribution $Binom(N, k, p)$ can be approximated by a **Poisson distribution** with parameter $\lambda = pN$. Indeed, given that $\langle k \rangle = p(N-1)$, if we have $k \ll N$ then it implies that $p \ll N$. Hence we can write the following:

$$(1-p)^{N-1-k} \approx e^{(N-1-k)\log(1-\langle k \rangle/(N-1))} \overset{N\to\infty}{\longrightarrow} e^{-\langle k \rangle}$$

and

$$\binom{N-1}{k} \approx \frac{(N-1)^k}{k!}$$

Obtaining the **Poisson distribution** for the *degree* we were looking for:

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \tag{2.5}$$

As before, the average and the variance scale exactly in the same way with the size of the network ($\sim \lambda = Np$). This actually tells us that **all** the **nodes are** more or less **the same**. Indeed when we observe a bounded variance, it means that all the nodes more or less have the same degree. In particular, as p increases the graph undergoes a **transition** from disconnected to fully connected one:

- if $Np < 1$, the graph will almost surely have no connected components of size larger than $O(\log(N))$;
- if $Np = 1$, the graph will almost surely have a giant component of size $O(N^{2/3})$;
- if $Np \to c > 1$, the graph will almost surely have a giant component comprising a large fraction of the nodes;
- if $p < \frac{(1-\varepsilon)\ln N}{N}$, the graph will almost surely contain isolated vertices;
- if $p > \frac{(1-\varepsilon)\ln N}{N}$, the graph will almost surely be connected.

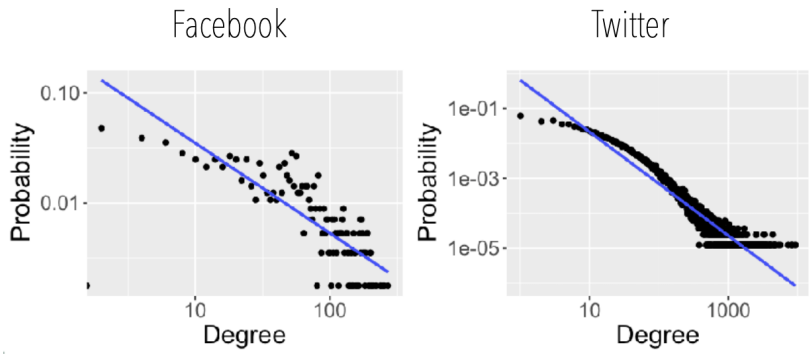


Figure 2.3: Real network of Facebook and Twitter.

2.2.2 Scale-free networks

However, so far we have not discussed how real networks look like, in particular what is their degree distribution. In the last decades we started to have really complex and large networks, whose structure really differs from the structure we usually see for a random network. In Fig. 2.3, as an example, we show two real social networks we know pretty well: Facebook and Twitter. Note that both plots are in log-log scale. Generalizing, we can say that most of the real networks scales in the same way.

We now want to understand how the **degree distribution** looks like. Let us consider Fig. 2.4: black dots follow the Poissonian distribution that we were mentioning before, while the squares follow a power-law $P(k) \sim k^{-\gamma}$, which is **heavy tailed distribution**, in the sense that possibility for large degrees is not null. One should note that the Poissonian distribution is not able to reproduce the heterogeneity we can see in the data, while the power-law is. Hence, in most contexts real networks are **highly heterogeneous** and degrees can span **several orders of magnitude**. In particular, the γ coefficient of the power-law has an important role, since it represents the **slope** of the curve in log-log scale. Since we observe similar structures for different scales, these networks are said to be **scale-free** networks. In most real networks γ has small values, i.e. $\gamma \leq 3$.

Heterogeneity means that almost all nodes have a very low connectivity, way less than a random net. However, the probability of having very large degrees is

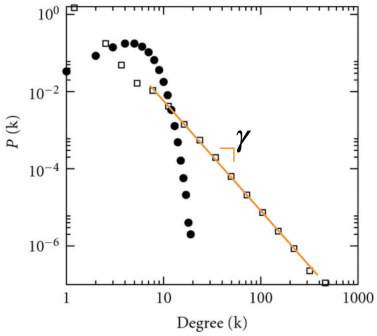


Figure 2.4: Difference between random networks and scale-free networks.

not zero (**hubs**): even for relatively small networks we can observe large hubs. One should take into account that this is something really important for the spreading of diseases: thanks to these large hubs we can see shortcuts for spreading, or the so called **super-spreaders**.

We want now to study the **limiting cases** of these scale-free networks. For instance, we want to see how the **average degree** behaves, or prove that the **largest degree** scales with the size of the network. Let us consider the power-law:

$$P(k) = C_0 k^{-\gamma} \qquad \text{with} \quad C_0 = (\gamma - 1) k_{min}^{\gamma - 1} \tag{2.6}$$

To understand how k_{max} scales with N, we have to study the case where:

$$\int_{k_{max}}^{\infty} P(k)\,\mathrm{d}k = \frac{1}{N} \quad \rightarrow \left(\frac{k_{min}}{k_{max}}\right)^{\gamma-1} = \frac{1}{N}$$

In the lhs formula, we exploited the fact that, on average, a single node drawn uniformly (prob 1/N) will have a degree that is k_{max}. Thus, when:

$$k_{max} = k_{min} N^{\frac{1}{\gamma-1}} \tag{2.7}$$

The last formula is known as the *natural cut-off*. Since in most of networks $\gamma \sim 2-3$, it is easily to understand that k_{max} scales **sub-linearly** with N, but still faster than random graphs, in the sense that the larger the number of nodes the higher is k_max. Hence it is possible to observe highly connected nodes. This is valid for previous plots, such as in Fig. 2.3 as well.

Recalling the definition for the general n^{th} moment of a distribution:

$$\langle k^n \rangle = \int_{k_{min}}^{k_{max}} k^n P(k)\,\mathrm{d}k = \int_{k_{min}}^{k_{max}} C_0 k^{n-\gamma}\,\mathrm{d}k = C_0 \frac{k_{max}^{n-\gamma+1} - k_{min}^{n-\gamma+1}}{n-\gamma+1} \tag{2.8}$$

$$\langle k^n \rangle = C_0 k_{min}^{n-\gamma+1} \frac{N^{\frac{n}{\gamma-1}-1} - 1}{n-\gamma+1} \tag{2.9}$$

Where we used 2.7. We note as it converges only if $\gamma - 1 > n$. This gives an hint on how the **average degree** scales as the size of the network: a very important result. If instead we consider the variance $\sigma^2 = \langle k^2 \rangle - \langle k \rangle^2$, we it holds that:

- if $\gamma < 2$, both $\langle k \rangle$ and $\langle k^2 \rangle$ diverge as $N \to \infty$;
- if $2 < \gamma < 3$, the average degree $\langle k \rangle \to c$ but $\langle k^2 \rangle \to \infty$ as $N \to \infty$, and $\sigma^2 \to \infty$.

Remembering that most real networks have $\gamma \leq 3$, hence the **variance of the degree also diverges**. The result is that we have extremely **heterogeneous networks** and not homogeneous ones. This is indeed coherent to our observations. It has indeed a very strong **implication**: all the models we have been using before, in which we assumed that **all the people** in the population were **equal**, does **not hold** anymore.

2.2.3 Barabási-Albert Model

So far we have discussed about scale-free networks, but actually we have not created a single one yet. Therefore, an **algorithm** to create such network we can rely on, is the **Barabasi-Albert model**. This topic is discussed in a paper that is the second, chronologically speaking, that gave birth to modern Network Science.

The **idea** behind this paper is extremely simple: once some real networks had been analyzed they assumed that the degree distribution $P(K) \sim k^{-3}$, in order to create a model to reproduce the behaviors observed. Moreover, their model was based on the concept of **growing** for random networks. We start with a small number of nodes, named **clique**, and, at each time-step, a new node enters the network and connects with pre-existing nodes but according to a **preferential attachment**. Therefore, at each step the network grows in size.

The principle on which **preferential attachment** is based on is a very simple concept: *rich gets richer*. That is to say: the more connected a node is, the more likely it is for it to receive new links. The probability for a node i to attract a new link at time t, is proportional to its degree k_i at time t:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{2.10}$$

If we speak about **influencers**, having them a lot of followers, the probability for them to increase their connections is very high. Actually, this idea is not even new, and it is something already known. Indeed this model is just a modification of the *Price model*: if we published a paper and more than someone has found it interesting, it will be more likely for it to receive much more attention in the future.

Specifically for this model, we are drawing links at random, according to some probability that indeed is not uniform. Let us briefly summarize the**main steps** of the algorithm:

- we start with a clique of m_0 nodes;
- at each time step t, we add a new node to the network;
- we create m (i.e. $m = 2$) links between the new node and the existing ones according to the preferential attachment (remember to update the connection probability after each link);
- repeat until the desired size N is reached.

In particular, let us consider Fig. 2.5. We start with a small number of nodes connected via some links. At the *first time step* we add a new node, and then we need to draw connections to the other nodes. Let us assume that every time we add a node, we are adding two links. First, we need to compute the set of probabilities of connecting to each node and, at the first time-step, is equal for all the nodes. Then we pick up one node at random and we draw the link. The following step is to update the set probabilities for each node, according to their degree. We see that the node on the left has got an higher probability of getting new connections since the last node inserted has linked to it. Then, we iterate this procedure by introducing a new node and draw connections following the same procedure, until we end up with a total number of N nodes.

This algorithm is indeed able to create networks with some **interesting properties**. Indeed we can approximate the **degree distribution** as:

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \sim k^{-3}$$

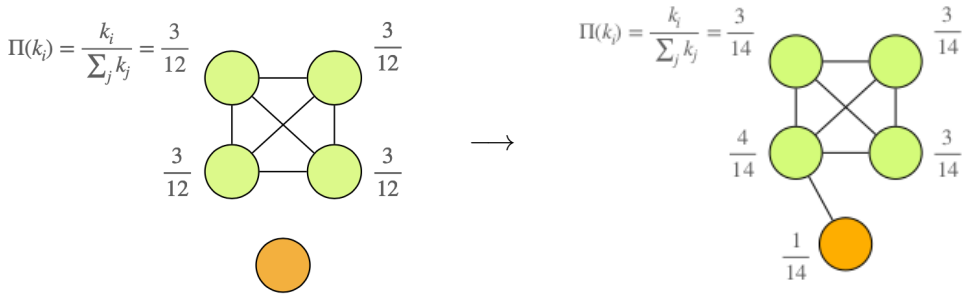


Figure 2.5: Example of Barabási-Albert algorithm.

where m is the number of links we are adding at each step. Note that m is a parameter that is related the minimal degree of the network. However, this approximation is valid for **large** k.
An important result is that $\gamma = 3$ and it is **independent** of m and m_0. Hence, the **maximum degree** of the network scales as $k_{max} \sim N^{1/2}$. Moreover, it holds that $\langle k \rangle \to c$, but $\langle k^2 \rangle \to \infty$ with N, as we have seen before. Finally, the **average length** of the network is:

$$\langle l \rangle \sim \frac{\ln(N)}{\ln(\ln(N))}$$

which tells us that the small-world property holds as well.

3

Epidemic Spreading on Networks

Now it is time to drop the assumption of the well-mixed population, and start taking into account **contact networks**. In other words we are considering that **individuals can be connected in different ways** one another. The main idea is that:

- **all** individuals are **equivalent**;
- we remove the assumption that all individuals have the same number of contacts and we assume that each node **do not interact at random**. This reflects the reality, since we usually have more contacts with some people (friends, family, colleagues...) rather than others. The fact that we may have repeated contacts with someone else has strong effects on the dynamics: we are somehow constraining the way how the disease will spread.

3.1 SIS model in a network

Let us try to build a general model for a general network, without making any assumption on the latter. In order to do that, we start by introducing the equations of SIS model for a generic network.

The first step is to define a **binary variable** for each node i: $\sigma_i(t)$. This variable can only take two values:

- $\sigma_i(t) = 0$, if the individual is **susceptible**;
- $\sigma_i(t) = 1$, if the individual is **infected**.

As one can easily see, this variable describes the state of a generic i-th node at time t. Defining another variable $\rho(i,t)$:

$$\rho(i,t) \equiv \text{Prob}[\sigma_i(t) = 1]$$

which represents the **probability** of that node i is infected at time t. Using this formalism, we can recall the general equation for the SIS in a network:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(i,t) = \overset{\text{Recovery}}{-\mu\rho(i,t)} + \beta \sum_j A_{ij} \overset{\text{Infection}}{\text{Prob}[\sigma_i(t) = 0, \sigma_j(t) = 1]} \tag{3.1}$$

The most problematic part is to compute the two nodes infection probability (in green). Since we are in a network, the probability of being infected depends on my neighbours: the (i,j) infection probability depends on the status of all the other neighbors l of j and i and so forth. Therefore we would have to follow the entire

chain of connections, but this would turn out to be a problem: we cannot obtain a closed form for this expression, since it actually depends on the probabilities of all its neighbors. In turn, they would depend on their neighbors probability and so and so forth.

We want to stress one more time that if we want to predict what is going to happen in the system, we would need to consider the entire network and the time evolution for all the nodes. This approach however is **feasible** only for **small graphs** (i.e. 4/5 nodes) and **few compartments**.

This argument reminds us that we may need some sort of an **approximation**: indeed we need to **cut down** this **probability chain**. That is to say that, at some point, we require a closure of our equations, by the mean of approximation: we are not going to take into account the entire structure of the network. At some point we will take the **average**, and after that we will be able to solve the problem. In physics this kind of arguments are called **mean-field approximations**. Since we are not able to solve many body problems, at a certain point we will consider a **random field** which **acts on the entire system** and we will consider its average effects on the system.

Tailoring this procedure to our specific problem, we are substituting in some way the probability $\text{Prob}[\sigma_i(t) = 0, \sigma_j(t) = 1]$ with some average probability. Obviously, depending on the assumption we are making for this approximation, we will obtain different results.

There are actually many different types of approximations based on different features:

- **Network structure**:
  - **Homogeneous** mean-field (all the nodes are equal);
  - **Heterogeneous** mean-field;
- **Coarsening level**:
  - **Degree-based** mean-field theories (DBMF) in which we assume that all the nodes of the same degree are equal;
  - **Individual-based** mean-field theories (IBMF) in which we assume that all the nodes are different and that we will take individual connections between individuals;
- **Where to cut the chain**:
  - **Individual** level;
  - **Pair** approximations;
  - **Triangles**, etc...;

3.1.1 Homogeneous Networks

Let us start by taking the simplest approximation: we assume **homogeneous network**, **DBMF** and we cut the chain at an **individual** level.

It means that we are considering networks where **nodes degree** is **bounded**, hence:

- we have that $k_i \simeq \langle k \rangle$;
- we have also that the standard deviation is bounded $\frac{\sigma_k}{\langle k \rangle} = \sqrt{\frac{1-p}{p(N-1)}} \overset{N\to\infty}{\longrightarrow} 0$.

All the **nodes** can **be assumed to be equal**, so their position on the network does not matter anymore. This implies the **spatial homogeneity** it holds that: $\rho(i,t) \equiv \rho(t)$.

In addition, cutting at the individual level means that the two terms of the **joint probability** of one being infected and the other one being susceptible $\text{Prob}[\sigma_i(t) = 0, \sigma_j(t) = 1]$ are **statistically independent**. This implies that the joint probability can be factorized as follows:

$$\text{Prob}[\sigma_i(t) = 0, \sigma_j(t) = 1] \qquad \rightarrow \qquad \text{Prob}[\sigma_i(t) = 0] \cdot \text{Prob}[\sigma_j(t) = 1]$$

But now we recall that:

$$\rho(t) = \text{Prob}[\sigma(t) = 1]$$

is the density of infected at time t. Hence, putting everything together, we derive the equation:

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\mu\rho + \beta \sum_j A_{ij}(1-\rho)\rho \qquad \rightarrow \qquad \frac{\mathrm{d}\rho}{\mathrm{d}t} = -\mu\rho + \beta(1-\rho)\rho \sum_j A_{ij}$$

Actually, this last term is the degree of a generic node. Recalling that we have assumed that all nodes are equal, the degree of a generic node is therefore the average degree of the network:

$$\sum_j A_{ij} = k_i \simeq \langle k \rangle \tag{3.2}$$

and by replacing it, we can obtain the same expression that we derived before for SIS model in a well-mixed population:

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \beta \langle k \rangle (1-\rho)\rho - \mu\rho \tag{3.3}$$

This is a very important result and one should recall that is the same we previously found in 1.8. One should keep in mind that now we are considering all the **nodes statistically independent** and we are back again to exactly the same result of well-mixed population. The only **difference** is that when we were considering well-mixed population, we assumed that the **probabilities** were *exactly* **statistically independent**. Now, this is just an **approximation**.

Obviously, all the results derived for SIS model in well-mixed populations are still valid, for instance the epidemic threshold.

Recap. Let us summarize what we have seen at the end of this lecture. We moved from well-mixed populations to contact networks, so we added more complexity in order to make the model is more realistic. We also derived the equations for SIS dynamics on a generic network and then considered its adjacency matrix. Since for us was impossible to write down a closed equation for this model, given the expression for the infection joint probability that involves two nodes, we were not able to compute exactly the probability for a single node of being infected (ρ_i). It would take into account the probability of three nodes i, j, k at the same time. This is actually unfeasible for all the models and all the possible graphs: it has been done in the literature up to only 4/5 nodes. Hence we end to somehow approximate this probability, in order to cut this infinite chain to a certain value. This is exactly why we introduce mean-field approximation: in this way we take into account the effects of all terms on a specific quantity, not individually, but on average therefore reducing the complexity of our problem. We are switching from a many body problem to a one body problem. The simplest approximation we have seen is the one of homogeneous network in which all the nodes are equal, used on SIS model. According to this

argument, for each node there is the same probability of getting infected, so we can approximate the probabilities to be statistically independent. After, we derived all the equations. Their solutions were the same as the ones we had found for well-mixed population. However, in that case the solutions found were *exact*, while now are the result of an approximation.

Lecture 7.
Thursday 22nd October, 2020.
Compiled: Saturday 14th August, 2021.

3.1.2 Heterogeneous Networks

Now, we want to understand what is the effect of **heterogeneity** in the spread of the disease. That is to say that we drop the following assumption $k_i \sim \langle k \rangle$: all **nodes are not equal** any more.

Let us consider now the **heterogeneous mean-field approximation**. Let us use a **DBMF model** and let us cut the chain at an **individual level**. This last assumption means that we consider the probability for a single individual to get the infection. Let us follow the thread of paper *"Epidemic Spreading in Scale-Free Networks"*, written by Pastor-Satorras and Vespignani. It actually provides a **SIS model on scale-free networks**. The main idea behind this paper is the following. Since nodes are not equal anymore, *the probability of getting the infection strongly depends on their position (i.e. degree) in the network.* Authors' intuition is that **nodes** with **the same degree behave in the same way**. In order to do that, we need to divide the network in **degree classes**: that is to say we group together all the nodes with the same degree.

In order to write down the equations, we need to consider the number of compartments we have and introduce a density for each of them:

$$s_k = \frac{S_k}{N_k}, \qquad \rho_k = \frac{I_k}{N_k}$$

where s_k and ρ_k are the fractions of susceptible and infected nodes of degree k in the network. We have that N_k represents the number of nodes with degree k. As before, we introduced the fractions of susceptible and infected individuals (s_k, ρ_k) in the system, but in this case depending on each degree k. Obviously, the total fraction of ρ and s in the system is:

$$\rho = \sum_k \rho_k = \sum_k \frac{I_k}{N_k} \qquad s = \sum_k s_k = \sum_k \frac{S_k}{N_k}$$

While their average, sticking to the definition, is:

$$\langle \rho \rangle = \sum_k P(k)\rho_k, \qquad \langle s \rangle = \sum_k P(k) s_k \tag{3.4}$$

The **equation** that describes how the **probability of being infected** changes in time for the nodes that belong to the **same degree class**:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_k(t) = -\mu\rho_k(t) + \beta k \,(1 - \rho_k(t))\Theta_k(t) \tag{3.5}$$

where we can distinguish as usual a "recovery" term and an "infection" term. In particular, the probability of a contact between a susceptible individual that has degree k and an infected one is highlighted in green. This product consists in two terms: the probability for a node i to be susceptible $(1 - \rho_k(t))$ and the probability of having contact with an infected $\Theta_k(t)$.

We want now to dwell deeper and explain better this last term. The probability that a generic node with degree k has an infected neighbor can be expressed as:

$$\Theta_k(t) = \sum_{k'} P(k'|k)\rho_{k'} \tag{3.6}$$

where we sum over all the possible degree classes k'. In this way we expect to obtain the probability of connecting with any one of them, multiplied by the probability for that specific node to be infected. Note, however, that we are making no assumption about the function $P(k'|k)$, which may change according k. In principle, it could be anything, in the sense that it strongly depends on the structure of the network. However, in order to simplify the problem and derive some results, there are cases where we can make some assumptions on the structure of the latter.

Picking a node at random, the probability to be connected to a node of degree k' given the node degree we start from is k, is the following:

$$P(k'|k) = \frac{k'P(k')}{\sum_{k'} k'P(k')} = \frac{k'P(k')}{\sum_k kP(k)} = \frac{k'P(k')}{\langle k \rangle} \tag{3.7}$$

Note that $P(k')$ is the generic probability of getting a connection at random, times k', which is the number of connections that point toward a node of degree k'. Finally we normalize over all possible degrees of the network [1]. What we obtain is the probability that a generic node in the network is linked to k'. Note as $P(k'|k)$ does not depend on k.

After replacing this last result in 3.6:

$$\Theta_k(t) = \frac{\sum_{k'} k'P(k')\rho_{k'}(t)}{\langle k \rangle} = \Theta(t)$$

Let us take a look closer to the different terms. In the numerator: there is the product between the probability that a link, randomly picked, points to a node of degree k', times the probability of being infected, Finally, we then we sum over all the possible degrees. On the other hand the expression in denominator is related only to the structure of the network. In addition, one should note that $\Theta_k(t)$ **does not depend on** k anymore. Since we are just picking up at random it should be the same for all the nodes.

The method that we are going to exploit to **solve** the differential equation $\frac{\mathrm{d}}{\mathrm{d}t}\rho_k(t)$ is similar to the ones previously used in other models. The first assumption is to be in the **steady state**:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho_k(t) = 0 \qquad \rightarrow \qquad \rho_k = \frac{\beta k \Theta}{\mu + \beta k \Theta}$$

The next step is then to substitute the expression for ρ_k, obtained thanks to Θ:

$$\Theta_k(t) = \frac{\sum_{k'} k'P(k')\rho_{k'}(t)}{\langle k \rangle} = \Theta(t) \qquad \rightarrow \qquad \Theta = \frac{1}{\langle k \rangle} \sum_k \frac{k^2 P(k) \beta \Theta}{\mu + \beta k \Theta}$$

This is the **self consistent equation** for Θ.

However, in order to solve this last equation, we need some workaround. First of all one should note, as what happens in statistical mechanics, this expression has different solutions depending on the value of Θ:

- the **trivial solution** $\Theta = 0$, that of course is not in our interest;
- the **non trivial solution**. We can rewrite the self consistent equation as follows:

$$\Theta = \frac{1}{\langle k \rangle} \sum_k \frac{k^2 P(k) \beta \Theta}{\mu + \beta k \Theta} = f(\Theta)$$

[1] One should keep in mind that $\sum_k kP(k) = \sum_{k'} k'P(k')$.

Hence, the solutions are the values for which it holds $\Theta \equiv f(\Theta)$. These, geometrically, are the interceptions between the line Θ and the function $f(\Theta)$ and have to be found graphically (or using computational algorithms).

Since Θ is a probability, it holds that $0 < \Theta \leq 1$. This means that, it is required for a non trivial solution to exist, the slope of $f(\Theta)$ must be greater than 1. Mathematically, it means that:

$$\frac{\mathrm{d}}{\mathrm{d}\Theta}\left[\frac{1}{\langle k\rangle}\sum_k \frac{k^2 P(k)\beta\Theta}{\mu+\beta k\Theta}\right]_{\Theta=0} \geq 1$$

that leads to the following condition:

$$\frac{\beta}{\mu\langle k\rangle}\sum_k k^2 P(k) \geq 1 \qquad \rightarrow \qquad \frac{\beta\langle k^2\rangle}{\mu\langle k\rangle} \geq 1 \tag{3.8}$$

which is the **condition** for the **existence** of an **endemic state**. Since the network has become more complex, also the structure for the condition of the endemic state acquires in complexity. Indeed, for the **epidemic threshold**:

$$\frac{\beta\langle k^2\rangle}{\mu\langle k\rangle} = 1 \qquad \rightarrow \qquad \beta_c = \frac{\mu\langle k\rangle}{\langle k^2\rangle} \tag{3.9}$$

which is pretty similar to the one previously found, but also includes a term that increases its complexity.

The first check one can make is to verify whether this last result holds also in the case of homogeneous networks. For such networks $\langle k^2\rangle = \langle k\rangle^2$, therefore:

$$\beta_c = \frac{\mu\langle k\rangle}{\langle k^2\rangle} = \frac{\mu}{\langle k\rangle}$$

which is exactly the expression we previously found.

Recalling what we were discussing last lectures, in **scale-free networks** with $2 < \gamma \leq 3$, we have $\langle k\rangle \to c$ and $\langle k^2\rangle \to \infty$ as $N \to \infty$. As the network becomes larger also its variance increases, that is:

$$\beta_c = \frac{\mu\langle k\rangle}{\langle k^2\rangle} \to 0$$

hence the **epidemic threshold vanishes** for $N \to \infty$. This is a quite important result: **if** our **network is large enough every disease will spread no matter its infectivity** (see Fig. 3.1). The converse is still valid: if we have disease with a very low infection rate in a small part of the network, it will not disappear if the network is large enough[2]! That is to say we **always** find ourselves in an **endemic state**, while the threshold becomes very small. These results are actually valid for the most real epidemic models, given the networks are large enough.

Obviously, **real networks** are not infinite: therefore we need some **finite-size corrections**. For example, we may want to derive an expression for epidemic threshold when the size of the system does not diverge.

Let us consider the degree distribution for **scale-free networks**: since the degree cannot go to infinity, it is convenient to introduce an **exponential cut-off** at some point. For instance, let us consider the air transportation network: we see that until a certain point a certain trend is followed, but then the slope of the curve starts to change and resembles to an exponential. This implies that we cannot have an infinite number of connections: the line starts out as a power law and then ends up

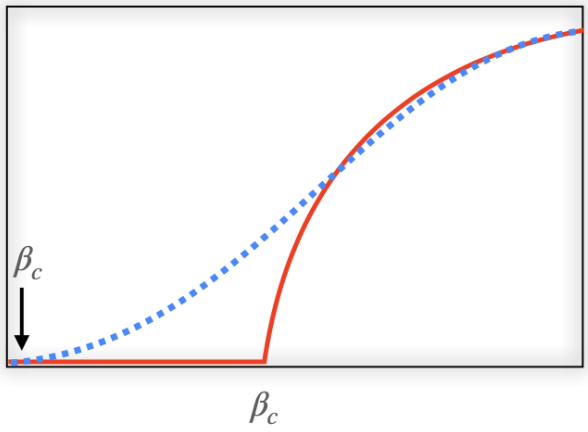


Figure 3.1: In scale-free networks (and many heavy-tailed distributions) the epidemic threshold vanishes in the thermodynamic limit.

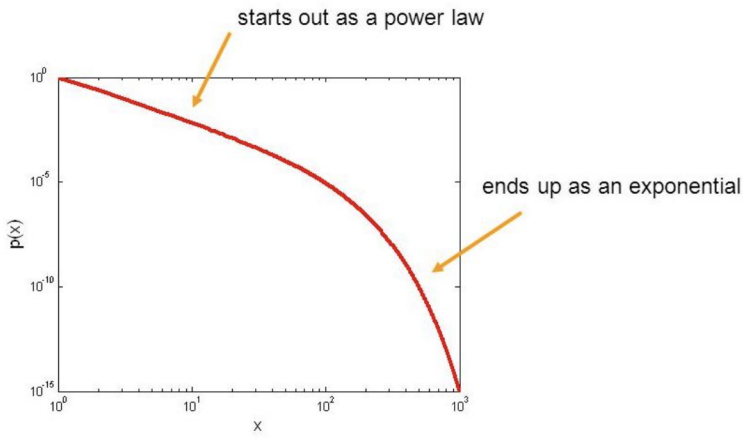


Figure 3.2: Power-law with an exponential cut-off.

introducing some sort of exponential cut-off. The behavior is similar to the one in Fig. 3.2.

We introduce our considerations into our model by adding an exponential term:

$$P(K) \sim k^{-\gamma} e^{-k/k_c} \tag{3.10}$$

where k_c is a **characteristic degree**. At some point, the term we just added will become the dominant term and what happens is that, for large k_c and $2 < \gamma < 3$, the epidemic threshold can be approximated as:

$$\beta_c \simeq \left(\frac{\mu k_c}{k_{min}} \right)^{\gamma - 3} \tag{3.11}$$

we are not going to prove the computations. However, in the lab, we will compare the epidemic thresholds for a random and for a scale-free networks in order to see how they differ. This was the last consideration about the study of the SIS model in a network.

[2]Physically, we refer to this as taking the thermodynamic limit.

3.2 SIR model in a network

3.2.1 Degree-based mean-field theories (DBMF)

The same as before equations can be derived for the SIR model under the assumption of **heterogeneous mean-field**. The main difference is that we need **one more equation** to take into account also the compartment related to **recovered** individuals. Their densities are $\rho_k^S(t)$, $\rho_k^I(t)$ and $\rho_k^R(t)$, and it holds that $\rho_\infty^R = \lim_{t\to\infty} \sum_k P(k)\rho_k^R(t)$. Equations take the form:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\rho_k^I(t) &= -\mu\rho_k^I(t) + \beta k \rho_k^S(t)\Gamma_k(t) \\ \frac{\mathrm{d}}{\mathrm{d}t}\rho_k^R(t) &= \mu\rho_k^I(t) \end{aligned} \tag{3.12}$$

with $\rho_k^S(t) = 1 - \rho_k^I(t) - \rho_k^R(t)$ and where:

$$\Gamma_k(t) = \sum_{k'} \frac{k'-1}{k'} P(k'|k)\rho_{k'} \tag{3.13}$$

is the probability of a contact with an infected node, and plays exactly the same role of Θ before. Actually it represents the link from which the infection arrived to that node, however we will not show how to derive this expression beside one small consideration: the $\frac{k'-1}{k}$ term that is the main difference from the SIS model. It is present due to the fact that we cannot infect a node that has already transmitted us the disease: either because it has already recovered or because it is still infected. In this way we are taking into account that the disease is coming "from one side", therefore for us is forbidden to spread the infection towards that specific direction: recovered (or already infected) individuals cannot be infected twice.

The **epidemic threshold for random networks** results:

$$\beta_c = \frac{\mu \langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \tag{3.14}$$

and the important thing to notice is that $\beta_c^{SIS} \neq \beta_c^{SIR}$. This is the first time so far that the **epidemic thresholds** for these two models **differ**!

3.2.2 Individual-based mean-field theories (IBMF)

Up to now we were assuming that all the nodes with the same degree were equal. Now, since we are going to study the **individual based mean-field** theories, we will not consider a specific instance of the network, but an average over all the possible networks we can obtain **given** that **degree distribution**. That is to say, that under the **Heterogenous Mean-Field framework** we are solving the epidemics problem for an **ensemble of networks** whose common feature is the degree distribution $P(k)$[3].

In the degree based approach we previously assumed that all the nodes with the same degree to be equal. We were therefore analyzing not a specific instance of networks, but its *average*. This is actually what in physics we refer as **annealed networks**. On the opposite, we call **quenched networks** when we consider a *particular realization* of one network. The idea is really simple: instead of considering the average, we consider a particular instance network. This is the main difference between a degree based (i.e. annealed networks) or an individual based approach (i.e. quenched networks).

[3] the so called "ensemble" of networks!

Let us write down the equations for the **quenched mean-field**. We are going to introduce a **discrete time** framework in order to make equation simpler. However, nothing prevents us to use differential equations, where time is a continuous variable.

Let us consider $\rho_i(t)$, that is the probability for node i of being infected at time t. The total fraction of infected individuals is given by $\rho(t) = \sum_i \rho_i(t)$.
At the following time-step, the probability of being infected at time $t+1$ is:

$$\rho_i(t+1) = \rho_i(t)(1-\mu) + (1-\rho_i(t))q_i(t) \tag{3.15}$$

which is the sum of the probability of being infected and not get cured (green term) and the probability of being susceptible multiplied by the probability of contracting the disease (yellow term).
We now need an expression for $q_i(t)$, that is the **probability** for node i to **be infected by, at least, one neighbour**. The basic idea for doing this is:

$$q_i(t) = 1 - \prod_{j=1}^{N} [1 - \beta A_{ij}\rho_j(t)] \tag{3.16}$$

Let us consider Fig. 3.3, in green we have susceptible nodes, which include node i itself, and in red its infected neighbours. The probability of getting infected, at least, by a generic node j is:

$$\beta A_{ij}\rho_j(t) \tag{3.17}$$

Its complementary to 1 is the probability of NOT get the infection by node j.

$$[1 - \beta A_{ij}\rho_j(t)] \tag{3.18}$$

Repeating this argument for all neighbors that are actually infected, we can obtain the probability of NOT contracting the disease from ANY neighbor, namely:

$$\prod_{j=1}^{N} [1 - \beta A_{ij}\rho_j(t)] \tag{3.19}$$

Again, we previously introduced $q_i(t)$ as the **probability of getting infected by at least one neighbor**. Finally, the probability of getting infected is the *complementary* to one of the probability of not getting infected by *any* neighbor:

$$q_i(t) = 1 - \prod_{j=1}^{N} [1 - \beta A_{ij}\rho_j(t)] \tag{3.20}$$

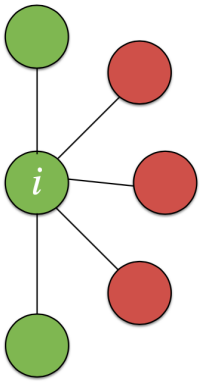


Figure 3.3: In green susceptible nodes, while in red the infected neighbours.

Note as the system of $(\rho_i(t+1))$ equations can be solved numerically by iteration. This results to be precise for the entire epidemic diagram, and faster than numerical simulations: there is no need of averages and reproduces individual nodes probabilities. Indeed, in this framework we will obtain two equations for each of the nodes: we have 2^N equations, where N is the size of the system.

Remark. One should have noted that this last approach differs from the degree based mean field theories by the fact that now we are including adjacency matrix A_{ij}, while before we took only the average.

The equation (3.21) therefore takes the form:

$$\rho_i(t+1) = \rho_i(t)(1-\mu) + (1-\rho_i(t)) \left(1 - \prod_{j=1}^{N} [1 - \beta A_{ij}\rho_j(t)] \right) \tag{3.21}$$

We can also **solve analytically** the system at the **steady state** in order to estimate the **epidemic threshold**. Assuming that we find ourselves in the steady state:

$$\lim_{t\to\infty} \rho_i(t) = \rho_i^* \quad \rightarrow \quad \rho_i(t+1) = \rho_i(t) = \rho_i^*$$

it follows that:

$$\mu\rho_i^* = (1-\rho_i^*)q_i^* \quad \rightarrow \quad q_i^* = 1 - \prod_{j=1}^{N} \left[1 - \beta A_{ij}\rho_j^*\right] \tag{3.22}$$

Now, if we think about what happens when we are in **proximity of the epidemic threshold** (*epidemic onset*), it happens that ρ_i^* can be assumed to be small for all the nodes $\rho_i^* = \varepsilon_i^* \ll 1$. Therefore, the product in q_i^* can be approximated by a sum:

$$q_i^* = 1 - \prod_{j=1}^{N} \left[1 - \beta A_{ij}\varepsilon_j^*\right] \simeq \beta \sum_{j=1}^{N} A_{ij}\varepsilon_j^* \tag{3.23}$$

Substituting what we have just found in the lhs of 3.22 we obtain:

$$\mu\varepsilon_i^* = \beta(1-\varepsilon_i^*) \sum_{j=1}^{N} A_{ij}\varepsilon_j^* \tag{3.24}$$

that is a linear system where the interaction is given by the adjacency matrix:

$$\mu\varepsilon_i^* = \beta \sum_{j=1}^{N} A_{ij}\varepsilon_j^* - \beta\varepsilon_i^* \cancel{\sum_{j=1}^{N} A_{ij}\varepsilon_j^*}$$

Neglecting second order terms, we have that:

$$\frac{\mu}{\beta}\varepsilon_i^* = \sum_{j=1}^{N} A_{ij}\varepsilon_j^* \tag{3.25}$$

This linear system has solution only if $\frac{\mu}{\beta}$ is an **eigenvalue** of the **adjacency matrix** A_{ij}. Here we should understand why last lecture we stated that the spectrum of the adjacency matrix is something we may be interested in. Hence:

$$\beta = \frac{\mu}{\Lambda_i} \tag{3.26}$$

where Λ_i is a generic eigenvalue of the adjacency matrix A_{ij}. However, since we are interested in the **smallest** possible **value** of β for which there exists solution, we need to take the **largest eigenvalue** of the adjacency matrix A:

$$\beta_c = \frac{\mu}{\Lambda_{max}} \tag{3.27}$$

The last one is the **expression** for the **epidemic threshold**, and it is a **general result** that is valid not only while using this approximation, but for a more general framework in a generic network.

3.2.3 DBMF vs IBMF: Epidemic treshold

One may wonder now what is the relation between the two values for the epidemic thresholds we have found for the different mean-field theories, that is DBMF and IBMF. We have found that:

- for **DBMF**:

$$\beta_c^{DBMF} = \frac{\mu \langle k \rangle}{\langle k^2 \rangle}$$

- for **IBMF**:

$$\beta_c^{IBMF} = \frac{\mu}{\Lambda_{max}}$$

For **scale-free networks** $P(k) \sim k^{-\gamma}$ it holds that:

$$\Lambda_{max} \sim \max\left(\sqrt{k_{\max}}, \frac{\langle k^2 \rangle}{\langle k \rangle}\right) \tag{3.28}$$

And in particular:

$$\beta_c \sim \begin{cases} \mu/\sqrt{k_{\max}} & \gamma > 5/2 \\ \mu \langle k \rangle / \langle k^2 \rangle & 2 < \gamma < 5/2 \end{cases} \tag{3.29}$$

We can conclude that **IBMF** is **more accurate** than DBMF. Due to the approximation, indeed, the **DBMF** is **accurate only** in the **proximity of the epidemic threshold**, while IBMF is accurate for the entire epidemic diagram.

Lecture 9.
Thursday 29th October, 2020.
Compiled: Saturday 14th August, 2021.

We recall now one of the most important result of the last lecture: if the network is large enough, for $N \to \infty$, the **epidemic threshold** tends to zero:

$$\beta_c \xrightarrow{N \to \infty} 0 \tag{3.30}$$

Moreover, the epidemic threshold for IBMF depends on the largest eigenvalue of the adjacency matrix Λ_{max}. The last relations which contain β_c for IBMF and DCBMF, can tell us more about the accuracy of the model: **DBMF** is accurate only in the proximity of the epidemic threshold, while **IBMF** is accurate for the entire epidemic diagram. See the figure 3.4.

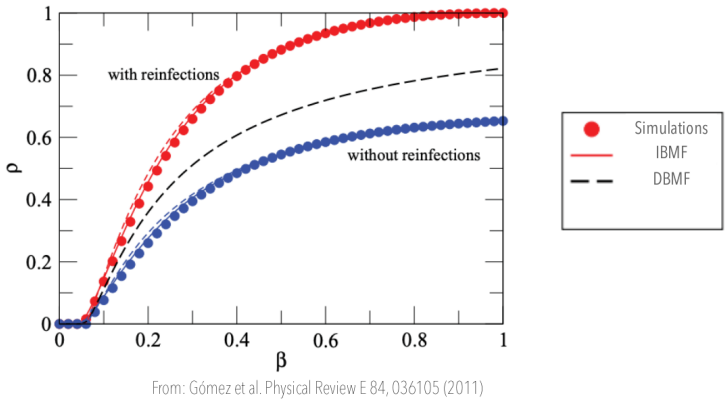


Figure 3.4: The quenched mean field (IBMF) curve follows exactly the simulation, while DBMF the is precise only around the epidemic threshold.

We want now to discuss what are the reasons behind this important result. Since we know there is a strong connection between these two theories, it would be of our interest to derive DBMF from IBMF. Once again, we shall repeat that in **annealed networks** we are not considering a single network but an **average** of all the *possible random networks* that can be generated *from a degree distribution*. Instead, in the **quenched networks** we pick a particular **one**, and we compute the result for *that specific network*. Then, nothing prevents us to run our model on that network and iterate multiple times.

Let us try to characterize the **annealed network**[4], in particular we want to see the *adjacency matrix* looks like. In this case, the **adjacency matrix** is a **weighted matrix**, whose general form is:

$$\bar{A}_{ij} = \frac{k_j P(k_i|k_j)}{NP(k_i)}$$

and for **random networks** becomes:

$$\bar{A}_{ij} = \frac{k_i k_j}{N \langle k \rangle} = \frac{k_i k_j}{2L}$$

where the probability $P(k_i|k_j)$ of picking a random node is k_j and we can define $N_{k'} = NP(k')$. This is the number of trials we have available in order to create this specific connection, over all the possible connections that the network can return.

Then, we have now simply to substitute the last result in the expression 3.23 of q_i:

$$q_i = 1 - \prod_{j=1}^{N} \left[1 - \beta \frac{kP(k'|k)}{N_{k'}} \rho_j \right]$$

And starting from individual nodes and heading towards more general degree classes:

$$\dot{\rho}_k = -\mu\rho_k + (1 - \rho_k) \left[1 - \prod_{k'} \left[1 - \beta \frac{kP(k'|k)}{N_{k'}} \rho_{k'} \right]^{N_{k'}} \right]$$

this is the most general expression that we can obtain for DBMF. The multiplication can be replaced by a sum, only if assuming that $\beta\rho_k \ll 1$:

$$\dot{\rho}_k = -\mu\rho_k + \beta k (1 - \rho_k) \sum_{k'} P(k'|k) \rho_{k'}$$

And recalling that the expression for $\Theta_k = \sum_{k'} P(k'|k)\rho_{k'}$:

$$\dot{\rho}_k = -\mu\rho_k + \beta k (1 - \rho_k) \Theta_k$$

That actually is accurate only under the assumption $\beta\rho_k \ll 1$. But one should note that this is exactly what is depicted in the plot above! Hence, we are able to switch from IBMF to DBMF and, in this way, we can even explain why there is such difference in the accuracy between the two models.

3.2.4 IBMF and pair approximation

Let us make a very brief overview about what means to **cut down** the chain to **pair approximation**. Up to now, all the models we have seen were cut at the *individual level*. Now, instead, let us consider the joint probability of being infected, given that we were susceptible and given that a neighbor of ours was infected. We look for an approximation for this joint probability, and this time we choose to cut the chain at the level of a single link (i, j). We want to see actually how $P\{\sigma_i(t) = 0 \ , \ \sigma_j(t) = 1\}$ changes in the equation for $\dot{\rho}$. Hence we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}t}\rho(i,t) = -\mu\rho(i,t) + \beta \sum_j A_{ij}\rho(j,t) - \beta \sum_j A_{ij} \mathbb{E}[X_i(t) X_j(t)]$$

[4]Guerra, Gardenes - Annealed and Mean-Field formulations of Disease Dynamics on Static and Adaptive Networks, 2010

where $\mathbb{E}[X_i(t)X_j(t)]$ is the two nodes expectation probability to be infected or, in other words, the expectation of both X_i and X_j being infected. This interpretation is valid since $\rho_i = \mathbb{E}[X_i(t)]$, where $X_i(t)$ is a Bernoulli r.v.

However, we need to look for an expression for the $\binom{N}{2}$ equations for $\mathbb{E}[X_i(t)X_j(t)]$, since we have to take into account one expression for each node multiplied by all the possible link we can have in the network. The main idea is:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[X_i(t)X_j(t)] = -2\mu\mathbb{E}[X_i(t)X_j(t)] + \beta\sum_k A_{ik}\mathbb{E}[X_j(t)X_k(t)]+ \tag{3.31}$$

$$+\beta\sum_k A_{jk}\mathbb{E}[X_i(t)X_k(t)] - \beta\sum_k (A_{ij}+A_{jk})\mathbb{E}[X_i(t)X_j(t)X_k(t)] \tag{3.32}$$

Let us analyze the terms on the rhs. The first one is the **recovery term** and needs both of *them to be infected*. On the other hand the second and third terms are the **infection terms**, where either one of the nodes is already infected and the susceptible one gets infected from any other neighbour. The last term has to be put in order to discard the three nodes expectations, and here comes the need for an **approximation**, namely a **closure**.

The most used closures used in the literature are:

$$\mathbb{E}[X_i(t)X_j(t)X_k(t)] = \mathbb{E}[X_i(t)X_j(t)]\mathbb{E}[X_k(t)]$$

where in this case the third term we factorize out is the *mean-field* term.Alternatively:

$$\mathbb{E}[X_i(t)X_j(t)X_k(t)] = \frac{\mathbb{E}[X_i(t)X_j(t)]\mathbb{E}[X_j(t)X_k(t)]}{\mathbb{E}[X_i(t)X_k(t)]}$$

where the second is similar to the first, but now we are considering the two extremes and the probability that j, that is the node in between, is infected.

4

Epidemic spreading on networks: more advanced models

4.1 Non-Markovian Epidemic Spreading

Despite it is difficult to find it discussed in literature, we surely need to take into account that **both** the **infection process** and **recovery process** in reality **DO NOT** have a constant rate.

Up to now, we have assumed the other way around: at each time step the probability of being recovered is always the same, no matter how much we have stayed in the *Infected* compartment. Therefore, our process is **memoryless**. Since the jumps are memoryless, we can characterize our problem as if we were running a Markov chain. Let us recall what is its **main property**: the **jump probability does not depend on time**. Hence, it is not needed to take into account the time we spent there, and so time spent inside each compartment follows an **exponential distribution**. We refer to the average time that we spend there as $\tau = \frac{1}{\mu}$ and the underlying pdf is shown in 4.1:

$$P(x) = \tau e^{-\tau x}$$

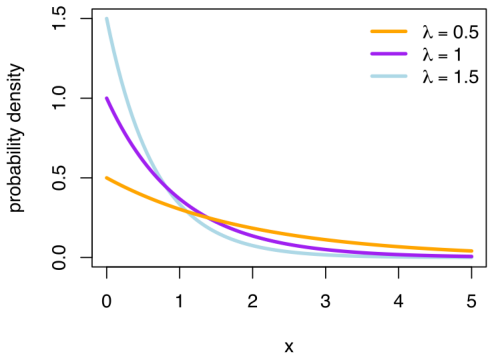
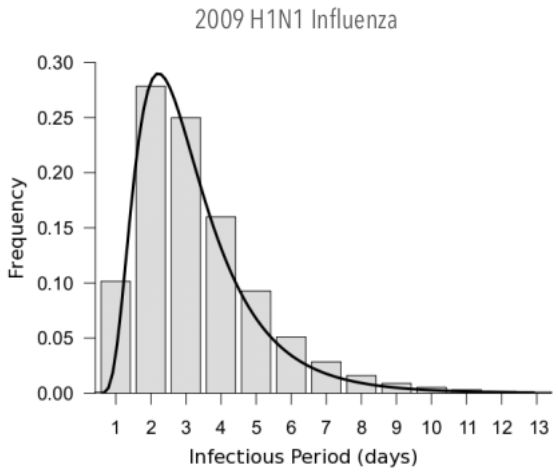


Figure 4.1: Exponential probability density function for different $\lambda = \tau$. Note as the most probable value is when we start our observation, namely at the moment in which $x = 0$.

We want now to discuss what are the implications of these assumptions we have made so far. The most immediate one is that the **mode**, namely the most probable duration of being infected, is *null*. Obviously, the "height" does depend on the mean value, but in any case the most probable moment when we can make the jump is at the beginning of our observation window (see fig 4.1). Obviously, this is something

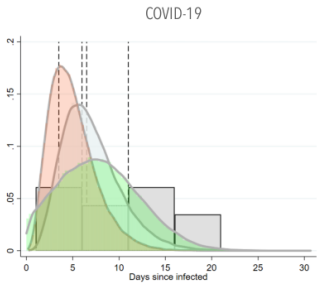
that is **not realistic**. If we got influenza, we do expect to spend at least some time infected, and we do not expect the probability to decrease wrt time. If we wanted to see how infectious periods are distributed in real life, it is indeed something quite different, and do not behave like an exponential. For any disease, we can know (almost) exactly when it starts, but we are obviously not aware when it is going to end. For instance, let us consider the plot for 2009 H1N1 Influenza. For this specific strain of influenza, the **mode** is around 2 days and an half, so it is not 0 as we would expect from an exponential!



From: Mostaço-Guidolin, Luiz et al. (2011). A classical approach for estimating the transmissibility of the 2009 H1N1 pandemic. Canadian Applied Mathematics Quarterly. 19.

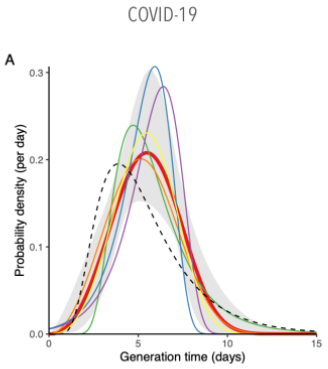
Figure 4.2: Infectious period distribution for the 2009 H1N1 influenza. One should note how it does not follow and exponential, therefore the mode is not located at $x = 0$, hence the approximation of the recovery rate constant in time is not realistic.

However, we can make estimates also for Covid-19, see figure 4.3.



From: Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases
Andrew W. Byrne, et al. medRxiv 2020.04.25.20079889 (2020)

(a) Some of the inferred recovery distributions for the Covid-19 disease.



from: Ferretti *et al.*
https://www.medrxiv.org/content/10.1101/2020.09.04.20188516v1

(b) Probability distributions for the recovery rate, note as this is not unique and it may vary according to data scientists are looking at. However, this is not an exponential at all

Figure 4.3

An observable easier to measure is the so called **serial interval**, namely the one that is between the onset of symptoms from an individual, to the onset of symptoms

for another individual. Despite this is just an **approximation**, it still can tell us some useful information.

At this point we should have been convinced by last considerations that these kind of diseases are *not Markovian*. Hence, the **recovery times depends on the time** we spend in that specific compartment. Now another problem arises, that is how to model this "non Markovianity". The family distribution that better describes our empirical data is the **Gamma distribution**:

$$P(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$$

the curve in fig. 4.4 starts to resemble somehow our observations. A similar family of distributions is the so called **Erlang distributions**, where the factorial replaces the Gamma function in the denominator.

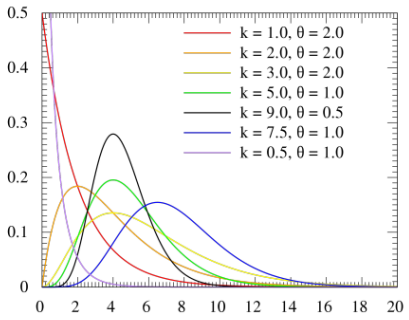


Figure 4.4: Gamma distribution for different changes of parameter.

Note that both distribution families introduced so far are able to reproduce quite well the shape of histogram, found by mean of empirical data.

We need now to discuss on how to include this **non-markovian** behavior into the **classical** epidemiological **models** we have introduced so far. One should note that, when analyzing them, we were assuming the markovian property. A **first** approach is the following. Let us consider a *trick* for the **infectious period**. We want to exploit the theoretical result that the sum of exponential random variables obeys to a gamma distribution. Specifically for our model, instead of considering only one transition with a constant rate (i.e. a single exponential distribution), we are going include **many transitions**, every one with its own rate. Therefore instead of having only a single infectious state, there will be many. Individuals will be able to move sequentially from one compartment to another. In this way, they will be forced to spend **at least some time infectious** before being recovered. Hence, we obtain a markovian model. Despite we are forced to spend some time infected before being recovered, the underlying model is still markovian. The only constraint is that we are imposing that these **stages** must be **sequential**.

Writing formally the equations with these last considerations:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = -\beta si \qquad \frac{\mathrm{d}i_1}{\mathrm{d}t} = \beta si - K\mu i_1 \qquad \frac{\mathrm{d}i_2}{\mathrm{d}t} = K\mu i_1 - K\mu i_2 \quad \ldots \quad \frac{\mathrm{d}r}{\mathrm{d}t} = K\mu i_K$$

Where we introduced K different infected compartments and the variable $i = \sum_{k=1}^{K} i_k$ that takes into account the prevalence for all the compartments. Hence a generic transition rate for each I transition is $K\mu$.

The infectious period distribution is the sum of these "intermediate" exponentially

distributed random variables, namely:

$$P(\tau) = \frac{(\mu K)^K}{\Gamma(K)} \tau^{K-1} e^{-\mu K \tau}$$

which is a gamma distribution, whose "scaling parameter" is $1/\mu_K$ Some limiting cases are:

- if $K = 1$, we obtain an *exponential* distribution;
- if $K \to \infty$ fixed, we obtain a *delta* distribution.

And this conclude the first approach in order to deal with non-markovian problems.

A *second* approach, which is **more general**, allows us to include *non-markovian* property both in **recovery** and **infections**. We want to discuss whether there is the possibility to write a more general model on networks as possible. Practically, this is feasible, despite at some points some sort of approximations will be needed. In order to do this, we have to slightly change our point of view and modify our approach: instead of probabilities, now we will be considering **events**. The idea is that we need to **model infections** and **recoveries** according to **two random numbers** which are drawn from a distribution as general as possible. Every time we extract a random number, which actually represents the time when we recover (i.e. the time we will spend as infected). This is defined as $R_i(t)$. Then, an other random number we need to extract is $M_{ij}(t)$ which represents the number of trials that i makes while trying to infect node j. This is to be repeated for all the neighbours. The sequence generated will be the following:

$$T_{ij}^{(1)} \leqslant T_{ij}^{(2)} ... \leqslant T_{ij}^{(M_{ij}(t))} \leqslant R_i(t)$$

where $T_{ij}^{(1)}$ is the first time at which node i tries to infect node j, then we have the second time, $T_{ij}^{(2)}$ and so forth. Hence, the **transmissibility** of a disease depends on how many trials we have available to infect.

The **algorithm** for infections looks like the following: once we have fixed a node i, we *draw* a number $R_i(t)$ that is *not* going to change unless we consider the following time step $t+1$. One should note that $R_i(t)$ is the time at which node i will recover. Later we consider one of its neighbours, for instance j, and draw the number that defines how many attempts i will have in order to infect j: $M_{ij}(t)$. For each of these attempts, we will draw $T_{ij}^{(n)}$ and if the latter is less than $R_i(t)$ the node j will be considered to have been infected. Otherwise, we keep going extracting $T_{ij}^{(n)}$, with the constraint that n must be at most $M_{ij}(t)$. We iterate this procedure for each of the neighbors of i and see which nodes are going to be infected next time step $t+1$. At $t+1$, then, we fix another node k that is infected and draw $R_k(t+1)$, choose a node j' among its neighbors, draw $M_{kj'}(t+1)$ and repeat what we have stated before. One last remark is that R and M can follow **any distribution** and not only the exponential one. How we extract T is not important, because the only point that matters is how R's and M's are distributed.

Let us make some **assumptions** in order model these distributions reasonably, and finally try to solve our problem analytically. We can consider both R's (recovery times distribution) and I's (infection times distribution) to be **peculiar** of the disease, therefore must **not depend** both on the **node** and on **time**. Obviously they should take into account the background information, for instance lockdown, particular restrictions...but momentaneously we will skip this part.

In other words we are just **reducing** the **complexity** of our problem by stating that $R_i(t) \to R$ and $M_{ij}(t) \to M$. We define the long run probability v_i to be the **probability** that node i is *infected* in the **steady state**.

Now it is time to build our model. Let us suppose that we are in the **steady state**, that is to say that there is no more transiency. In the long run, for a period of time $[0, S)$ and S large enough, the number of times that node j was infected is proportional to S. Therefore we are introducing in our model the property that the number of times that we contract the infection is linear wrt time. *On average*, the length of each infected period is $\mathbb{E}[R]$ (expected time to recover). Then, in the long run, the *number of times* that node j *has been infected* in a period of length S can be rewritten as:

$$\frac{v_j S}{\mathbb{E}[R]}$$

Recall now that for every time step in which node j is infected, it will attempt to infect its neighbour i on average $\mathbb{E}[M]$ times. *On average*, the *total number* of infection **attempts** from node j to i in the long run is the following:

$$\frac{v_j S \mathbb{E}[M]}{\mathbb{E}[R]}$$

Now we want to make a *mean-field* assumption. We recall that it refers to the joint probability that j is infected while i is susceptible and allows us to factorize in the following way:

$$P[\sigma_i(t) = 0 \ , \ \sigma_j(t) = 1] \sim P[\sigma_i(t) = 0] P[\sigma_j(t) = 1] = (1 - v_i) v_j$$

Where we replaced the two factors by the probabilities of the respective node for large time intervals.

Let us consider now the number of times which i actually received the infection from j. *On average*, in the long run, the **number of successful attempts** made by j to infect its neighbour i i is the following:

$$S \frac{\mathbb{E}[M]}{\mathbb{E}[R]} v_j (1 - v_i)$$

That is nothing more than the total number of attempts we computed before, times the probability that node i was not infected. If we sum over all the neighbors of i, we obtain the the **total number of successful infections** i will receive during time interval $[0, S)$:

$$S \sum_{j=1}^{N} a_{ij} \frac{\mathbb{E}[M]}{\mathbb{E}[R]} v_j (1 - v_i) = v_i \frac{S}{\mathbb{E}[R]}$$

Where the last equivalence asymptotically holds only in the **steady state**, and the rhs is the number of infected periods experienced by i. Simplifying last formula we have that:

$$v_i = \mathbb{E}[M](1 - v_i) \sum_{j=1}^{N} a_{ij} v_j$$

hence the probability of i being infected depends on the sum over all its neighbours, times the term $\mathbb{E}[M]$ which is the average number of infection attempts it will experience during that time interval.

This last expression should sound familiar: it is exactly what we obtained from the **linearization** of the quenched mean field approach (IBMF). The only difference is that before we had:

$$\mu \varepsilon_i^* = \beta (1 - \varepsilon_i^*) \sum_{j=1}^{N} A_{ij} \varepsilon_j^*$$

While now a generic infection term $\mathbb{E}[M]$ replaces the β term in the $IBMF$. This generic infection term therefore encodes all the distributions with the same expected

value $\mathbb{E}[M]$. Note that this last result holds *only* in the *steady state* and *does not* depend on the shape of the distribution M, but only on its expected value!

Let us briefly discuss the **implications** for this. One should note that the definition of the distribution M already includes the recovery term R, since at the end of the day it sets a "lower-bound" for it[1]:

$$T_{ij}^{(1)} \leqslant T_{ij}^{(2)} ... \leqslant T_{ij}^{(M_{ij}(t))} \leqslant R_i(t)$$

Moreover, the expected number of infection events in a Poisson process with intensity β and an exponential recovery time whose expectation is $1/\mu$, we can prove that:

$$\mathbb{E}[M] = \frac{\beta}{\mu}$$

Last information we can obtain is that the epidemic threshold m_c can be derived from:

$$m_c = \mathbb{E}[M_c] = \frac{1}{\Lambda_{max}}$$

4.2 Interacting diseases

Lecture 11.
Thursday 5th November, 2020.
Compiled: Saturday 14th August, 2021.

In this lecture we will make another step further and take into account something which is quite common in reality: diseases usually does not spread independently of each other, but interact with the other ones.

One should know that there might be many variants of the the same disease, with special regards to the ones we have seen so far. For instance, considering seasonal influenza there are many viruses which are similar, i.e. they form a **family**. In addition one must take into account that these viruses may **interact** among each other. The fact that there might be many variants for a single virus is an important feature to be taken into account.

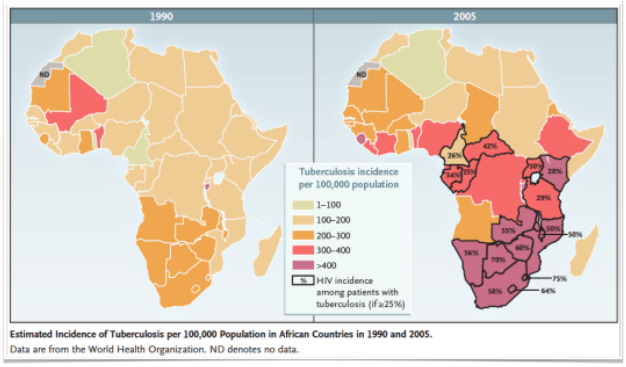
Let us now consider the simplest case, where we have only two different diseases that **cooperate**. That is to say that one disease **boosts** the spread of the other one. The map fig.4.5 represents the spreading of Tuberculosis (TB) in 1990. The latter is a disease which has a very long latent period: it can stay in a latent state actually for many many years or, sometimes, we may have contracted it without it never showing up. Let us compare the same the same map some years later, in 2005. One should note that the disease has exploded, especially in southern regions: numbers almost have doubled. The reason is that, during that time window, HIV reached the African continent. HIV is a disease which compromises our immune system and makes it less efficient: the probability of getting any other disease is higher than the normal.

This should explain us what is depicted in the map: when people contract HIV, their immune system becomes inactive and consequently Tuberculosis can activate. As one may have understood, HIV has provided much help for the spreading of TB: numbers shown represent the fraction of patients which get both HIV and TB. This actually was the **first example** of **interaction** of two diseases.

The downside, when **competition** between diseases may arise, is shown as an example by seasonal influenza. From fig. 4.6 we can see the distribution of the different influenza viruses that are out during a single season. Individuals can be infected by different strains of influenza, therefore the distribution of the prevalence of total infected people is the sum of the values regarding the three strains.

It can be shown that our immune system has a sort of **memory** which provides a **long lasting** immunity. The main consequence is that if we accidentally get a

[1] indeed when we draw M, the recovery time R can be any number greater than M itself: in a certain sense R is fixed once M has been fixed.



From "Tuberculosis in Africa, combating an HIV-driven Crisis" Chaisson, R.E. & Martinson, N.A., New Eng. J. Med., March 2008

Figure 4.5: Map of the prevalence of TB for different years (1990 and 2005).

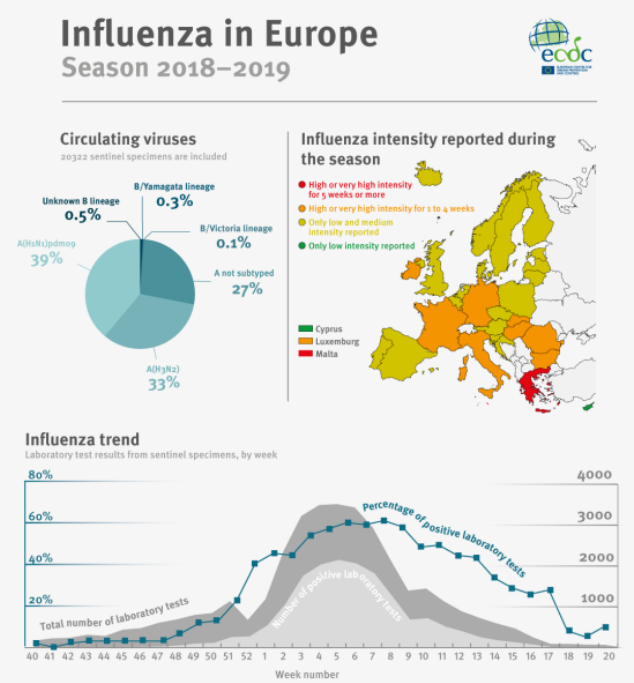


Figure 4.6: Distribution of different strains of influenza for a single season.

disease, there is a possibility that we end up not contracting it once again for some time in the future. Moreover, it can provide also a sort of **cross-immunity** for some other diseases that are similar to the one we contracted, even though our immune systems had never faced them. As one can imagine, this implies that the actual susceptible population is just a reduced fraction of the whole one, and should explain why we do not usually observe influenza pandemics.

However, **evolution** applies to viruses as well, since they may mutate during time. For instance, HIV is an extremely volatile virus, and is known as one of the fastest evolving entities known: science has categorized different types and subtypes, which are distributed among different regions.

We are now going to focus only on the simplest settings, in which we consider competition and cooperation between only two diseases. We want to discuss how it is possible to **model** these kind of **interactions** between diseases/strains and their behaviors. The *simplest solution* to our problem is to **couple different dynamics**, let us see how.

For instance, the simplest case one can think of is to couple two different diseases

whose dynamics are respectively **SIS** and **SIS**. We end up having twice the number of states (see fig 4.7), since we take into account all the possible combinations between compartments. A single node i at time t can be be either one of the following states $\{SS, SI, IS, II, \}$ with its own probability density $\{[\rho^{SS}]_t^i, [\rho^{SI}]_t^i, [\rho^{IS}]_t^i, [\rho^{II}]_t^i\}$. Therefore, every disease will be defined thanks to its own **parameters** $\{\beta_1, \beta_2, \mu_1, \mu_2\}$. However, some more are to be introduced, namely the ones that **encode** the **interaction** between diseases.

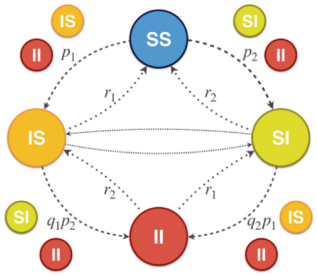


Figure 4.7: Coupled SIS model and all possible combinations and different transition probabilities between each compartment.

As an **example**, let us discuss one of these models. Let us consider a **classical heterogeneous mean-field** in which we have *two different diseases* that can spread inside its own network. We are dealing now with the most general case: diseases may spread in different manners and by different means. For example, one may contract a disease orally, while an other one by blood contact. This is the reason why we need to take into account different networks in which diseases spread: every network is peculiar of the disease and can be different one from another, having its own degree distribution or topology. As an example, for HIV we have a network that is defined by its degree distribution $P(k)$, while for TB an other network $P(l)$. However, when dealing with both of them we shall use the joint probability for the two distributions $P(k, l)$.

Recalling now that we are doing a degree based mean field, we are able to divide our network in four different classes according to the compartment they belong to and their degree distributions: $\{SS(k, l), SI(k, l), IS(k, l), II(k, l)\}$. As an example, $SS(k, l)$ is the fraction of nodes of degree k in the first network and l in the second, that is susceptible for both diseases.

At this point it comes to take into account, and therefore model, the interaction effects between the two diseases (see fig. 4.8). These interactions may result in three effects:

- **modified susceptibility** $\lambda_\mathbf{a}$: being infected of one diseases makes an individual more ($\lambda_a > 1$) or less ($\lambda_a < 1$) probable to be infected by a second one. It is the case for the HIV that increases the probability of contracting TB, or for a strain of a seasonal influenza that inhibits individuals to get influenza of another kind

- **modified infectivity** $\lambda_\mathbf{b}$ once we get infected by one disease, we are less/more infectious wrt the second one. Hence, we infect less/more other people with the second disease.

- **modified infectious period** η if we are infected of one disease, it can favor/hinder the recovery from the other disease. As for the second case, it may happen when our immune system is compromised.

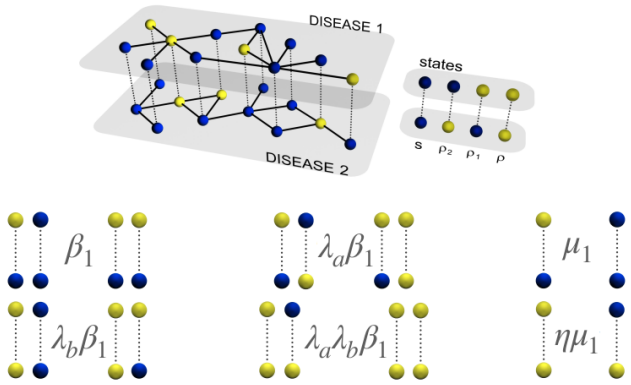


Figure 4.8: Coupled SIS networks and their effects on each other.

In this way we have covered all the possible interactions between diseases, given this simple model. Obviously, the interactions may result in an increasing susceptibility or infectivity, along with the tuning aforementioned parameters. Actually, despite in this last model we have introduced all possibilities and cases that we may observe, only a subset of them at time is meaningful and has some sort of biological sense.

Now we may wonder how do the individuals flows between compartments look like:

$$\dot{SS}(k,l) = -(k\sigma_1 + l\sigma_2)SS(k,l) + \mu_1 IS(k,l) + \mu_2 SI(k,l) \tag{4.1}$$

$$\dot{IS}(k,l) = k\sigma_1 SS(k,l) - l\lambda_a\sigma_2 IS(k,l) - \mu_1 IS(k,l) + \eta\mu_2 II(k,l) \tag{4.2}$$

$$\dot{SI}(k,l) = l\sigma_2 SS(k,l) - k\lambda_a\sigma_1 SI(k,l) - \mu_2 SI(k,l) + \eta\mu_1 II(k,l) \tag{4.3}$$

$$\dot{II}(k,l) = k\lambda_a\sigma_1 SI(k,l) + l\lambda_a\sigma_2 IS(k,l) - (\eta\mu_1 + \eta\mu_2)II(k,l) \tag{4.4}$$

Where we have introduced the *infection terms* σ_1, σ_2 for the disease 1 and 2. The former can be propagated by both IS and II individuals and the infection term is:n

$$\sigma_1 = \beta_1(\Theta_1^{IS} + \lambda_b\Theta_1^{II})$$

while the disease 2 can be propagated by both SI and II:

$$\sigma_2 = \beta_2(\Theta_2^{SI} + \lambda_b\Theta_2^{II})$$

For the sake of completeness we write how Θ_i look like: they are the probabilities that a link of network $1/2$ points to an infected. For network 1 it holds that:

$$\Theta_1^{IS} = \frac{\sum_{k,l} P(k,l)kIS(k,l)}{\sum_{k,l} P(k,l)k} \qquad \Theta_1^{II} = \frac{\sum_{k,l} P(k,l)kII(k,l)}{\sum_{k,l} P(k,l)k}$$

While for network 2:

$$\Theta_2^{SI} = \frac{\sum_{k,l} P(k,l)lSI(k,l)}{\sum_{k,l} P(k,l)l} \qquad \Theta_2^{II} = \frac{\sum_{k,l} P(k,l)lII(k,l)}{\sum_{k,l} P(k,l)l}$$

The structure is absolutely the same as the one obtained for the one disease framework, and also the procedure to solving these equations. However, we will not start computations and derive all the results since some passages are extremely tedious.

In order to solve these equations, firstly we need to assume that we are in the **steady state**. Later we are able to write down a couple of self-consistent equations

for σ_1 and σ_2, and then solve them by finding the intersection. Finally, the **epidemic threshold**:

$$\beta_1^c(\sigma_2) = \mu_1 \frac{\langle k \rangle}{\sum_{k,l} P(k,l) k^2 \frac{l^2\sigma_2^2\lambda_a^2\lambda_b + l\sigma_2(\eta\mu_2\lambda_a + \lambda_b(\lambda_a\mu_1 + \lambda_a\mu_2)) + \mu_2(\eta\mu_1 + \eta\mu_2)}{l^2\sigma_2^2\lambda_a\eta + l\sigma_2(\eta\mu_1 + \eta\mu_2 + \lambda_a\eta\mu_2) + \mu_2(\eta\mu_1 + \eta\mu_2)}}$$

that is quite complex, but the form resembles the one of before.

One should see from the formula that the **epidemic threshold** of the **first** disease **depends** on the **prevalence of the other**. An other way to see it is that we are assuming that one disease is already there, and then insert a second one and check the effects on the epidemic threshold.

Let us see what happens to the epidemic threshold for different cases.

For instance, let us consider the 3D epidemic diagram shown for **cooperating diseases** (see fig. 4.9) with $\lambda > 1$ and $\eta < 1$.

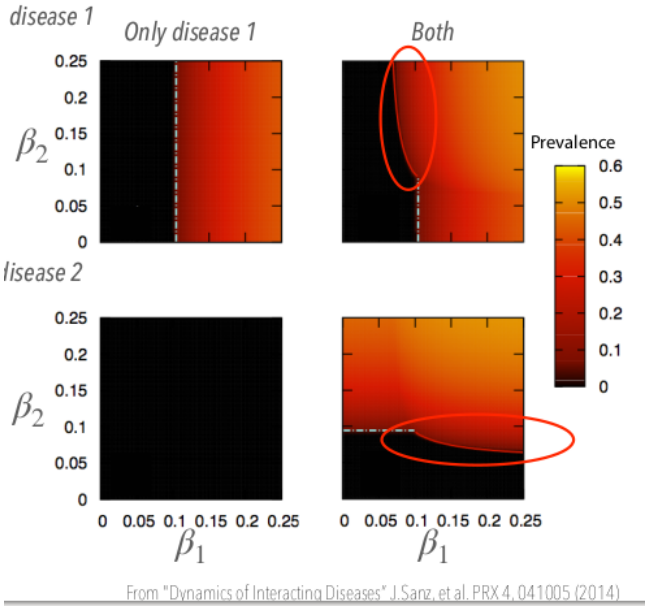


Figure 4.9: Cooperating diseases, $\lambda > 1$ and $\eta < 1$

If we consider a single disease, one see that we get back the old results (in 2D): since we have not introduced yet a second disease we are not able to see it. On the other hand, if the parameter β_2 is below the critical threshold, we cannot note any difference and the spreading is the same as before. However, when we overcome the threshold for a disease, one should note the that other's decreases and therefore the disease spreads easier and with a larger prevalence.

The opposite actually occurs when we consider **competing diseases** (see fig. 4.10), that is the case $\lambda < 1$ and $\eta > 1$. For either one of them it is more difficult to spread once we have overcome the critical threshold for the other's. This concludes the discussion when taking for heterogeneous mean field degree assumption.

One may wonder now how **quenched mean-field** equation look like (individual based formulation). For the sake of simplicity we are going to discuss only a single network, and we will take into account only its main effect, that is the one on the **modified susceptibility**. Also under this assumptions equations $\{[\rho^{SS}]_i^{t+1}, [\rho^{SI}]_i^{t+1}, [\rho^{IS}]_i^{t+1}, [\rho^{II}]_i^{t+1}\}$ can be written, whose structure is exactly the same one as before for a single disease case. Each term contributes and plays a role in the probabilities $[\rho^{IS}]_i^{t+1}, [\rho^{SI}]_i^{t+1}$. Since these expressions are really long and complex, we are going to skip this part. Nonetheless we will briefly discuss an interesting **assumption** we make during our computations: we did insert the functions f_{SI}, f_{IS}. It was done to consider the fact

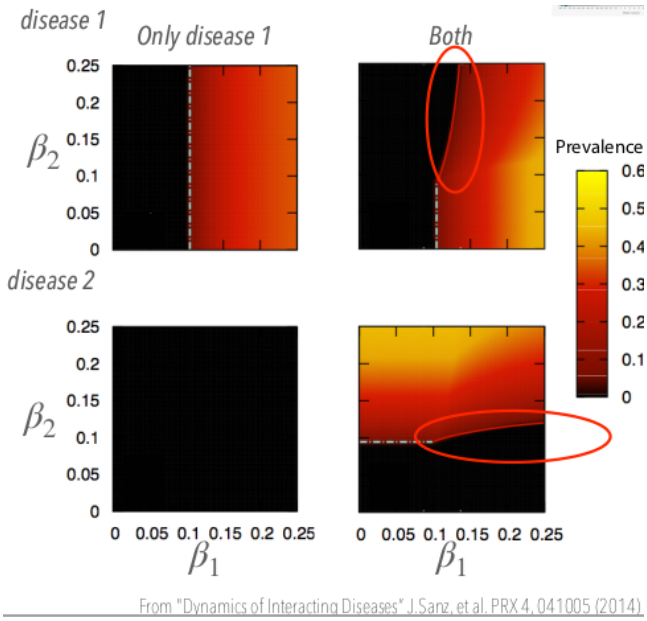


Figure 4.10: Competing diseases, $\lambda < 1$ and $\eta > 1$

that we cannot contract two diseases at the same time, it would be unrealistic. However, if during our simulations it happens, we pick only one disease at random. The function f_{IS} is the following, for the $[\rho^{IS}]_i^{t+1}$ expression:

$$f_{IS} = \frac{q_{IS}(1 - 0.5q_{SI})}{q_{IS}(1 - 0.5q_{SI}) + q_{SI}(1 - 0.5q_{IS})} \tag{4.5}$$

where:

$$q_{IS} = 1 - \prod_j^N \left[1 - A_{ij}\beta_1 \left([\rho^{IS}]_j^{t+1} + [\rho^{II}]_j^{t+1}\right)\right] \tag{4.6}$$

and:

$$q_{SI} = 1 - \prod_j^N \left[1 - A_{ij}\beta_2 \left([\rho^{SI}]_j^{t+1} + [\rho^{II}]_j^{t+1}\right)\right] \tag{4.7}$$

These equations can be solved numerically by iteration as we did for the single disease scenario. One should have noticed that when $\lambda = 1$ we get back to the classical case. Let us see now what happens when the two diseases **cooperate**. Recalling that:

$$\rho = \frac{1}{N}\sum_{i=1}^{N}(\rho_i^{IS} + \rho_i^{SI} + \rho_i^{II}) \tag{4.8}$$

If the probability of getting the disease is $\lambda = 2$ one can see in fig. 4.11 that the curve becomes steeper and, as λ increases more this behavior accentuates even more and looks like exploding. Moreover one should note that, despite the infectivity β is exactly the same, the prevalence ρ shows some discontinuities that becomes larger as more the two diseases cooperate more and more.

Let us now see what happens when two diseases **compete**. The resulting effect is the so called full **cross-immunity**: once we contracted a disease we cannot get the other one. Accordingly the prevalence is:

$$\rho = \frac{1}{N}\sum_{i}^{N}(\rho_i^{SI} + \rho_i^{IS}) \tag{4.9}$$

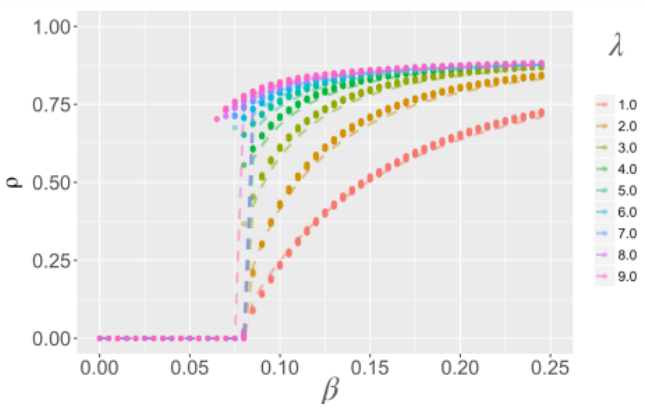


Figure 4.11: Cooperating diseases, numerical simulations results for different λ.

As expected, the prevalence shown in left figure 4.12 does not change at all and is the same as before. While, plotting the difference $|\rho^{IS} - \rho^{SI}|$ we observe some oscillations in values slightly after the critical threshold. This is explained by saying that only a disease can survive, and which is given by chance, being the two symmetrical (see right fig 4.12). As β increases, we see as the difference approaches zero: both of them here survive each with the same prevalence. That is to say that a half of infected population has contracted a disease, while the other half the second one. For large β we see as the two diseases coexist.

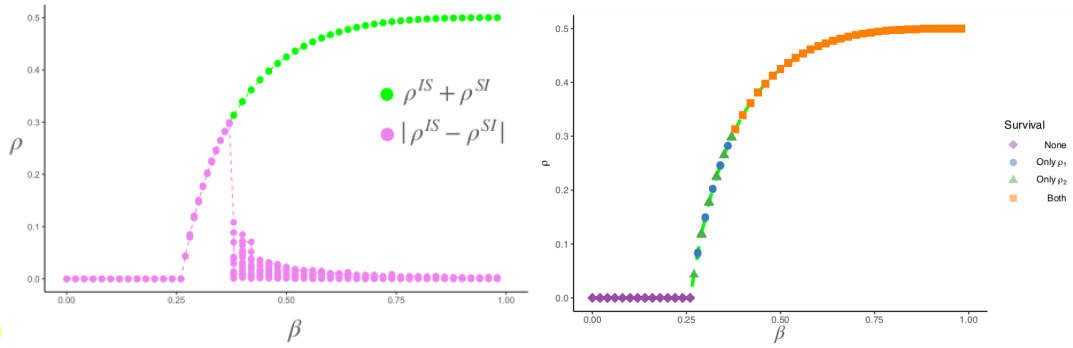


Figure 4.12: **Left:** Competing diseases, numerical simulations results. **Right:** Competing diseases, numerical simulations results with particular attentions to the **survival disease**. In the intermediate regime, the survivor is chosen by chance.

This system actually can be studied also in terms of its dynamics. We start by noting that it has two stable points. They might for instance coincide in the origin in the (β_1, β_2) space: here obviously both of them are absent. We can continuously move these points and increase (β_1, β_2) that are not anymore degenerate: the system will settle in either one of these attractors after some time. Later, when we have overcome a certain threshold, the stable points will again coincide and find ourselves exactly in the middle. Here, both diseases are coexisting with exactly the same prevalence. However if we introduce a slight difference between them, i.e. they are not symmetrical anymore, the stable point will not be in the middle anymore but slightly move according to what we have changed.

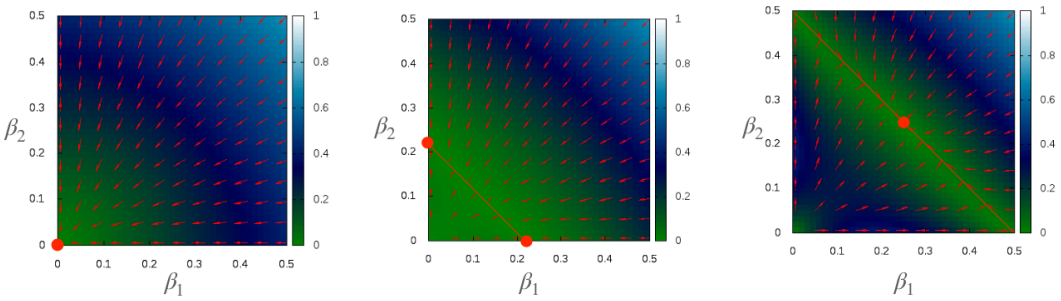


Figure 4.13: Competing diseases, manyfold that describes the dynamics of the prevalence according to different values of β_1 and β_2.

5

Spreading in social systems

Let us now discuss about how we can apply the epidemic models we have studied so far to other scenario, especially in **social systems**. Indeed, spreading of information inside social systems shares many similarities with epidemic spreading (i.e. "viral information"). This is why one can find a huge literature about these topics, where **epidemic models** are **adapted** in order to include also *social aspects*.

However, there are also some differences: the **communication aspect** actually behaves in its own way and leads to some effects that cannot be totally included in simple epidemic models. In social contexts things are a bit more complex:

- **information transmission** is an *intentional* act for both sender and receiver;
- often **beneficial** for senders and receivers (e.g. **reinforcement**), that is to say we are "invited" to acquire more information while feeling like spreading it. Indeed, the information can be replicated by different sources: for instance we often see the same information different times in different places;
- influenced by cognitive and **psychological factors**;
- content of information matters (e.g. **homophily**), so we tend to cluster among people that share the same knowledge or type of information.

According to these last properties, there are different instances of **spreading**, and a single one can be **defined** as:

- **simple contagion** if there is **no memory**, **no reinforcement**
- **complex contagion** if **multiple exposures** and reinforcement are involved, i.e. we keep trace of past interactions that can be either independent of each other.

In addition, we shall introduce the so called **threshold models**, that present some sort of threshold effect. Empirically, we may say that if only few friends bought a certain product we would not feel like buying it, as it would be if nearly half or hall of them have bought it. In the old fashioned models, when we were susceptible, our infected neighbours tried to infect us along with a certain probability. But, in this case, if our **neighbours number** is *lower* than a certain *threshold*, we **cannot be infected** by them. Conversely, if their number is either equal or above that threshold we are going to change our state.

In particular, threshold models lead to **information cascades**, that is to say that if someones acquires some information then he will start spreading it to other people. In turn, they will keep on sharing it until it will have reached a huge amount of individuals as if in a sort of *cascade*.

5.1 Complex contagion

We want now to generalize what we have told so far, and introduce some sort of more abstract model which is able to mediate between complex and simple contagions.

The main **assumptions** we will make for such general contagion model, able to reproduce both simple and complex contagions, are the following:

- we will assume to deal with **well-mixed population**;
- we will introduce the usual three **compartments** as in the SIR models: thus we will end up with different classes S *(susceptible)*, I *(infected)* and R *(recovered)* of individuals;
- we will **keep trace of past interactions** up to time T, since we want to take into account the effects given by different exposures in different period of times;
- we will change the **way information is spread**: given a successful interaction with an infected j, a susceptible individual i gets a "dose" of infection $d_i(t)$. This is done to take into account that the more we see a certain information, the more likely it will be spread in the future;
- Once the **accumulated dose** $D_i(t) = \sum_{t'=t-T+1}^{t} d_i(t')$ exceeds a fixed threshold d_i^*, then the i-th susceptible individual becomes in turn infected.

As one can see, we actually started from a SIR dynamics and applied some changes to it, in order to create the general contagion model that can be applied to social networks. Whereas, regarding the dynamics of the **infection** process:

- at each time step t:
  - each individual i contacts a random individual j;
  - if $i = S$ and $j = I$, with probability p, individual i gets a "dose" of infection $d_i(t)$ that is distributed according to a dose size distribution $f(d)$;
  - with probability $1-p$, that is to say if the contact has not been successful, $d_i(t) = 0$.
- each individual keeps trace of the doses acquired in T timesteps via the **"cumulative" dose**: $D_i(t) = \sum_{t'=t-T+1}^{t} d_i(t')$.
- if the cumulative dose $D_i(t)$ is larger than the individual threshold d_i^*, individual i gets infected.

As for the **recovery** process, the dynamics is more or less resembling the classical dynamics:

- if $D_i(t)$ gets below d_i^*, i recovers with probability r;

If one would like to create a more complex model, it is also possible to add an $R \to S$ transition that occurs with probability r'. This could be done to simulate an **SIRS** model with reinfection dynamics. The limiting case, namely the one with $r = 1$ and $r' = 1$, we end up again having an SIS-like dynamics.

Having said so, let us now summarize the main parameters we introduced so far:

- p and r are **infection** and **recovery** probabilities. They actually play the same role as β and μ in the epidemiological models;
- $d_i(t)$ **"dose" per infection**, which distributes according to $f(d)$;

- d_i^* **threshold**, in turn distributed following $g(d^*)$.

Note that $f(d)$ and $g(d^*)$ can be *any* distributions. By varying them we can reproduce different behaviors, i.e. obtain different dynamics. However, with specific choices of p, $f(d)$ and $g(d^*)$, it is possible to tune the effects of the threshold dynamics to our system. For instance, for either low or null threshold (or relatively high-valued doses distributions) we observe a dynamic really resembling the ones we have studied so far. Conversely, we may end up to models where the "threshold dynamics" is the one that characterize and strongly determines the behavior of the system.

Let us now **formalize** mathematically what we have said up to now and finally try to solve it. Firstly, let us define the **probability** for an **individual** with $K < T$ contacts to be **infected** as:

$$P_{inf}(K) = \sum_{k=1}^{K} \binom{K}{k} p^k (1-p)^{K-k} P_k \tag{5.1}$$

Let us discuss the single terms. $p^k(1-p)^{K-k}$ is the probability for the contact to be successful. Note as it a *Bernoulli distribution* with K trials and k successes. This is multiplied by the Binomial coefficient that takes into account all the possible combinations of k successes in K trials $\binom{K}{k}$ and, finally, it is multiplied by the factor P_k. In particular P_k is the *average fraction of infected individuals* after having received k doses in T time steps:

$$P_k = \int_0^\infty \mathrm{d}d^*\, g(d^*) P\left(\sum_{i=1}^{k} d_i \geq d^*\right) \tag{5.2}$$

which one can easily note, does depend on the thresholds. Indeed, $P\left(\sum_{i=1}^{k} d_i \geq d^*\right)$ is the probability that k doses exceed d^*.

This model can actually be solved numerically for any distribution of $f(d)$ and $g(d^*)$. However, for some specific cases we can recover classical dynamics. Indeed let us consider:

- if the probability of a successful contact $p < 1$, the dose has a fixed size $f(d) = \delta(d-1)$ and fixed threshold $g(d^*) = \delta(d^*-1)$, we observe **epidemic spreading** where interactions are independent (see fig. 5.1). In particular, all contacts share the same infection probability and the threshold is $d^* = 1$, i.e. one successful contact is enough to contract the disease. In addition, if we want to recover an SIS dynamics, we must constrain the dose to be the same for everyone and the threshold to be unique

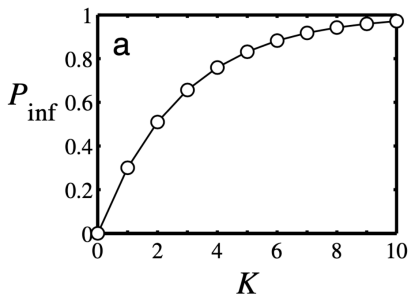


Figure 5.1: Generalized Complex Contagion model: epidemic spreading.

- if the probability of a successful contact $p = 1$, the dose has fixed size $f(d) = \delta(d-1)$ and fixed threshold $g(d^*) = \delta(d^*-5)$, we obtain a **deterministic**

threshold model (see fig. 5.2). In particular, we arbitrarily fixed the threshold at $d^* = 5$, that is to say that we need at least 5 encounters to be infected (they do happen with $p = 1$, so every contact is actually successful!). Hence, despite the dose size is exactly the same as before being the distribution peaked in 1, in this case we actually need more than a single contact to contract the disease.[1]

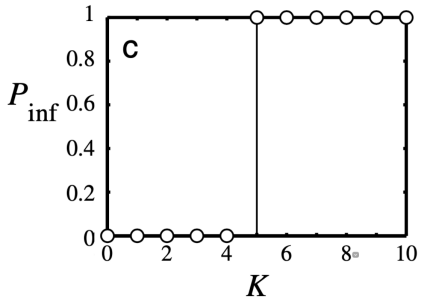


Figure 5.2: Generalized Complex Contagion model: deterministic threshold model.

- if the probability of a successful contact $p = 1$, the dose size $f(d)$ distributes log-normally and the threshold is fixed $g(d^*) = \delta(d^* - 5)$ we obtain the so called **stochastic-threshold model** (see fig. 5.3). In particular, we observe that the threshold is still fixed at $d^* = 5$, but now the "dose" per successful contact varies. In this case we assume contacts to not be equal among them, so despite the threshold being fixed, the dose size is actually different.[2]

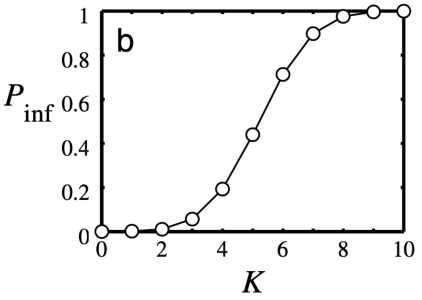


Figure 5.3: Generalized Complex Contagion model: stochastic threshold model.

5.2 Applications to Online Social Networks

We want to apply our framework to real social networks, thus we want to understand how hashtags or memes spread in online social communities. Let us consider the analysis of real data extracted from Twitter. In the latter we may have different types data: however we are going to focus only on retweets (RT) and mentions (@) that contribute to the diffusion of hashtags in different communities. In particular we want to see whether the structure, namely reinforcement and homophily, inside communities does have a role in the spreading.

To quantify the fraction of information that flows inside and outside of a community, we will introduce the following weights:

- $\langle w_{\circlearrowleft} \rangle_c$ is the average weight (number of tweets) per link inside the community;

[1] for instance 5 friends that show to me the same information.

[2] for instance we may trust some friends/people more than others, hence we give more importance to their information rather than acquaintances' one.

- $\langle w_{\frown} \rangle_c$ is the average weight (number of tweets) per link outside the community.

And the same for users activity:

- $f_{\circlearrowleft}$ is the fraction of activity inside the community;
- $f_{\frown}$ is the fraction of activity outside the community.

If information in Twitter spread like a *simple contagion*, there should not be any noticeable differences in the spreading process inside and outside a community (e.g. we observe *no reinforcement*). Conversely, if we saw that the average weights inside a community would be larger than outside ones, actually we should take into account some sort of reinforcement.

In Fig. 5.4 we can see the results showing the average weight inside and outside a community. They are actually pretty similar, but if we take a look more closely, we may notice that the averages for spreading inside a community are little higher, therefore we can conclude that homophily and reinforcement do play a role.

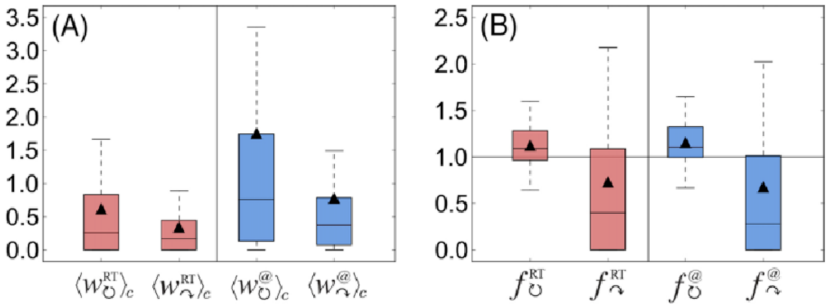


Figure 5.4: Spreading inside a community is favored (effects of homophily and reinforcement are noticeable).

To clarify this last point, we introduce a new metric on the level of single hashtags (h). Hence we measure the average popularity of an hashtag inside and outside a community, i.e. for every hashtag we measure the popularity that a tweet exploiting the latter had. In particular, for each hashtag:

- we measure the **usage dominance** $r(h)$. This is the ratio of tweets produced inside the "main" community of h and the total number of tweets containing h, namely $T(h)$. We expect that this metric is low for viral spreading and high for complex contagions;
- we measure the **usage entropy** $H(h)$: how h is distributed across communities. It is high for viral spreading and low for complex contagion;
- we measure the **average exposure** $N(h)$: which is the average number of exposures needed to adopt hashtag h. It is low for viral spreading and high for complex contagion.

In order to analyze it, we use some reference models (4 models $M_{1,\dots,4}$) in order to represent different baseline behaviors (see Fig. 5.5 for more details):

- the simplest model is M_1 where, for a given hashtag h, we randomly sample the number of tweets or users using the averages we have from real data (i.e. we assume that data is extracted at random). In such way, we can obtain some sort of **average behavior** for all the hashtags where we do not consider any community, neither network structure;

- the second model M_2 is simply an epidemic model. In particular, it takes into account the *network structure* while neglecting social reinforcement and homophily. Each hashtag therefore starts from some random users (the "seed") and, at each timestep, it spreads to other users according to a certain probability. This is indeed a reference model for **simple contagion**;
- However we can have more complex models which can take into account *network structure*, *reinforcement* and *homophily*. This is a reference model for **complex contagion**.

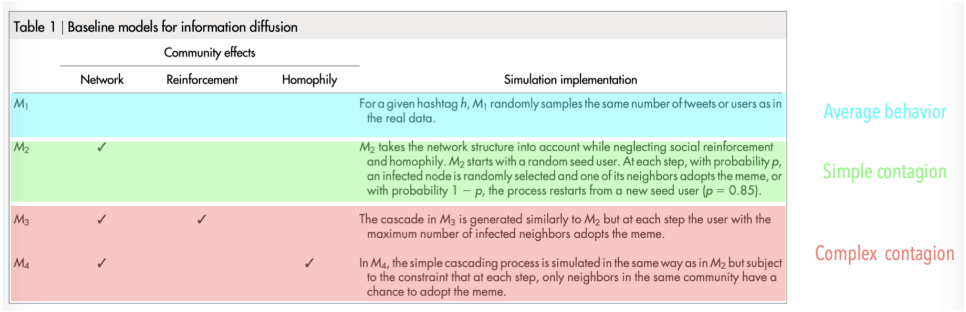


Table 1 | Baseline models for information diffusion

	Community effects			
	Network	Reinforcement	Homophily	Simulation implementation
M_1				For a given hashtag h, M_1 randomly samples the same number of tweets or users as in the real data.
M_2	✓			M_2 takes the network structure into account while neglecting social reinforcement and homophily. M_2 starts with a random seed user. At each step, with probability p, an infected node is randomly selected and one of its neighbors adopts the meme, or with probability $1 - p$, the process restarts from a new seed user ($p = 0.85$).
M_3	✓	✓		The cascade in M_3 is generated similarly to M_2 but at each step the user with the maximum number of infected neighbors adopts the meme.
M_4	✓		✓	In M_4, the simple cascading process is simulated in the same way as in M_2 but subject to the constraint that at each step, only neighbors in the same community have a chance to adopt the meme.

Figure 5.5: Reference models for information spreading in online social networks.

Let us consider Fig. 5.6 where we can see $r(h)$, $H(h)$ and $N(h)$ in function of the number of tweets T and number of users U. Black lines represent the real data, while the dashed line represents the theoretical results returned from model M_1 (average behavior), whereas the red square M_2 (simple contagion model) and last the blue and green lines given by models M_3 and M_4 (complex contagion).

One can note as popular (in grey) and not popular hashtags do result in two different behaviors, therefore can be easily distinguished and separated into two different classes. In particular, it holds that:

- popular hashtags (large T and U) spread like epidemics (viral);
- less popular ones follow a complex contagion.

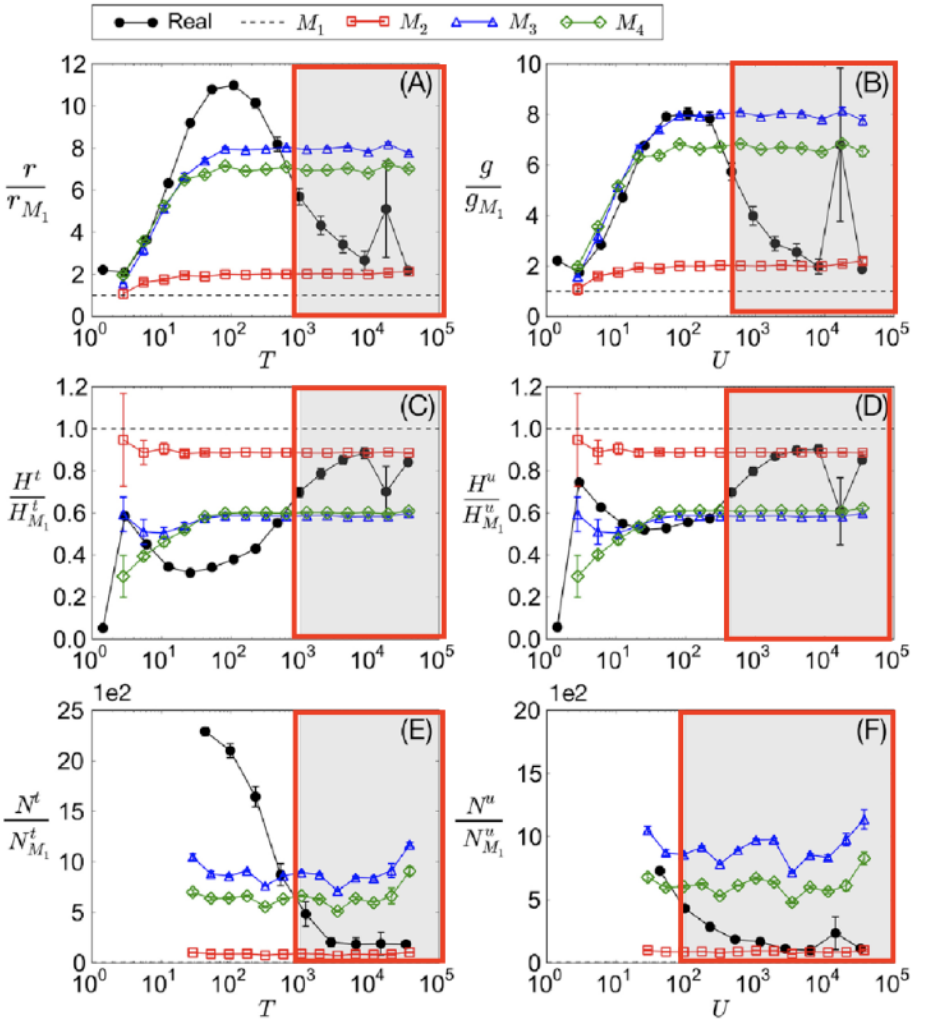


Figure 5.6: Results of information spreading in online social networks.

Part II

Poletto's Lectures

6

Introduction to metapopulation models

Lecture 14.
Friday 13^{th} November, 2020.
Compiled: Saturday 14^{th} August, 2021.

6.1 Spatial spread of epidemics

Let us start now to dig deeper adding some complexity to our models. In particular now we want to understand why *spatial spread* of epidemics is important, and its possible effect over public policies. It allows us to estimate the **invasion risk** for a given territory, hence understanding whether a place is more likely to develop an epidemic. In this way we are able to model and realize the **conditions for containment**, since containing an epidemic spatially helps in its management. Moreover, we will discuss about the so called **spatial coupling**, that is to say how the epidemic in a given area influences the epidemic in another. This should have explained us why **spatial information** is an **essential ingredient** in epidemiology: we need to know where the epidemic is at a given moment in order to take the proper countermeasures.

There might be different **drivers** of spatial transmission:

- **direct** transmission among humans: the spatial spreading of epidemics is strongly affected by the *human mobility*. Hence the pathogen spreads carried by travelling individuals;
- **vector borne**: the spatial propagation requires both *human mobility* and the local presence of *competent vector* (mosquitos, rats...). Mobility for vectors is also possible, too;
- **different drivers**, such as food borne, environmental diseases, zoonotic pathogens, etc.

As said, **human mobility**[1] behavior determines the spatiotemporal pattern of spreads. This should take into account that there exist **different types** of mobility, and the actual kind becomes relevant according to the epidemic and the epidemiological questions we are facing.

6.1.1 Human mobility

There are actually different types of data for the human mobility network. For instance, **air travelling** data is collected by the International Air Transport Association (IATA). It can be actually purchased, since the information publicly available is limited. There are two **types** of this data:

- **segment**: number of seats for each company between two airports;

[1] Human mobility: Models and applications, Barbosa et al. Physics Reports 734 (2018)

- **origin-destination**: number of passengers travelling between origin-destination, obtained from the tickets purchased.

Doing a similar analysis to the one we have done so far for the "air" network, one can note two facts related to the number of connections and the number of passengers: both **topology** and **traffic distribution** are **heterogeneous**.

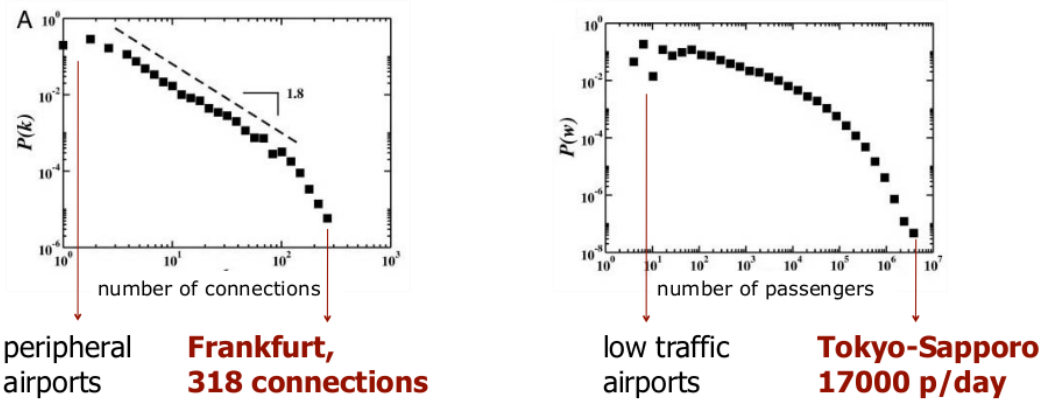


Figure 6.1: Whole segment network worldwide, 2002. Topology of the network and flux of passengers graphical analysis.

We can indeed find some **scaling relations** between fluxes, number of connections and population, as one can see from Fig. 6.2. The average number of route $i \to j$ is determined by a non linear function of the traffic in airports, and it has form $w_{ij} \sim (k_i k_j)^{\theta}$ with $\theta = 0.5$ and $N_i \sim k_i^{\phi}$ with $0.5 \leq \phi \leq 1.5$, where k_i is the traffic at the origin and k_j at the destination. Values for θ and ϕ are computed empirically.

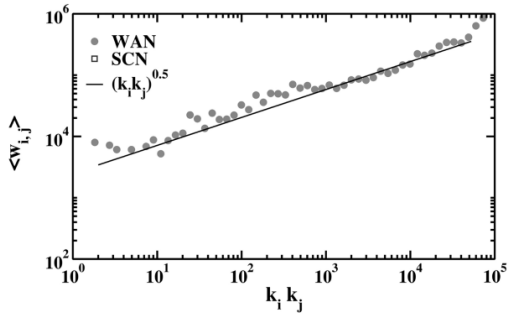


Figure 6.2: Whole segment network worldwide, 2002. Scaling relation.

Another type of mobility one may encounter is the so called **commuting**. This information is obtained from census of different countries and it mainly deals with locations of residence and work. For this reason, *spatial resolution* is highly *variable* and depends actually on the country: local/regional administrations can be actually organized differently within states. Moreover, we can make the aforementioned graphical analysis also for commuting network.

We want now to point out the **differences** between the *air travelling* and the *commuting* networks. In order to do so, first we obviously need to use the same spatial resolution: this is why researchers defined **macro urban areas** centered around airports. Now we are able to look at what differs one network from the other. The first feature one may want to analyze is the average **daily number of travellers** in a certain area: it is about 1000 for the air travel, while 20'000 for the commuting network. Order or magnitudes are indeed different. Another important point is the **fraction of daily travellers**: it is defined as the probability to travel per time unit,

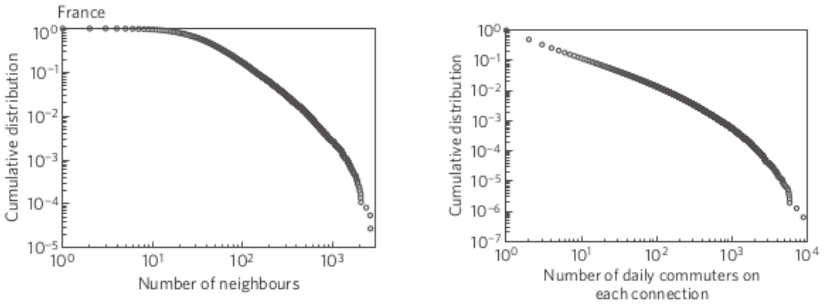


Figure 6.3: Topology of the network and flux of passengers graphical analysis for commuting network.

namely the *daily* flux of travellers normalized to the catchment population of an area assuming that all people have the same probability to travel. Empirically, we observe that is more probable to commute ($10^{-2} days^{-1}$) rather than air travel ($10^{-3} days^{-1}$). Finally, **time scales** for the two processes are different. One should first define what is a *time scale*: we can refer to it as the *length of the stay in the destination*, or more generically as the *time elapsed from one trip to another*. This quantity when one travels by plane is usually of order days/weeks, whereas commuting is a matter of hours since at workplace we spend order of hours. In conclusion, **commuting** has a **faster dynamics** and leads to a **higher** level of **mixing**.

One another type of data that is available is the one shared privately by **telephone** providers. Information is recorded for each call and/or SMS: for instance time, caller ID, recipient ID, call duration and cellular tower. Using the position of the latter ones one may be able to reconstruct individual level trajectories. Obviously, for privacy, users are anonymized. Some **challenges** that may arise using this kind of data are the following: for **statistical reliability** the analysis is restricted to users that call *more frequently*, but this actually can turn out to be not precise since many locations can still be missed. Another issue is that the area covered by the cell tower is highly variable: towers are more dense in densely populated area, whereas the **spatial resolution** is *rural area* is very **poor**. This kind of data is really important and accurate: one can reconstruct individual trajectories and the mean of transport used and, eventually, why. For many low income countries this is actually the **main source of information** regarding mobility, despite it is available worldwide. Some drawbacks are that statistically this data must be treated carefully and poses statistical challenges, also from a numerical point of view. Moreover, data cannot be shared across groups for a matter of validation.

Other data can be collected from **GPS**, exploited by some apps or from research projects. This comes with the greatest level of accuracy on movement trajectories: **spatial resolution** is about few *meters* and **temporal resolution** is about *seconds*. However, devices with GPS are a really small subset (10^3) compared to $\sim 10^6$ mobile phones. Data can be collected by other **mobile application services** (e.g. Google, Twitter, Facebook...), it can give high spatial resolution, being it based on GPS, but the population may not be representative. Note that, because of some special events, data can be **donated**: Google, Apple have been sharing their data for good initiatives in helping against the fight or COVID19. An other **historical way** to collect migration was to trace the position of some US Federal banknotes [2]: their trajectories are likely a convolution of the mobility of several individuals. Finally, annual information of residence from individual tax return files in the US can describe **migration**, which has a time scale that spans over years.

[2] www.wheresgeorge.com

As one may imagine **data** is very **heterogeneous**, being heterogeneous the sources where we collect it from. Heterogeneity can be seen in spatial resolution, individuals-level/origin-destination fluxes/seats, broken down per transportation media or per purpose of the trip. Since every dataset provides **partial information**, one may think to try to **combine** some of them. For instance, this cannot be done for air-travel and commuting network, being the spatial ranges very different. In addition, combining cell-phones data and commuting we are able to extract commuting proxies from cell-phone data. Finally, we should state that we cannot spot clearly the differences that occur for people of different ages: among air travellers indeed there are few children and old people, and statistically they look like the same and cannot be distinguished.

6.1.2 Modelling Human Mobility

Let us now discuss what models all the data collected so far can lead to. There are many types of model we can think of, starting from *individuals-level* models or *population-level* ones. As one can imagine, in the **individuals-level** models we model trajectories of individual using mathematical tools that might include stochasticity: random walk, brownian motion, Levy flight or preferential return, but there are actually many others.

Regarding instead **population level models** we try to model fluxes of individuals, therefore adding some layers of abstraction and generalization: we want to find for instance the *Origin-Destination* matrices. There are two main families for these kind of models: **gravity models**, or **intervening opportunity models**.

The **Gravity Model** was first introduced by G.K. Zipf. He took inspiration from Newton's law of gravitation in order to describe **mobility flows**:

$$T_{ij} \propto \frac{N_i N_j}{d_{ij}} \tag{6.1}$$

where N_i is the population in i-th site and d_{ij} is the distance between nodes i and j. The last formula can be written in a more general formula as it follows:

$$T_{ij} = C M_i M_j F(d_{ij}), \qquad M_i = N_i^{\alpha} \quad M_j = N_j^{\gamma} \tag{6.2}$$

where $F(d_{ij})$ is a general function that is either a power law of the kind d_{ij}^{β} or exponential form $e^{-\beta d_{ij}}$ or a combination of both. This model actually permits to fit very well the data, despite there are no general values for the fitting parameters: they do vary according to the spatial granularity. A possible workaround for this problem was to introduce the formula of *Balcan et al PNAS 2009* ($T_{ij} = C\frac{N_i^{\alpha} N_j^{\gamma}}{e^{\beta d_{ij}}}$) which was fitted to 29 different countries spread across all continents. The main **result** was that, when data is aggregated at the *same level of spatial resolution*, the same parameters are able to model well the mobility fluxes in all countries.

Another historical model for the mobility was the so called **radiation model** (see Fig. 6.4). It was introduced by Stouffer (1940). He noticed that the *key* driver of migration was the number of **intervening opportunities** or the **cumulative number** of opportunities between the origin and the destination. However, definition of *"opportunities"* was intentionally left vague and they assume a different meaning with respect to the system we are dealing with.

Resulting fluxes in this way are independent of $p(z)$ and are parameters free:

$$T_{ij} = O_i \frac{1}{1 - \frac{N_i}{M}} \frac{N_i N_j}{(N_i + S_{ij})(N_i + N_j + S_{ij})} \tag{6.3}$$

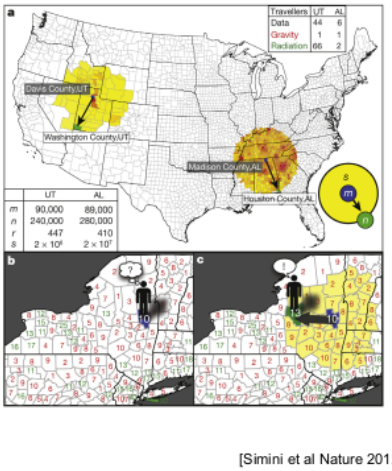


Figure 6.4: Graphical description for the radiation model. Individual is more likely to move where opportunities are more, since we use their number as a metric.

where S_{ij} is the population in the radius d_{ij} and $M = \sum_i N_i$. The advantage that we do not need any parameters in order to use this model: it is useful in epidemiology when the only information we have is the population distribution (it usually happens for low developed countries). However, the goodness of fit depends on the spatial resolution.

6.2 Integrating Human Mobility in Epidemic Models

Now we want to use the knowledge we have acquired so far to describe better how Human Mobility affects epidemic spreading. In order to do it, we will borrow from ecology the concept of **metapopulation models**. The last was introduced to study the interplay between stochasticity and spatial heterogeneities[3]: the entire population was divided in **patches** which are discrete entities (see 6.5). It follows that there may be two different levels of mixing: *local*, that occurs within a patch, or *global* that occurs among patches. It is a **coarse grained** description, and patches can be seen as the new elementary units for our network.

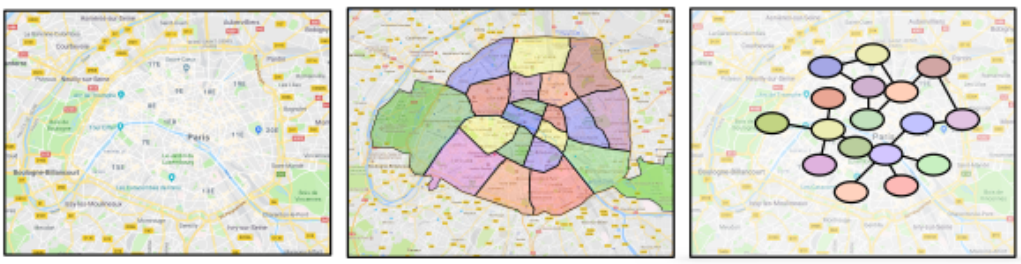


Figure 6.5: Population is divided in discrete entities, the so called *patches* that will become our new elementary units when dealing with network.

In ecology the dynamics is therefore driven by stochastic effects, which may lead for instance to extinction or recolonisation. This, obviously, will suggest us to find an analogy when dealing with epidemic models. However, one should keep into account that the **discrete nature of individual** is one of the most essential ingredients to describe the dynamics: it is meaningless to state that half an individual travels

[3]Levins Bull. Entomol. Soc. Am., 15 (3) (1969).

between two patches. The first models assumed that the mixing between patches occurred homogeneously, while more recently more complexity has been added: **mixing** among patches has started to be **mediated** by the **human mobility network** hence coupling the metapopulation perspective with network theory.

6.2.1 SIR metapopulation model

Let us discuss now how we can introduce the metapopulation concept into a model we have studied so many times: the *SIR* model (see fig. 6.6). The only difference here is that it is **not** an **individual-based** model any more. now we do not keep track of every indivudal, but we just monitor the occupation number of patches and compartments.

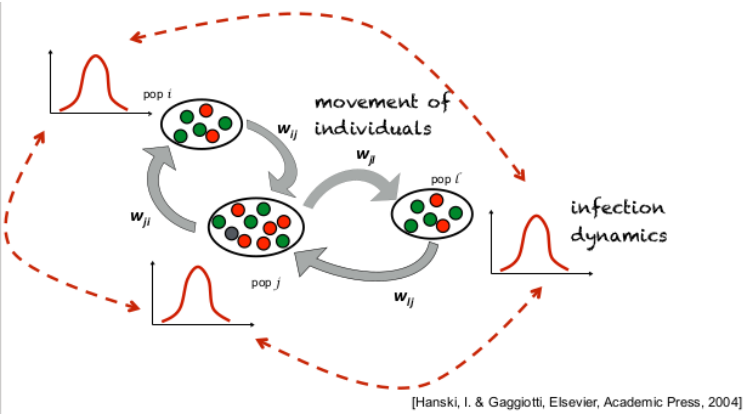


Figure 6.6: How we can model mobility and transmission dynamics using SIR metapopulation model

For each i-th patch we can define the following variables, as we did for the entire population before: $S_i(t)$, $I_i(t)$, $R_i(t)$, and obviously it holds that $N_i(t) = S_i(t) + I_i(t) + R_i(t)$ with clear meanings for each of them. Note that, given a total number of V patches, it follows the definition of global variables:

$$S(t) = S_1(t) + S_2(t) + S_3(t) + ... + S_V(t) = \sum_i S_i(t) \tag{6.4a}$$

$$I(t) = I_1(t) + I_2(t) + I_3(t) + ... + I_V(t) = \sum_i I_i(t) \tag{6.4b}$$

$$R(t) = R_1(t) + R_2(t) + R_3(t) + ... + R_V(t) = \sum_i R_i(t) \tag{6.4c}$$

$$N(t) = N_1(t) + N_2(t) + N_3(t) + ... + N_V(t) = \sum_i N_i(t) \tag{6.4d}$$

The set of equations related to i-th node and the compartments is the following:

$$\frac{\mathrm{d}S_i}{\mathrm{d}t} = -\beta \frac{I_i(t)S_i(t)}{N_i} + \Omega_i^S \tag{6.5a}$$

$$\frac{\mathrm{d}I_i}{\mathrm{d}t} = \beta \frac{I_i(t)S_i(t)}{N_i} - \mu I_i(t) + \Omega_i^I \tag{6.5b}$$

$$\frac{\mathrm{d}R_i}{\mathrm{d}t} = \mu I_i(t) + \Omega_i^R \tag{6.5c}$$

where we have introduced Ω_i^X that is a measure of the *in-flow* or *out-flow* of people in compartment X. Note as the set of equations really resembles the SIR model we previously studied.

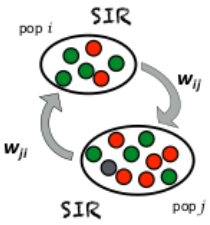


Figure 6.7: Mobility between two different patches occurs with probability

We will now discuss how to compute Ω_i^X for which we need to model human mobility. The first assumption we make is that the mobility is a **Markovian process**. Indeed, this is the easiest model one can think of. We need now to model human mobility, and this can be done in the following way: we now that N_i people live in i-th patch, and w_{ij} is the number of people that travel $i \to j$. Therefore, the **probability** for an individual in i **to travel** from i to j is:

$$p_{ij} = \frac{w_{ij}}{N_i} \tag{6.6}$$

The simplest possible model is when p_{ij} is the same for all individuals **regardless** their *infectious status* (S,I,R) and their *travel history*. This is the Markovian assumption we stated above. As soon as an individual enters into a new population, she mixes completely within it and cannot be distinguished from other individuals any more. Moreover, she will be considered as part of that population from now on.

Travelling is a **binomial process**. The average number of individuals in compartment X in i travelling from i to j at each t is:

$$\left\langle T_{ij}^X \right\rangle = p_{ij} X_i(t) = \frac{w_{ij}}{N_i} X_i(t) \tag{6.7}$$

Therefore, a formula for Ω_i^X can be the following:

$$\Omega_i^X = \sum_j \left(\frac{w_{ji}}{N_j} X_j - \frac{w_{ij}}{N_i} X_i \right) \tag{6.8}$$

where, for the flux, we consider all the people that are entering patch i from all possible other patches j, as well as all individuals exiting the node i being their destination any other patch j.

We now recall the **assumptions** we have done so far. We have modelled mobility as a **Markovian process**. In other words, we assume that travellers mix with the population at destination and forget about travel origin. The implications hence are:

- **travel trajectory** is *random*: patch $i \to$ patch $j \to$ patch $l \to$ ...;
- we do *not* take into account the **residence location**;
- we do *not* take into account the *length of stay* while travelling.

At the end of the day we are modelling a **migration process**. The Markovian assumption for mobility actually works well as long as **travels** are **not frequent**, that is to say that travelling rate is *negligible* wrt epidemic time scales ($p_{ij} \ll \mu$). Moreover one should choose the Markovian assumption when we want to model the **short term dynamics** of an epidemic. In the real world, some situations where this hold at a first approximation are for instance:

- air-travel and acute infections: for example flu or COVID 19. We have seen that travelling rate is about $10^{-3} days^{-1}$ that way smaller that their recovery rate $10^{-1} days^{-1}$;
- the early spread of COVID-19 or a flu pandemic. However, it does not work when we want to model the spreading in the long run (i.e. for large times).

Lecture 15. *Thursday 19th November, 2020. Compiled: Saturday 14th August, 2021.*

6.3 Application of metapopulation models

In the last lecture we have introduced the so called **SIR metapopulation model**, deriving it and understanding it *analytically*. Now we want study it further, focusing on the **spatial propagation** and **predictability** under the assumption of Markovian mobility. Note that beside it, there might be many other assumptions we can make to model mobility.

6.3.1 Spatial propagation dynamics

Let us discuss the **dynamics** of spatial spreading when we are *above* the **epidemic threshold**. Given that an epidemic has started in a given city i, we want to understand how it will spread to j, h... Let us define the **seeding time** (or *arrival time*) as it follows: it is the time of arrival of the *first* case in patch j. Let us focus on a 2 patches model (see fig 6.8). We consider that the travel events between patches $i \to j$ occur as instantaneous jumps (of probability p for each individual) at discretized times, and $I(t)$ is the number of infectious in patch i at time t. The **probability** that any infected individual arrives in j at time step n is $\left[1-(1-p)^{I(n\Delta t)}\right]$.

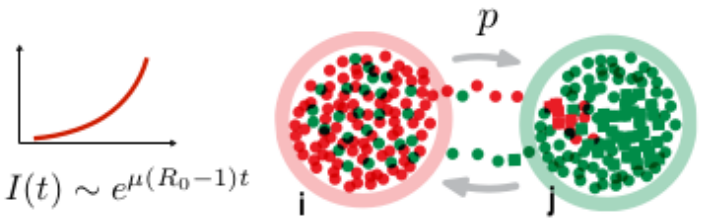


Figure 6.8: Dynamics of spatial spread: p is the traveling probability $i \to j$, and $I(t)$ is the prevalence in patch j at time t.

Consequently, the probability that the *first infectious* individual arrives at time $t_{\text{seeding}} = t = n\Delta t$ in city j is then given by:

$$P(t_{\text{seeding}} = n\Delta t) = [1-(1-p)^{I(n\Delta t)}] \times \prod_{i=1}^{n-1}(1-p)^{I(i\Delta t)} \tag{6.9}$$

which expresses the fact that at least one successful "jump" from i to j of an infectious individual occurs at time $n\Delta t$, and none at previous times. In order to obtain the probability density of the arrival time in city i, we first note that in real world systems the number of travelers is usually small with respect to the total population of a city: $p = w\Delta t/N \ll 1$. In this limit for $p \to 0$, we can rewrite:

$$P(t_{\text{seeding}} = t) = pI(t)e^{-p\sum_{0<i<n} I(i\Delta t)}$$

Using the standard approximation of $\Delta t \sum_{0<i<n} I(i\Delta t) \to \int_0^t I(\tau)d\tau$:

$$P(t_{\text{seeding}} = t) = pI(t)e^{-p\int_0^t I(\tau)d\tau} \tag{6.10}$$

Defining now $a = \mu(R_0 - 1)$ and knowing that $I(t) = e^{at}$ at early stages, we can rewrite the probability as:

$$P(t_{\text{seeding}} = t) = pe^{at}e^{-\frac{p}{a}e^{at}} \tag{6.11}$$

That is a **Gumbel distribution**, whose expected value is:

$$\langle t_{\text{seeding}} \rangle \simeq -\frac{1}{a} log\left(\frac{p}{a}\right) \tag{6.12}$$

This is the expected time needed for the infection, starting in i, to end up in j for a 2 patches model. Let us consider now the generalization to a **chain** of identical patches (see fig 6.9), denoting the population as N and the traveling weight among patches p.

Now it occurs that there is **correlation** among patches, since the time at which the patch i is infected does depend on the previous one, that in turn depends on

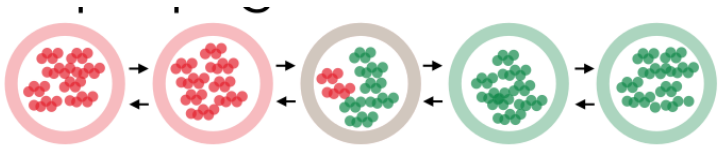


Figure 6.9: Dynamics of spatial spread: p is the traveling probability among patches, that are now more than two.

the one before and so forth. Infected people that travel between patches are actually important only at the starting of each infection process for every patch, but later these can be neglected since their contribution is small compared to the normal spreading dynamics. We can define as Δ_i the interval that interoccurs between two consecutive seedings, obviously related to different patches:

$$\langle t_{\text{seeding},i} \rangle - \langle t_{\text{seeding},i-1} \rangle = \Delta_i, \qquad \langle t_{\text{seeding},n} \rangle = \sum_{i=1}^{n} \Delta_i \tag{6.13}$$

Consequently Δ_i are correlated and not identically distributed: incidence dynamics in city i is mainly given by two contributions. The first is a term related the introduction of the infection from $i-1$-th patch (travelling rate, infected individual travelling...), while the second one is given by the infection transmission dynamics within patch i itself. Every patch is therefore correlated to the previous ones. However, the simplest approximation one can think of is:

$$\langle \Delta \rangle = \langle t_{\text{seeding},1} \rangle$$

which can be shown to not work so bad[4].

It is straightforward to introduce a **metric** that describes how nodes are close to each other. It is function of connection "weights" (see Fig. 6.10) that are in turn given by the number of travellers, of flights and connections are present between a pair of nodes. In this way we are able to introduce an **effective distance**[5] $-\ln(p_{ij})$ between nodes i and j: despite the geographical distance is way larger, it follows that New York and London are actually closer than New York and any other rural place in the US Midwest using this metric.

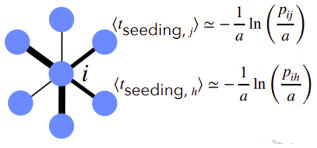


Figure 6.10: Effective distance between node i and other nodes connected to it. The larger the edge, the stronger the connection weight.

So, the existence of **pathways**[6], through which it is more likely that the spread of a disease can occur, comes in helpful when we are requested to make risk assessment analysis for some regions. Indeed, it allows us to better **predict** the possible path through which the evolution of a spreading might occur. We now introduce the *overlap function* $\Theta(t) \in [0,1]$ that describes the similarity between 2 outbreak realizations among different numerical simulations: the closer to 1, the more similar they are. Obviously, the **higher** the **overlap**, the **higher** the **predictability**.

As one can clearly understand from Fig. 6.11, introducing the **degree heterogeneity** decreases the predictability: in this way we are including hubs in our model.

[4] Gautreau et al JTB 2008.
[5] Brockmann, Helbing, Science 2013.
[6] Colizza, Barrat, Barthelemy et Vespignani, PNAS (2006).

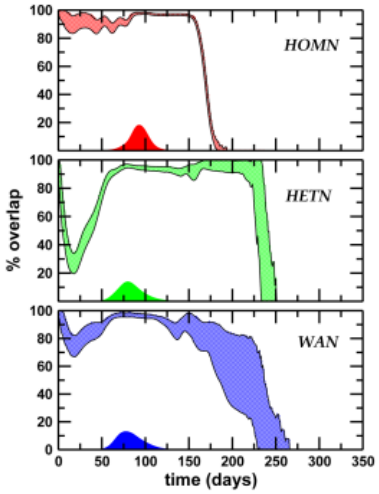


Figure 6.11:
Top: numerical simulations with no degree fluctuations and no weight fluctuations.
Middle: numerical simulations with only degree fluctuations and no weight fluctuations.
Bottom: numerical simulations with both degree fluctuations and weight fluctuations.

On the other hand, if we consider also **pathways** hence bringing in **weight fluctuations** numerical simulations are more likely to return similar outputs (the predictability is increased).

One of the most important questions when dealing with epidemics and mobility is whether introducing **travel bans** would allow us to gain some time wrt the spreading of the disease. In other words we want to see if a restriction of traffic would be **effective** in either containing or delaying the propagation[7]. We therefore *rescale* the travelling probability by a factor of ω:

$$\langle t_{\text{seeding},T.R.} \rangle \simeq -\frac{1}{a} \ln\left(\frac{p\omega}{a}\right) \tag{6.14}$$

We can now compute how much time we "gain" when we introduce some traffic restrictions:

$$\langle t_{\text{seeding},T.R.} \rangle - \langle t_{\text{seeding}} \rangle \simeq -\frac{1}{a} \ln\left(\frac{p\omega}{a}\right) + \frac{1}{a} \ln\left(\frac{p}{a}\right) = -\frac{1}{a} \ln(\omega) \tag{6.15}$$

where $0 < \omega \leqslant 1$. This result is indeed really important: to have a **consistent delay** in the propagation, we must **cut** the **flights** about **order of magnitudes**, with all possible consequent social and economical implications since the dependence is logarithmic. Therefore, the moral is that it is way *better to help the metapopulation where the epidemic started*, rather than cutting flights, being it quite useless. This would mean to try to decrease the slope of the exponential growth in the patch where the disease originated. For instance, during $H1N1$ pandemic in 2009, traffic volume from Mexico to Europe decreased by a factor of 40%, and the only gain in delay was really poor (order of days). Indeed, the same result is obtained thanks to numerical simulations with a global spreading model for influenza and it turn out that the delay is negligible.

We want now to briefly discuss how these arguments can be applied to **epidemic assessment**, in particular wrt **COVID-19** situation[8]. One of the first early warnings emanated by the Wuhan Municipal Health Committee was the following: *"urgent*

[7] Gautreau et al JTB 2008; Hollingsworth et al Nature Med 2006; Scalia Tomba et al Math Biosci 2008.

[8] `https://www.nytimes.com/interactive/2020/03/22/world/coronavirus-spread.html`.

notice on the treatment of pneumonia of unknown cause". This was actually picked up by ProMED-mail, which is an independent program for emerging diseases, that raised an alert to the situation in Wuhan at the beginning of January. Later, some cases of pneumonia were discovered far from the epicentre among travellers, namely 2 in Thailand and 1 in Japan: compared to the 40 *local* cases and to the traffic volume (i.e. p to travel) it was really out of scale: local cases should have been many more. This argument was actually developed by the Imperial College, stating that[9] this new pneumonia had already started spreading worldwide and provided an estimate for the number of cases. We want now to follow the argument developed by the just mentioned report. Clearly, the number of infected that travel is $I_{trav} = I_{Wuhan} \cdot p_{travel}$, where the probability that an infectious travels p_{travel} is the product between the daily probability to travel p_{daily} and the time to detect a case T_d:

$$p_{travel} = p_{daily} T_d \tag{6.16}$$

where p_{daily} is the probability of travelling out of Wuhan, namely $p_{daily} = \frac{w}{N}$ and the time to detect a case T_d is the sum of two terms: *incubation period* and *time to hospitalization* (first cases were detected only after hospitalization, i.e. when symptoms started to be severe). Using real parameters, namely the passenger per day at Wuhan airport $w = 3301$ $person/day$, the catchment population of the just mentioned airport $N \sim 10^6$ $individuals$ and the incubation period $5-6$ $days$ and time to hospitalization $4-5$ $days$ we could infer that the estimated cases must have been around 1800 ($95\ \% C.I. : 427-4471$) individuals. This is much more than the 40 cases officially detected. After that particular moment, with some delay, WHO raised an alert, too, and surveillance was heightened in foreign countries. Since mobility data among countries is actually more reliable, we can use them in order to infer the epidemiological situation in the seed country.

A tool used in **numerical simulations**, for the worldwide spread of epidemics, we may ever encounter is the website `http://www.gleamviz.org/`. GLEaM[10] is indeed a site that allows us to both visualize and simulate numerically the evolution of a pandemic using different data mobility and networks.

6.3.2 SIR metapopulation model with memory

Let us now discuss the **SIR metapopulation** model in a different regime: **commuting**[11] [12], and not air travel any more. We therefore **drop** the **assumption** that the traveling rate is *negligible* with respect to the epidemic time scale, i.e. $p_{ij} \ll \mu$. The interoccurrence time between two trips, from now on, will be of order of *hours*: much less than the recovery rate.

In general, treating mathematically the interplay between mobility and transmission is very difficult. The problem can be solved by using **time scale separation**, that works as a sort of *mean field approximation.* Hence, either the **epidemic unfolds faster than mobility**, that is the case of flu ($\mu \sim 10^{-1}$ $days^{-1}$) and air traveling (rate $\sim 10^{-3} days^{-1}$) or alternatively **mobility is faster than the epidemics**, that is the case of commuting (usually we travel twice a day, its rate is $\sim 3 days^{-1}$) and flu ($\mu \sim 10^{-1} days^{-1}$). **Timescale** in *commuting case* is mainly given by travel duration, whereas the **probability** p instead determines the fraction of people commuting, i.e. the fraction of workers/students that daily travel vs pensioners that do not.

[9] `https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-1-case-estimates-of-covid-19/`

[10] **Gl**obal **E**pidemic **a**nd **M**obility Model.

[11] Sattenspiel, L. & Dietz, K. Math. Biosci. 128, 71–91 (1995)

[12] Keeling, M. J. & Rohani, P. Ecol. Lett. 5, 20–29 (2002)

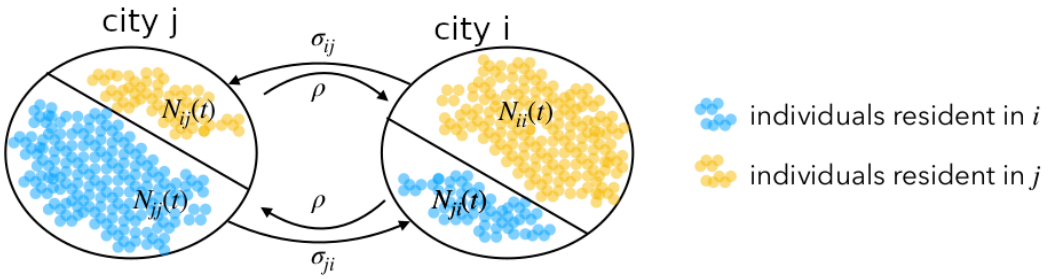


Figure 6.12: Graphical representation for the SIR metapopulation model with memory, this is actually more realistic since we usually back and forth in such trips home↔workplace. We distinguish the population inside patches on their residence place.

We want now to take into account that, when **commuting**, we usually **return** to the place where we started. A typical trip in this case would be $i \to j \to i$. Nothing however prevents us to leave once again for k-th patch as soon as we are back to i, but this would not not commuting any more. The first modification we make is to divide every patch (here for simplicity we show only two in Fig. 6.12) distinguishing on the individuals and whether they are resident in i or j. In this way we can consider people that, despite they live in another patch i, travelled to patch j with a certain **leaving rate** σ_{ij} and will return home according to a **returning rate** $\rho^{-1} = \tau \approx 8h$. that is independent of the destination. $N_{ij}(t)$ are the individuals resident in i and traveling to j, while we approximate the **population** for a given patch i as *constant*:

$$N_i = N_{ii}(t) + \sum_j N_{ij}(t) \tag{6.17}$$

The latter is the expression that describes the number of people **resident** in patch i. In other words we are assuming no immigration and no births nor deaths. In this way we can keep track of from where and individual left.

We want now to consider the the change in time of people resident in i and either staying in i or leaving for j:

$$\partial_t N_{ii}(t) = -\sum_j \sigma_{ij} N_{ii}(t) + \rho \sum_j N_{ij}(t) \tag{6.18a}$$

$$\partial_t N_{ij}(t) = \sigma_{ij} N_{ii}(t) - \rho N_{ij}(t) \tag{6.18b}$$

The linear ordinary **differential equation** of order 1 to be solved can be rewritten, replacing $\sum_j N_{ij}(t)$ from 6.17 into $\partial_t N_{ii}(t)$, as:

$$\partial_t N_{ii}(t) + (\rho + \sigma_i) N_{ii}(t) = N_i \rho \tag{6.19}$$

The first solution is:

$$\begin{aligned} N_{ii}(t) &= e^{(\rho+\sigma_i)t} \left(C_i i + N_i \rho \int_0^t e^{(\rho+\sigma_i)s} ds \right) = \\ &= \frac{N_i}{1+\sigma_i/\rho} + \left(N_{ii}(0) - \frac{N_i}{1+\sigma_i/\rho} \right) e^{-\rho(1+\sigma_i/\rho)t} \end{aligned} \tag{6.20}$$

While for the individuals resident in i and travelling to j:

$$\begin{aligned} N_{ij}(t) = \frac{\sigma_{ij} N_i/\rho}{1+\sigma_i/\rho} - \frac{\sigma_{ij}}{\sigma_i} \left(N_{ii}(0) - \frac{N_i}{1+\sigma_i/\rho} \right) e^{-\rho(1+\sigma_i/\rho)t} + \\ + \left[N_{ii}(0) - \frac{\sigma_{ij} N_i/\rho}{1+\sigma_i/\rho} - \frac{\sigma_{ij}}{\sigma_i} \left(N_{ii}(0) - \frac{N_i}{1+\sigma_i/\rho} \right) \right] e^{-\rho t} \end{aligned} \tag{6.21}$$

More important quantity to be defined is the **time of relaxation** to the equilibrium, which is defined as τ and dominated by:

$$[\rho(1+\sigma_i/\rho)]^{-1} \sim \rho^{-1} = \tau \qquad \text{since } \rho \gg \sigma_i \tag{6.22}$$

where the probability of leaving i regardless the destination is $\sigma_i = \sum_j \sigma_{ij}$ Therefore, the **equilibrium solutions** one may find are:

$$N_{ii} = \frac{N_i}{1+\sigma_i/\rho}, \qquad N_{ij} = \frac{\sigma_{ij}N_i/\rho}{1+\sigma_i/\rho} \tag{6.23}$$

Since also in the **steady state** number of people that are *resident* in i is conserved, then we can compute the number of people **present** in i as:

$$N_i^* = N_{ii} + \sum_j N_{ji} = \frac{N_i}{1+\sigma_i/\rho} + \sum_j \frac{N_j\sigma_{ji}/\rho}{1+\sigma_j/\rho} \tag{6.24}$$

The ratio actually σ_i/ρ quantifies the proportion of time spent outside and in the residence population.

Let us discuss now some **limiting cases**:

- $\sigma_i \to 0 \implies N_{ii}(t) \to N_i;\ N_{ij}(t) \to 0;\ N_i^* \to N_i$ in this case people rarely leave their residence, thus the non travelling individuals approach the population of residents;
- $\rho \to \infty \implies N_{ii}(t) \to N_i;\ N_{ij}(t) \to 0;\ N_i^* \to N_i$ people return home immediately thus the non travelling individuals approach the population of residents;
- $\rho \to 0 \implies N_{ii}(t) \to 0;\ N_{ij}(t) \to \frac{\sigma_{ij}}{\sigma_i}N_i;\ N_i^* \to \sum_j \frac{\sigma_{ji}}{\sigma_j}N_j$ here the case collapses to **migration** process: people never get back and the population of resident in i is distributed among the neighbouring destinations j.

One should recall now that the time scale of commuting $\tau \approx 8\,\text{h}$, while the duration of an acute infection (i.e. flu) is way more: $\mu^{-1} \approx [1-3]\,\text{days}$. We could adopt, as said, a *mean field* paradigm: we assume that the person can be partially in a place and partially in another one: on average it happens that he might be at the same time in different places[13]. **Transmission dynamics** is therefore **slower than mobility**: we can assume that compartments occupation numbers are at the equilibrium with respect to mobility dynamics. Formally, the **occupation numbers** are:

$$X_{ii}^{[m]} = \frac{X_i^{[m]}}{1+\sigma_i/\rho}, \quad X_{ij}^{[m]} = \frac{\sigma_{ij}X_i^{[m]}/\rho}{1+\sigma_i/\rho}, \quad X^{[m]} = S, I, R \tag{6.25}$$

We want now to understand how many infected individuals a susceptible person resident in i is exposed to, that is to say we want to compute the **force of infection** λ:

$$\lambda = \beta\frac{I(t)}{N(t)} \tag{6.26}$$

This parameter is present in the set of differential equations related to compartments numbers:

$$\partial_t S = -\beta\frac{I(t)}{N(t)}S(t) \tag{6.27a}$$

$$\partial_t I = \beta\frac{I(t)}{N(t)}S(t) - \mu I(t) = \lambda S(t) - \mu I(t) \tag{6.27b}$$

$$\partial_t R = \mu I(t) \tag{6.27c}$$

[13] in terms of reality it is meaningless to state that fractions of person are present in different places, unless we are speaking about Voldemort and his Horcruxes. Aside jokes, Mean Field approximation allows us to attach a meaning to these fractions since we are considering them "on average".

Instead of explicitly modelling mobility, we now directly compute the effects of the other patches on the risk of infection, i.e. we break down the force of infection in its different contributions. Having assumed we are at equilibrium, S_i is distributed among patch i and all possible destinations j according to the set of proportions:

$$\left\{ \frac{1}{1+\sigma_i/\rho}, \dots, \frac{\sigma_{ij}/\rho}{1+\sigma_i/\rho}, \dots \right\}$$

whereas the **risk of infection** has the form:

$$\lambda_i = \frac{\lambda_{ii}}{1+\sigma_i/\rho} + \sum_j \frac{\lambda_{ij}\sigma_{ij}/\rho}{1+\sigma_i/\rho} \tag{6.28}$$

where the first contribution is the contribution of people that stay at home, and the second is the contribution of people infected that visit patch i from any other patch j. In the formula above we have that λ_{ii} is the force of infection that a person resident i experiences from infected people that are in i, formally:

$$\lambda_{ii} = \frac{\beta_i}{N_i^*}\left[I_{ii} + \sum_j I_{ji} \right] = \frac{\beta_i}{N_i^*}\left[\frac{I_i}{1+\sigma_i/\rho} + \sum_j \frac{I_j\sigma_{ji}/\rho}{1+\sigma_j/\rho} \right] \tag{6.29}$$

Whereas λ_{ij} is the contribution of a susceptible resident in i that is travelling to any other patch j and there is exposed to infected people:

$$\lambda_{ij} = \frac{\beta_j}{N_j^*}\left[I_{jj} + \sum_l I_{lj} \right] = \frac{\beta_j}{N_j^*}\left[\frac{I_j}{1+\sigma_j/\rho} + \sum_l \frac{I_l\sigma_{lj}/\rho}{1+\sigma_l/\rho} \right] \tag{6.30}$$

Note as some of the terms are similar to the steady state ones in equation 6.24. These last expressions are actually useful in **numerical simulations**, and they allow us to understand the relative role of mobility and infection parameters on the epidemic dynamics in order to speed up the simulations, too. One should actually remember that both *infections* and *travels* are **stochastic processes**, and last results hold only when mobility is faster than the epidemics time scale.

Lecture 16.
Friday 20th November, 2020.
Compiled: Saturday 14th August, 2021.

6.4 Global Invasion Threshold

Following the formalism introduced for the SIR metapopulation model and markovian mobility, we want now to analytically derive the **epidemic threshold** for metapopulations. The latter is the *threshold* for a pathogen that, when overcomed, allows it to spread among the entire population. In other words, we want to find the **conditions** for a **local outbreak** to **spread** at a **global** scale.

In order to pursue our goal, we make the so called **coarse graining**: starting from the formalism typical of a single subpopulation we decrease the number of degrees of freedom by making some averages or finding some more general rules. Quantities must therefore be scaled accordingly: R_* is the correspondent to R_0 for metapopulations (if $R_* > 1$ we observe an outbreak), and the typical **timescale** is now the *duration of* an *outbreak* in a patch. Hence, we follow the spread from one subpopulation to another, by the mean of **mapping**[14] [15] the **spreading dynamics among subpopulation** into the spreading on a **network**.

We are indeed approximating as a **branching process** and try to follow the invasion dynamics at the subpopulation level, denoting by D^n the diseased subpopulations at the n-th generation. Hence, we do not follow the dynamics in term of times, but in terms of *generation*.

[14] Colizza & Vespignani, PRL 2007, JTB 2008
[15] Cross, et al. JRSoc Interface 2007

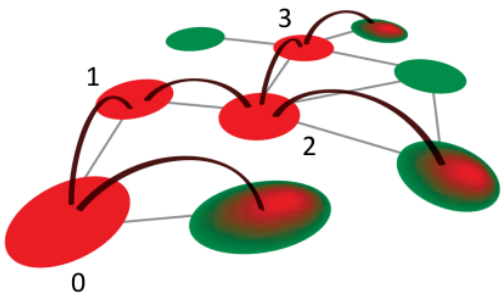


Figure 6.13: Dynamics of spatial spread at subpopulation level and the generation time n they contract the disease.

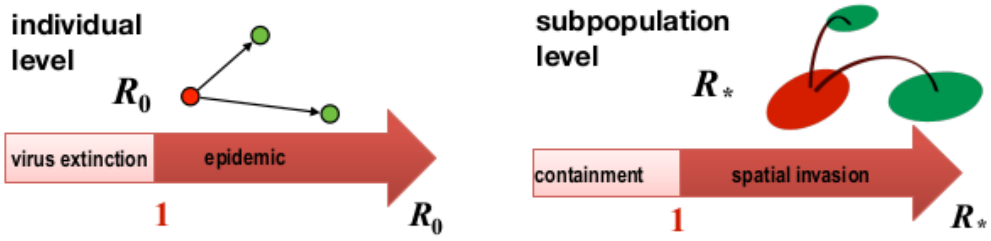


Figure 6.14: **Left:** Dynamics of spread on a individuals-in-a-network level. **Right:** Dynamics of spreading among subpopulation, once we made a *coarse graining*. Graphically, they look like similar, despite the change in value, and meaning, of parameters as output of the mapping. R^* is the analogous of the basic reproductive ratio R_0 at metapopulation level.

6.4.1 Homogeneous networks

Let us assume that we are dealing with **homogeneous systems**, for the sake of simplicity, and with only 2 patches: namely j and i as one can see in Fig. 6.15.

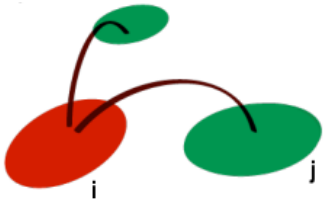


Figure 6.15: Dynamics of spatial spread at subpopulation level, but considering only two patches.

The variables one may want to introduce now are the following ones: w_0 is the **number of travellers** along each link, whereas $\langle k \rangle$ is on average the **number of connection** of each subpopulation, which coherently with our notation has **population** N. Let us introduce now α, that is the **epidemic attack rate**: it is the *"final size"* of the epidemic within each patch. We recall that the latter depends on the basic reproductive ratio R_0.

Using these quantities we can compute the probability of early extinction of the disease, that is the probability for the disease to be not able to spread to other patches once it started in patch i:

$$P_{ext} = \left(\frac{1}{R_0}\right)^{\lambda_{ij}} \tag{6.31}$$

where R_0 is the usual term and λ_{ij} is the total number of infectious individuals sent

from i to j during the local outbreak:

$$\lambda_{ij} = \frac{w_0}{\cancel{\mathcal{N}}} \frac{\alpha \cancel{\mathcal{N}}}{\mu} \tag{6.32}$$

Note that it is the product of two terms: the first factor is the probability to travel, while the second is the ratio between the total number of people that get the infection rescaled by the recovery rate μ. We recall taht α is the maximum value of the prevalence $i = I/N$ within a patch.

The number of diseased subpopulation at generation n, namely D^n, is computed iteratively from the number of diseased patches at generation D^{n-1}:

$$D^n = (\langle k \rangle - 1)(1 - P_{ext}) \left(1 - \sum_{m=0}^{n-1} \frac{D^m}{V}\right) D^{n-1} \tag{6.33}$$

where V is the number of patches. Moreover, in the factor $(\langle k \rangle - 1)$, the -1 is present in order to ignore the patch we have got the infection from. In addition, the one highlighted is the probability that the patch is disease free at $n-1$-th generation. The last factor D^{n-1} is the number of diseased patches at the previous generation: as one can imagine, if the remaining factor is **larger** than 1, then the epidemic has an outbreak, otherwise it gets extinct. Indeed:

$$D^n = R_* \left(1 - \sum_{m=0}^{n-1} \frac{D^m}{V}\right) D^{n-1}$$

where the condition for a global outbreak is $R_* > 1$.

Accordingly to what we have just said, the factor R_* is therefore defined *at the beginning of the infection*, i.e. when all patches are susceptible and so the sum gives no contribution and can be *neglected*, as:

$$R_* = (\langle k \rangle - 1)(1 - P_{ext}) \tag{6.34}$$

Moreover, if R_0 is really low[16], the probability of having an outbreak is:

$$1 - P_{ext} = 1 - \left(\frac{1}{R_0}\right)^{\lambda_{ij}} \simeq \lambda_{ij}(R_0 - 1) = \frac{\alpha w_0}{\mu}(R_0 - 1) \tag{6.35}$$

which actually simplifies our expression for R_*. Indeed, recalling that according to our assumption every patch has same number of connections and travellers, we obtain:

$$R_* = (\langle k \rangle - 1)\frac{\alpha w_0}{\mu}(R_0 - 1) \tag{6.36}$$

We want now to study the **dependencies** of the **invasion potential**. It is indeed a *growing* function of: R_0, both mobility related quantities *overall traffic rescaling* w_0 and *average number of connections* $\langle k \rangle$. In addition, it is *inversely proportional* wrt recovery rate μ or, in other words, it is a *growing* function of **infectious duration**. The more we stay infected the larger becomes R_* and if the latter is larger than $R_* > 1$ the epidemic is able to spread to other patches, which actually makes sense.

[16] Actually this last approximation does not hold for $COVID19$, since it was able to spread given that its R_0 was high.

6.4.2 Heterogeneous networks

Let us introduce a more realistic model and consider that networks are actually different from the homogeneous we have seen up to now. Indeed, **real systems** are **highly heterogeneous**. As we have already pointed out (see Sec. 6.1.1 and Fig. 6.1 and 6.2): the number of connections and travellers along the connections is heterogeneous[17]. This implies that average quantities, such as $\langle k \rangle$, are not representative of the properties of the patches: **homogeneous approximation is bad**. However, we were able to find some **scaling relations**: these are some approximate laws that make computations feasible and are related to the **degree-block** description (coarse-graining).

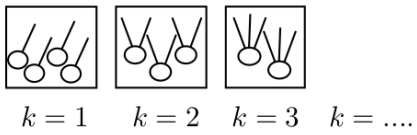


Figure 6.16: Patches are divided into different degree blocks, according to the number of connections k they have.

With **degree-block** description we mean that we group patches according to their degree, and consider patches within the same degree-classes homogeneous. In other words, we make the approximation, similar to the one we have made with networks, that patches with **same number of connections behave in the same way**, like when we used the *Mean Field Degree Approximation* with networks.

We can rewrite some quantities using some empirical laws found when analyzing air traveling data: catchment populations, number of connections for every airport etc. The number of individuals resident in the patch k is:

$$N_k = N_0 k^{\phi}$$

Whereas the flux between two nodes k and k':

$$w_{kk'} = w_0 (kk')^{\theta}$$

therefore the **probability** of **travelling** from a node of degree k to a node of degree k' is:

$$p_{kk'} = \frac{w_0}{N_0} \frac{(kk')^{\theta}}{k^{\phi}} \tag{6.37}$$

where the ϕ, θ are empirical values. Once we have as **input** the **degree distribution** $P(k)$, we can try to find the number of diseased subpopulations at generation n, with k mobility connections, namely D_k^n.

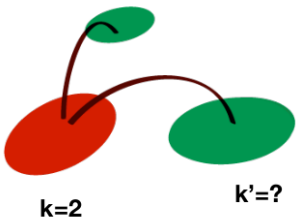


Figure 6.17: A patch that has $k = 2$ connections was infected. We want to compute what is the probability that it is connected to a disease-free patch of degree k'.

[17] Colizza & Vespignani, PRL 2007, JTB 2008

We want now to derive the invasion equation as before, but this time in a heterogeneous mean field approach framework. The question to be replied is therefore the following: *what is the probability that an infected patch with degree k' is connected to a disease-free patch of degree k*. The answer to this question is the following:

$$D_k^n = \sum_{k'} D_{k'}^{n-1} (k'-1)P(k|k') \left(1 - \sum_{m=0}^{n-1} \frac{D_k^m}{V_k}\right) (1 - P_{ext}(\lambda_{k'k})) \tag{6.38}$$

which looks like 6.33, except for the sum over all possible degrees k' that was introduced since we are dividing our patches in same-degree classes. Let us analyze every term: the $k'-1$ is the number of mobility connections through which the seeding might potentially occur. This is multiplied by the probability that contact has degree k, given we are starting from a node with degree k': explicitly *in networks* when we pick a node at random $P(k|k') = k\frac{P(k)}{\langle k \rangle}$. In this step, we have assumed that the network is uncorrelated. This however implies that when we are making connections at random we are more likely to connect to hubs, leading to the so called *friendship paradox*, a.k.a. "rich gets richer", where nodes that have already many connections tend to accumulate more. These last two terms are given by the **topology** of the network. The second last term (red) is equal to the one previously obtained: it is indeed the probability that the contact patch belonging to k class is disease-free. Last term (orange) is, as before, the probability for the epidemic to not get extinct before the global outbreak. One should note that, as previously done, this last term can be approximated to $\lambda_{k'k}(R_0 - 1)$ assuming this is a *branching process*, and moreover for heterogeneous networks:

$$\lambda_{k'k} = \frac{w_0(kk')^\theta}{\cancel{N_0 k^\phi}} \frac{\alpha}{\mu} \cancel{(N_0 k^\phi)} = w_0(kk')^\theta \frac{\alpha}{\mu}$$

Finally, 6.38 can be rewritten as it follows:

$$D_k^n = (R_0 - 1)\frac{\alpha w_0}{\mu} \frac{k^{1+\theta} P(k)}{\langle k \rangle} \sum_{k'} D_{k'}^{n-1}(k'-1)k'^\theta \tag{6.39}$$

If we define the highlighted factor in red as Θ^{n-1} for a more compact writing, we can multiply both rhs and lhs of Eq. 6.39 by $\sum_k (k-1)k^\theta$ thus obtaining:

$$\Theta^n = (R_0 - 1)\frac{\alpha w_0}{\mu} \frac{\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle}{\langle k \rangle} \Theta^{n-1} \tag{6.40}$$

In this way we obtain a function that is monotone as well and has the property that the epidemic spreads if Θ^n is greater than Θ^{n-1}. But, actually, this how R_* works, so we have found an expression for the **invasion potential**:

$$R_* = (R_0 - 1)\frac{\alpha w_0}{\mu} \frac{\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle}{\langle k \rangle} > 1 \tag{6.41}$$

If the last condition holds, then the epidemic will spread. As one can easily note, R_* is a growing function of R_0, the overall traffic rescaling and average number of connections (w_0), epidemic attack rate α, infectious duration ($\tau = \mu^{-1}$) and, finally, of the **moments of the degree distribution and its fluctuations** which are very large for random networks: $\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle \approx 7 \cdot 10^4$ while $\langle k \rangle = 10$. In this way, the network topology is indeed *helping* and *favoring* the spread of the disease (see Fig. 6.18)! One should keep in mind that we are dealing with travelling individuals as integers, and the stochasticity relies in the process of travelling, which might either happen or not (Bernoulli process). However, if they are treated as integers we are not sure that outbreak will happen for sure. Indeed, this would happen when dealing with continuous variables (i.e. fraction of individual) which would be quite unrealistic.

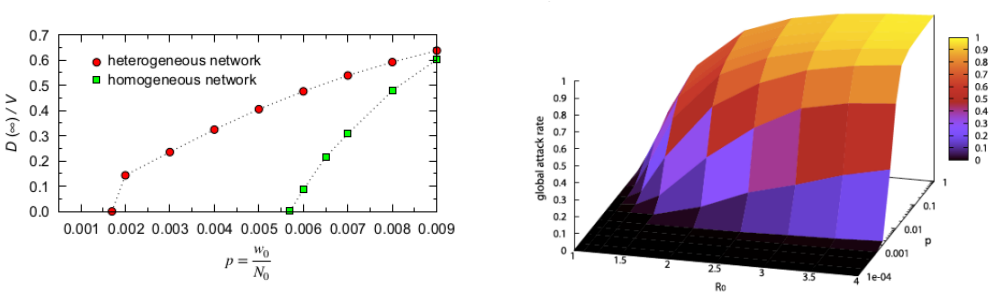


Figure 6.18: **Left:** Network topology favors the spreading of the disease and its global outbreak. **Right:** Attack rate in function of different traveling probabilities and the basic reproductive ratio R_0.

6.5 Spatial spread of competing diseases

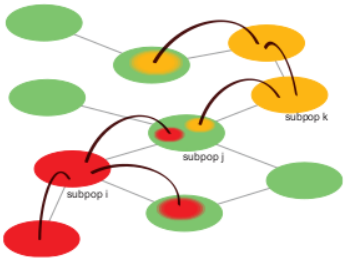


Figure 6.19: Two strains present in the same metapopulation network.

We will present now an application to understand better what we have explained so far: let us assume that there are two **competing pathogens** that are in the same metapopulation system. The **compartmental** model[18] that we obtain is the one in figure 6.20: there are two strains of a certain disease present, and one individual might be infected by either one of them. Once he recovers, he acquires the **full-cross immunity**, that is to say that he cannot be infected by both of them any more. The main difference between the two is that the infectious period of one is actually **faster** than the others ($\tau_{slow} > \tau_{fast}$), whereas they share the **same** R_0.

Since the two strains have actually different infectious periods, they lead to **different infectious dynamics**. Indeed, τ affects both the slope of the *exponential growth*, as well as the *peak* and the presence of the disease at longer times. For instance, keeping $R_0 \propto \beta/\mu$ **fixed**, if τ is small we stay infected for less time but the transmissibility β must be actually higher, hence the spreading explodes. On the other hand, if τ is large, we stay infected more and therefore the disease has more time to spread despite the lower β. At the end of the epidemic, we will have reached the same amount of population but ended up with very different dynamics.

If we let the two different disease spread in a **well-mixed population** with the same assumptions as before: the faster strain will infect faster much more people, being the transmissibility higher. In a certain sense, the "slow" disease will *die out*.

Let us now **introduce**, under the same assumptions of two competing diseases one faster than the others with fixed R_0, the **metapopulation network**. The probability to travel from a patch to another is p, regardless of the patches. The results of the simulation are depicted in figure 6.21.

At the end of simulation there might be a little chance for the two strains to coexist if they reach the same patch, however, more likely the fast strain will make

[18] Poletto, PLOS Comp Biol PRL 2007, JTB 2008.

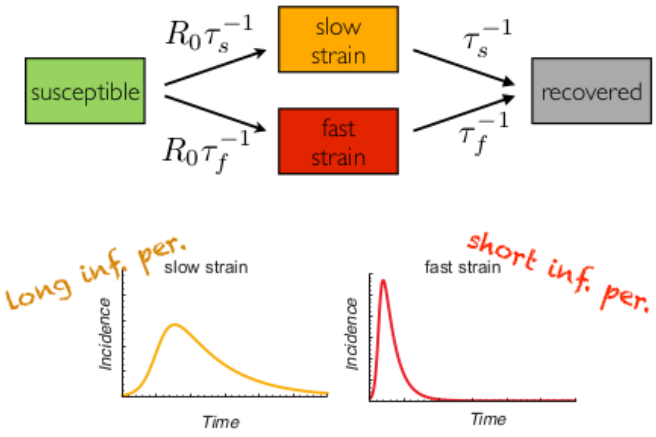


Figure 6.20:
Top. Compartmental model for two competing disease with same R_0 and different infectious time τ. Once one recovers, he acquires the full cross immunity. **Bottom.** Different epidemic timescales for the fast and slow strain of the disease.

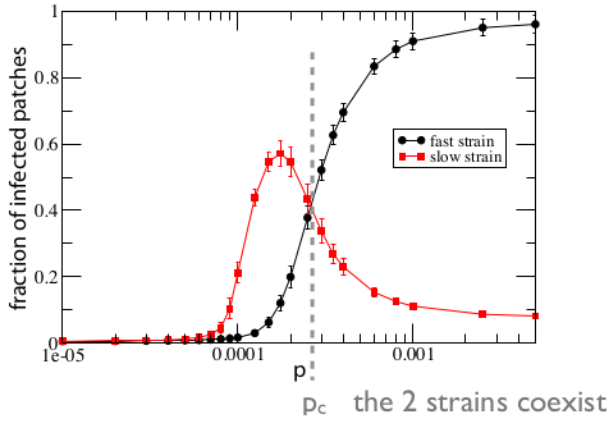


Figure 6.21: Numerical simulation results for two competing strains in metapopulation networks for different value of travelling probability.

the other die out when the *probability of travelling* is *higher*. On the other hand, if the *travelling probability* is *less* they even might not encounter each other, and therefore the slower strain is favored to last longer and still be present at the end of simulation time. We can find a **probability** p_c for which the two diseases show a *crossover*, having the same prevalence. One should note that, for higher traveling probability, we approach the *homogeneous mixing regime*.

Recalling what we have discussed, we can bring what is at the scale of individuals, namely the infection duration μ^{-1}, R_0 and the logistic curve $I(t) \sim e^{\mu(R_0-1)t}$ to a **patches scale**. The timescale will be the *outbreak duration* T and R_* will be:

$$R_* = (\langle k \rangle - 1)\frac{\alpha p N}{\mu}(R_0 - 1) \tag{6.42}$$

while the **number of diseased** patches is $D(t) \sim e^{\frac{1}{T}(R_*-1)t}$. One should note that, since we are assuming that all patches are the same, the outbreak duration T on average is the same for every patch.

According to the fact that R_* is an increasing function of μ^{-1}, we have that the invasion potential $R_*^s > R_*^f$. However, for *large* p it holds that both $R_*^s, R_*^f \gg 1$, but the faster strain is able to reach more rapidly new patches. Instead, for *small* p, we have $R_*^s > R_*^f$ and the slower strain is more able to percolate through the system.

We want now to understand how the probability p_c for the **crossover** changes with respect of R_0. We need to solve the following equation:

$$\frac{D_s(t)}{D_f(t)} \sim e^{\left(\frac{R_*^s-1}{T_s}-\frac{R_*^f-1}{T_f}\right)} = 1 \tag{6.43}$$

The solution is plotted in the left image of Fig. 6.22. It is given by replacing the value for R_*, keeping also in mind that in homogeneous mixing $T_s = T_f$. In this way, it becomes reasonable to be solved, otherwise a graphical approach would have been needed.

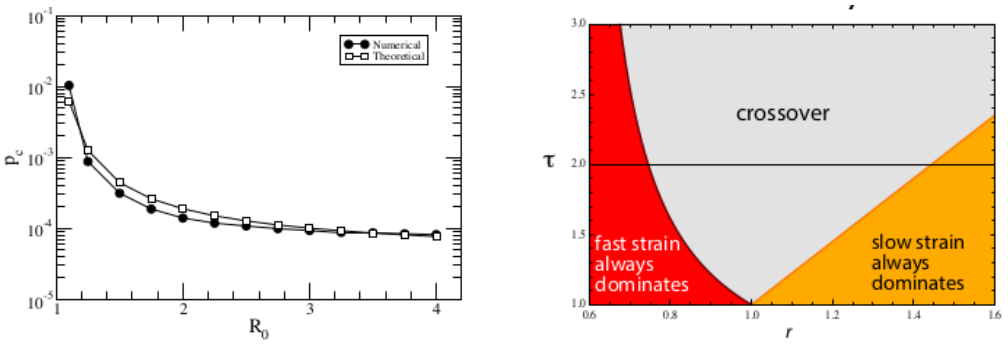


Figure 6.22:
Left: Crossover probability for different Basic reproductive rate R_0.
Right: Two competing strains in homogeneous metapopulations with different τ and different R_0.

Now let us **drop** the assumption that R_0 is **fixed**, and allow for different R_0[19]:

$$R_0^s = rR_0^f \tag{6.44}$$

where r is the ratio between the two R_0 and τ (see Fig. 6.22) is the ratio between the two infectious periods. Now, things start to change, indeed there are some regions where **regardless** of the level of **mobility** either one of the two strain dominates: this happens quite always for limiting cases wrt r. For *small* r fast strain *always* dominates regardless the time of infection, while for *large* r slow strain always dominates when τ is not so large. There is in addition a **crossover** region where the two may coexist and **mobility does** matter to determine which one spreads more than the other.

Analytically, if we introduce the exponential growth in homogeneous mixing as $G = \mu(R_0 - 1)$, we see that there might be two different cases:

$$G^s > G^f \implies R_*^s > R_*^f \quad or \quad R_*^f > R_*^s \implies G^f > G^s \tag{6.45}$$

that define our colored areas where either one strain dominates (the one that comes with larger R_*).

However it is present a third case: if $G^f > G^s$ *and* $R_*^s > R_*^f$ we may observe **crossover**. More particularly:

$$\begin{aligned} r > 1 \rightarrow \quad & G^f > G^s \quad \Rightarrow \quad rR_0^f - 1 < \tau\left(R_0^f - 1\right) \\ r < 1 \rightarrow \quad & R_*^s > R_*^f \quad \Rightarrow \quad \tau\alpha_s \log(R_0^s) > \alpha_f \log\left(R_0^f\right) \end{aligned}$$

Summarizing: depending on the relation between the pathogens' traits, mobility can play a determinant role or be non-influential for the outcome of the competition.

[19]Poletto et al. Sci Rep 2015.

In the space of epidemiological parameters, there exists a region for which lowering the traveling probability induces a cross over from the fast strain dominance to the slow strain one. This behaviour is determined by the trade-off between epidemic growth and potential for spread at the spatial level, the former being an advantage in a well mixed population while the second being relevant in a sparse environment with intermediate or low mobility coupling.

7

Temporal Networks

Lecture 17. *Thursday 26th November, 2020. Compiled: Saturday 14th August, 2021.*

Let us discuss about networks that change structure in time (see Fig. 7.1). This kind of networks are called **temporal networks**. There are actually many of them: for instance let us consider the *face-to-face interactions* network. It is obviously a temporal network: nodes are indeed individuals and an edge is present depending on whether we are talking to each other at the considered time instant. Since one usually does not have conversations that last so much time, edges vary in time. Data for this network can be collected through *RFID*: they are radio frequency detector devices than can be tuned to 1-2 meters distance. If two people come across into each other, then they send the data to a third antenna which keeps track of such contacts. Another example of these network is the *bovine displacement* among farms (see Fig. 7.2) in Italy. Nodes represent farms, and every node is connected to others if cattle were sold and moved from a farm to another.

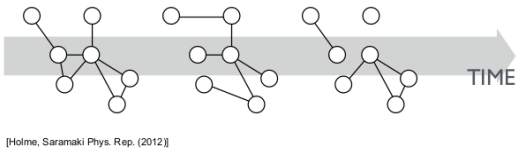


Figure 7.1: A temporal network is a network whose edges vary as time passes.

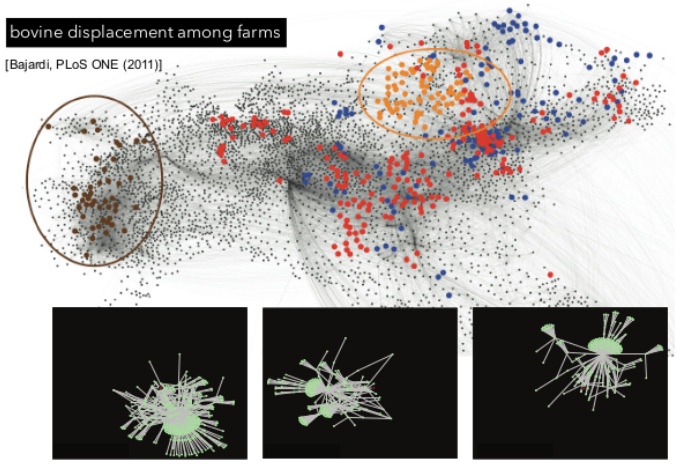


Figure 7.2: Bovine displacements among farms network.

Despite the unquestionable usefulness of such networks, there are some **issues** that arise.

The first problem regards **data visualization**. We are dealing with networks as the one in figure 7.3: we treat *time* as a *continuous variable* and we draw *edges* and

see how they change. However, the **sequence of links** is not a good representation to spot communities/topology: indeed we see what happens and at what moment, but in this way it is difficult to have a general overview. It allows us to preserve the whole information but, actually, it cannot be easily analyzed. As one can imagine, we must make a choice and understand what is really important to our problem.

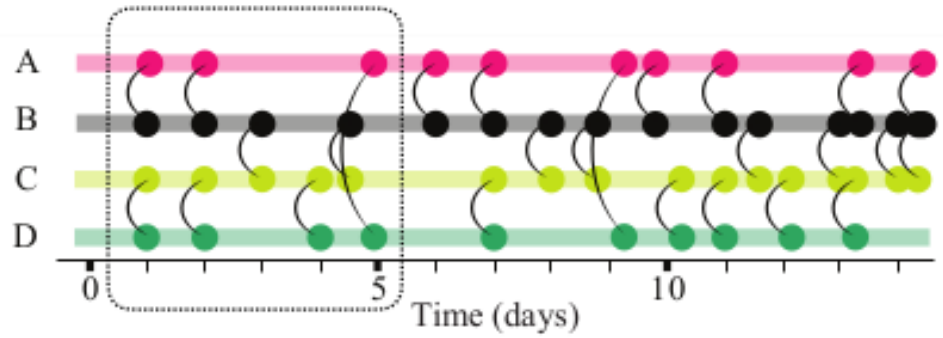
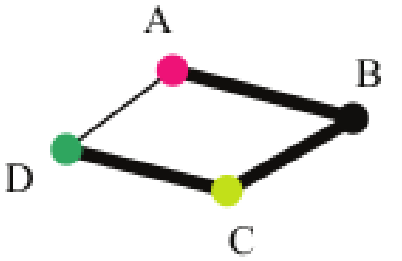


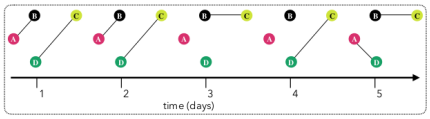
Figure 7.3: Example of temporal network using sequence of links in time representation.

Some possible solutions are the following ones and are depicted in Fig. 7.4:

- **aggregate** the **weights** and lose time dimensionality: the longer the time nodes are connected, the larger the weight. Building a static network makes us lose much information, however these are the simplest networks we may work with and shows clearly topology and communities;
- **discretize the time** that is to say **sample** the network according to rules (**daily snapshot**). This helps us thinning our network, but at the end we have still to deal with a temporal network (but discretized, so it can be seen as a set of static networks) even though a less complex one, and might lead to some misunderstanding.



(a) Solution 1: Aggregate the network as a static weighted one. Only first 5 days are considered as a time window.



(b) Solution 2: Discretize the time and sample according to some rule.

Figure 7.4: Possible solutions for visualization issue with temporal networks.

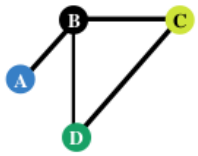


Figure 7.5: Reachability issue is not important in static networks.

The second **issue** one may face when dealing with temporal network is **reachability**. We say that i is **reachable** from j if it exists a path $i \to j$. This is indeed a very easy problem to tackle when dealing with *undirected static networks*: every node belonging to the same **connected component** is reachable from the other members. On the other hand, if the network where we start from is a temporal one, a static network might be the output we obtain after aggregating it. However in this we are losing some important information: if, for instance, two nodes share a link only at the very beginning, we see them as connected even though they are not present any more after a certain time instant. It actually shows lots of path that might be **not**

available any more. Hence, the **existence** of a *time respecting path* does depend on the window $[t, T]$ of observation (see fig. 7.7a)!

As said, in an **undirected temporal network**, j is reachable from i only if there exist a **time respecting patch** $i \to j$. That is to say that there is a **sequence of contacts** that connects $i \to j$ with each contact in the path coming sequentially one after the other from node $i \to j$ (see Fig. 7.6).

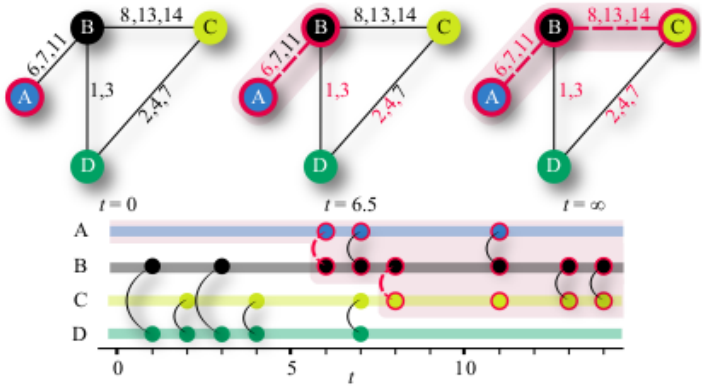
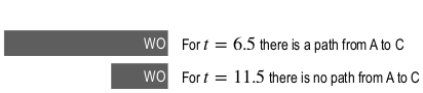
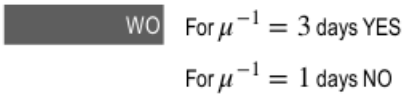


Figure 7.6: The existence of temporal paths between two nodes depends on time.



(a) Reachability depends on the choice of time window.



(b) Whether infection spreads depends both on the disease, in particular μ, and on the time window.

Figure 7.7

One should recall that we are dealing with **epidemic spreading**: if we aggregate network into a static one, possibilities to infect other nodes are much larger rather then what happens for temporal network. Moreover, infection has a certain duration that is given by the **recovery rate** μ: the longer it is, the more likely we stay infected and are able to infect other nodes spreading the disease. On the other hand, when infection duration is small we may recover without having the time to spread it to other nodes that at that moment are not connected to us (see fig. **??**).

The third **issue** that arises is the *activation frequency* or, in other words, the **contact heterogeneities**. In static network, we know that there are actually some nodes that have more contacts whereas some have few. The analogous for temporal networks is that, besides the *number of contacts* a node may have, we must consider how **often activation occurs**. That is to say: if a node activates often and has many contacts, for sure it will lead to a different spreading rather than a node that activates less often and has few contacts. It is also different in terms of epidemic spreading if, once fixed the number of contacts, we have them as "one-shot" (we go to a party and meet lot of people) or diluted wrt time (daily we meet few people): at the end the number of total contacts will be the same. Indeed, the **cumulative number of contacts** results from *activation frequency* and *number of contacts per activation*.

Another **issue** we have to tackle is the **non homogeneous activation**, that is to say that **inter-contact** times (i.e. *inter-occurence* time) between activations is not exponential and the number of contacts cannot be modeled as Poisson. Therefore, individuals do not have some sort of general *activation rate*: it is a more complex

process, since according to an exponential distribution longer periods are not allowed (see fig.7.8). Indeed, empirical data suggests us that human behavior is **bursty**, and this *burstiness* is reflected in broader-than-expected distribution of inter-contact times. Some datasets, for example face-to-face interactions, emails, or phone calls suggested that the best model that reflects human social activity is a power law with a cutoff:

$$P_E(\tau) = A\tau^{-\alpha}e^{-\tau/\tau_E}$$

Indeed, we usually stay inactive for a while, and then have a burst: we start reply to just received messages and eventually start a conversation (*causality effect*).

Figure 7.8: Inter-contact times can be modeled according different distributions, which lead to very different behaviors.

The last **issue** is the most complex one to be captured by models and involves **temporal correlations**. Every time a node activates, there is the chance that the **contacts** it makes are **correlated** to the contacts it had in the **past**. A possible workaround is to introduce the so called **social strategy**[1]:

$$\gamma_{i,t} = \frac{k_{i,t}}{s_{i,t}} \tag{7.1}$$

where $k_{i,t}$ and $s_{i,t}$ are respectively the *degree* and the *weighted degree* of i in the network aggregated over the interval $[t-\delta, t]$. For $\gamma \to 0$ we have a *memory-driven behavior*, so a node tends to make contacts always with the same nodes (social keeper), while for $\gamma \to 1$ the behavior is memoryless, therefore a node shows a more socially exploratory behavior.

7.1 Temporal networks dynamics

Let us consider now, once we have a dataset with temporal description of the network, how its main **properties affect** the **dynamics**. Up to some years ago, temporal networks were not considered of much interest: for instance indeed under some assumption, namely the **time scales separation**, the time dimension seemed not to be relevant in the spreading process and therefore was dropped.

Let us refer to the **average infectious duration** with μ^{-1}, while let τ be the **average contact time**. For $\tau \ll \mu^{-1}$ we do not need to take into account contacts with a high resolution in time, being the network faster than the disease: considering the average network would be actually sufficient. Conversely, if the network is slower than the disease (e.g. migration network), namely $\tau \gg \mu^{-1}$, one may want to exploit the static network and drop the time information. We want to understand when $\tau \sim \mu^{-1}$, so when **timescales** are **comparable**. This can be done obviously if and only if **timescales are definite**: there might be some distributions whose mean is

[1] Miritello, et al, Sci Rep 2013.

not informative, since its variance is high and a timescale cannot be defined out of them.

We can follow two main approaches when dealing with temporal networks in epidemiology: **bottom-up** perspective and **top-down** approaches. In the first one, we start *building mathematical* models for human interactions and the spreading dynamics. However, this is not at all a simple approach and requires too much work. The second approach is to take some *empirical networks* and *run numerical simulations* over them, by randomising (i.e. neglecting/throwing away) some properties whose we want to study the effects. In this way, we are able to understand the *impact of a specific property* in the spreading process.

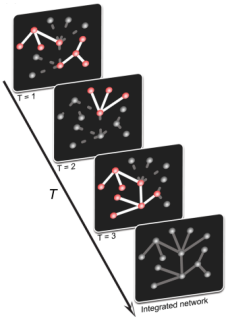


Figure 7.9: Activity driven model.

7.1.1 Activity driven model

It is the first and the simplest model[2] that was introduced to study human interactions, where nodes activate according to a certain rate. It follows the so-called **bottom-up** approach.

The main **ingredients** for this model are:

- Δt timestep, indeed time is discretized;
- N: number of nodes;
- x_i: activity potential, that is the number of activations of i during Δt and is normalized over the total number of activations ($\varepsilon \leqslant x_i \leqslant 1$);
- $F(x)$: distribution of the activity potential;
- $a_i = \eta x_i$ activation rate, where η is a rescaling factor chosen to tune the average number of active nodes per unit time in the system $\bar{N} = \eta \langle x \rangle N$;
- m: number of connections made at each activation, it is the same for all the nodes.

The **algorithm** for the model is indeed the following:

- at each time step t the network G_t starts with N nodes that are *all* disconnected;
- node i activates with probability $a_i \Delta t$ and makes m links with other randomly selected nodes. Non-active nodes are still able to receive connections from active nodes;
- at the next time step $t + \Delta t$ all the edges are deleted, so all links last $\tau_i = \Delta t$.

This model actually captures the issue introduced when talking about **heterogeinity in activations**: now we are indeed modelling Poisson activations, despite the rate of activation is heterogeneous. In addition, it does not take into account possible correlations between activations. Moreover, *network at each timestep* has *on average* $E_t = m\eta \langle x \rangle N$ **edges** and $\langle k \rangle_t = \frac{2E_t}{N} = 2m\eta \langle x \rangle$ **average degree**. At the end of the day the **topology** of the network is *homogeneous*[3], even though **activation rate** is *heterogeneous*!

Now we want to understand how properties change when we **integrate the network over a time window** T. The **degree** of a node i, in the aggregated network is:

$$k_T(i) = k_T^{OUT}(i) + k_T^{IN}(i) \tag{7.2}$$

[2] Perra et al, Sci Rep 2012.

[3] $P(k)$ is Poisson

Let us focus on the first term $k_T^{OUT}(i)$: i makes on average $Ta_i m$ links. We want to understand how many different nodes i connects to, without counting twice repeated links with the same node. This should resemble the *Urn problem*, where we want to count the number of *different* balls extracted from an urn with N balls after T extractions. We know that the probability for each ball (node) to be extracted at least once is $p = 1 - [1 - 1/N]^{Ta_i m}$ and the probability of extracting d balls is binomial with parameters $\text{Bin}(p, N, d)$. Hence, the average number of balls k_T^{OUT} is:

$$\left\langle k_T^{OUT} \right\rangle = Np = N[1 - e^{-\frac{Ta_i m}{N}}] \tag{7.3}$$

as $N \to \infty$ and the time window is such that $T/N \to 0$.

Let us focus on the other term $k_T^{IN}(i)$, that is the number of nodes that make connections with i among those that were not targeted by i. In this way we avoid double counting. The probability that a node was not a target of i is the complementary as before, that is: $[1 - 1/N]^{Ta_i m} \sim e^{-\frac{Ta_i m}{N}}$. Given that the average number of edges formed at each timestep is $mN \langle a \rangle$ and they can connect to node i with probability $1/N$, we have that:

$$\left\langle k_T^{IN} \right\rangle = \frac{m\cancel{N} \langle a \rangle}{\cancel{N}} e^{-\frac{Ta_i m}{N}} = m \langle a \rangle e^{-\frac{Ta_i m}{N}} \tag{7.4}$$

Therefore, the degree formula can be written as function of each activity potential a_i:

$$k_T(i) = N[1 - e^{-\frac{Ta_i m}{N}}] + m \langle a \rangle e^{-\frac{Ta_i m}{N}} \simeq N[1 - e^{-\frac{Ta_i m}{N}}] = N[1 - e^{-\frac{T\eta x_i m}{N}}] \tag{7.5}$$

where we assumed that $N \to \infty$ and $T/N \to 0$.

Rewriting now the **activity potential** as a function of the *degree* $x(k)$:

$$x(k) = -\frac{N}{\eta m T} ln\left(1 - \frac{k}{N}\right)$$

We can revert both the latter formula and $P_T(k)\mathrm{d}k \sim F(x)\mathrm{d}x$, thus obtaining **degree distribution** of the aggregated network for observation in window of length T:

$$P_T(k) \sim F[x(k)] \frac{\mathrm{d}x(k)}{\mathrm{d}k} = \frac{1}{Tm\eta} \frac{1}{1 - k/N} F\left[-\frac{N}{\eta m T} ln\left(1 - \frac{k}{N}\right)\right]$$

If the time window dimension is small, namely $T \to 0$, then also $k/N \to 0$, and the latter expression can be further approximated as:

$$P_T(k) \sim \frac{1}{Tm\eta} F\left[\frac{k}{Tm\eta}\right] \tag{7.6}$$

This implies that **nodes activate and form a heterogeneous network** despite they activate heterogeneously. In other words, heterogenous topology in the **aggregated** network (i.e. there are hubs), over a window T, results from a heterogeneous activity potential (i.e. they activate more often). To avoid any further confusion: at each *timestep* the static network is homogeneous, regardless the the activation rate which is heterogeneous. On the other hand the *aggregated* network results being heterogeneous, thanks to the heterogeneous firing rate distribution.

Let us try to understand what are the **effects** on the network dynamics on epidemic spreading. For instance, let us consider how the **epidemic threshold** changes according to the activation dynamics. In order to pursue our goal, let us use the **activity block approximation** that works as the same way as the *degree block approximation*, and consider a SIR model. Let us define the *probability of transmission*

per contact as β, moreover let us assume $m = 1$: every time a node activates, it makes a single connection. The SIR equation, classifying nodes according their activity, for infected nodes at time $t + \Delta t$ within class a is:

$$I_a^{t+\Delta t} = -\mu \Delta t I_a^t + I_a^t + \beta(N_a^t - I_a^t) a \Delta t \int da' \frac{I_{a'}^t}{N} + \beta(N_a^t - I_a^t) \int da' \frac{I_{a'}^t a' \Delta t}{N} \tag{7.7}$$

The *green* term describes the probability for a node in class a to activate and get in contact with infected nodes of any other classes a', from which it contracts the disease. The blue term, on the other hand, returns the probability we have to be infected by other infectious nodes that instead activate while we do not. Defining the last integral $\theta^t = \int da' \frac{I_a^t a' \Delta t}{N}$, we are able to write the **total number of infectious** as:

$$\int da I_a^{t+\Delta t} = I^{t+\Delta t} = I^t - \mu \Delta t I^t + \beta \langle a \rangle I^t \Delta t + \beta \theta^t \Delta t \tag{7.8}$$

Multiplying both rhs and lhs of the last equation by a and integrating over the latter, we obtain:

$$\theta^{t+\Delta t} = \theta^t - \mu \theta^t \Delta t + \beta \left\langle a^2 \right\rangle I^t \Delta t + \beta \langle a \rangle \theta^t \Delta t \tag{7.9}$$

Up so far, we obtained two equations in two variables I and θ. Rewriting them as differential equations:

$$\partial_t I = -\mu I + \beta \langle a \rangle I + \beta \theta \tag{7.10a}$$

$$\partial_t \theta = -\mu \theta + \beta \left\langle a^2 \right\rangle I + \beta \langle a \rangle \theta \tag{7.10b}$$

Let us understand now when these expressions return as a growth in the number of infectious. The tool we will use is **linear stability analysis**. The Jacobian is:

$$J = \begin{vmatrix} -\mu + \beta \langle a \rangle & \beta \\ \beta \left\langle a^2 \right\rangle & -\mu + \beta \langle a \rangle \end{vmatrix} \tag{7.11}$$

whose set of eigenvalues is $\Lambda_{1,2} = \beta \langle a \rangle - \mu \pm \beta \sqrt{\langle a^2 \rangle}$. We want the *largest eigenvalue* to be **positive**, in order to have the number of infectious growing.

The condition for the *largest eigenvalue* to be *positive* is thus:

$$\frac{\beta}{\mu} > \frac{1}{\langle a \rangle + \sqrt{\langle a^2 \rangle}} + \mathcal{O}\left(\frac{1}{N}\right) \tag{7.12}$$

and it sets a threshold for β that is function of the moments of the activity distribution. Note as if activity distribution is heterogeneous, then its variance becomes large and the threshold decreases, hence favoring the spread.

Summarizing, the **activity driven model** captures the realistic property of human behavior (face-to-face, sexual contacts, phone call, email, tweets) and takes into account heterogeneous activity rate. Moreover, the contact network at a **given instant** is *sparse* and has homogeneous degree. Once again, we want to stress that the **aggregated network** over a certain time window is really resembling a *heterogeneous degree* network!

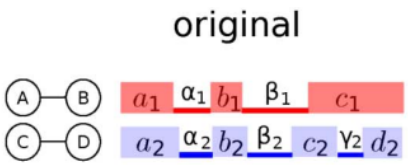
So if we do not make any assumption on the pattern of activation that may unfold at the same time scale of the spreading process, computations are made possible within the *activity-block* approximation, which follows the same scheme as the degree-block approximation. In conclusion, **contact heterogeneity lowers the epidemic threshold**.

7.1.2 Randomised Reference Models

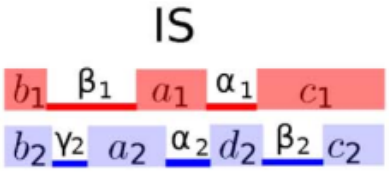
Let us now discuss an example of the other type of approaches one may want to use when dealing with *temporal networks*: **top-down** approaches[4]. Let us recall we want to understand how the **temporal structure** of the network **impacts** the **spreading**. We therefore compare the epidemics on real data with the outcome in suitable null models that randomize (i.e. destroy) some properties over some others. However, we keep the topology of the aggregated network fixed.

We are going to study 3 different types of randomizations. Let us define the following quantities:

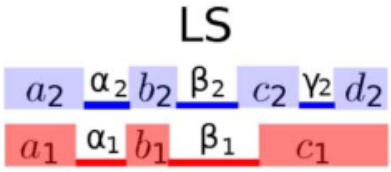
- $P(\tau)$: inter-contact time distribution;
- ω_{AB}: cumulated contact durations of an arbitrary link AB;
- $P(\omega)$: distribution of the cumulated contacts duration;
- n_{AB}: number of contacts per link of an arbitrary link. These are the total times A, B are in contact, regardless of how much;
- $P(n)$ distribution of the number of contacts per link.



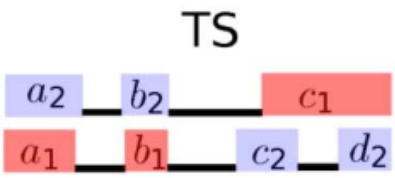
(a) Initial condition.



(b) IS: Interval shuffling applied.



(c) LS: Link Shuffling applied.



(d) GTS: Global Time Shuffling applied.

Figure 7.10: Red (light) segments indicate $A \breve{} B$ contacts, while blue (dark) segments indicate $C \breve{} D$ contacts. For each link, individual contact intervals are marked with latin letters and inter-contact intervals with greek letters.

Let us discuss the three different of shuffling (see Fig. 7.10) we may have:

- **Interval Shuffling**: the sequences of contact and inter-contact durations are reshuffled for each link separately. The only property we destroy is the *causality*: correlations between historical information of link are destroyed, since their chronological order (sequence) is randomized.
- **Link Shuffling**: the unaltered sequence of events (i.e. contacts) are swapped between link pairs. In this case we are destroying *causality*: by reshuffling links we destroy causal correlations between pair nodes. Moreover, we are also dropping ω_{AB} and n_{AB}, since we are assigning to each link a history that has been taken randomly from other nodes. We are destroying the exact structure of the networks, i.e. swapping labels.

[4] Gauvin et al Sci Rep 2013.

- **Global Time Shuffling**: we build a global list of the empirical contact durations and, for each link, we generate a synthetic activity timeline by sampling with replacement the global list of contact durations according to the original number of contacts for that link. We are preserving by construction the topology, the number of connections between each pair and their distribution. On the other hand, we are destroying the remaining ones: namely *causality*, $P(\omega)$, ω_{AB}, $P(\tau)$ since now we are drawing randomly activation times.

See Fig. 7.11 in which the main shuffling characteristics are summarized.

RRM	Topology	Causality	$P(\tau)$	ω_{AB}	$P(\omega)$	n_{AB}	$P(n)$
IS	V	X	V	V	V	V	V
LS	V	X	V	X	V	X	V
TS	V	X	X	X	X	V	V

Figure 7.11: Table depicting all different types of randomizations and what property we drop.

Comparing the empirical data and the other to the ones we obtain by using the aforementioned shuffles, we are able to understand what are the **main features** of our network. In addition, we can see how these either enhance or contrast the spreading of a disease. For instance, if we destroy a property and we note that spread does not strongly change, we are even allowed to not collect data related to that property any more, thus saving time, memory and resources. On the other hand, if after shuffling it strongly changes, then we can conclude that that property plays an important role in the spread process.

8

Model fitting

Lecture 19. *Thursday 3rd December, 2020. Compiled: Saturday 14th August, 2021.*

The next step one must make is now **make sense** of data. Therefore, we want to **extract meaningful information** out of it by the mean of a model. However, there are also many **challenges** when collecting data: one should think well about what kind of data is needed, as well as the measurement process and its interpretation. We are going to analyze these arguments using as paradigmatic example COVID-19, but our considerations can actually be applied to *any* spread process.

8.1 Data Collection

When dealing with epidemic processes one usually speaks about **incidence data**, like the one depicted in figure 8.1.

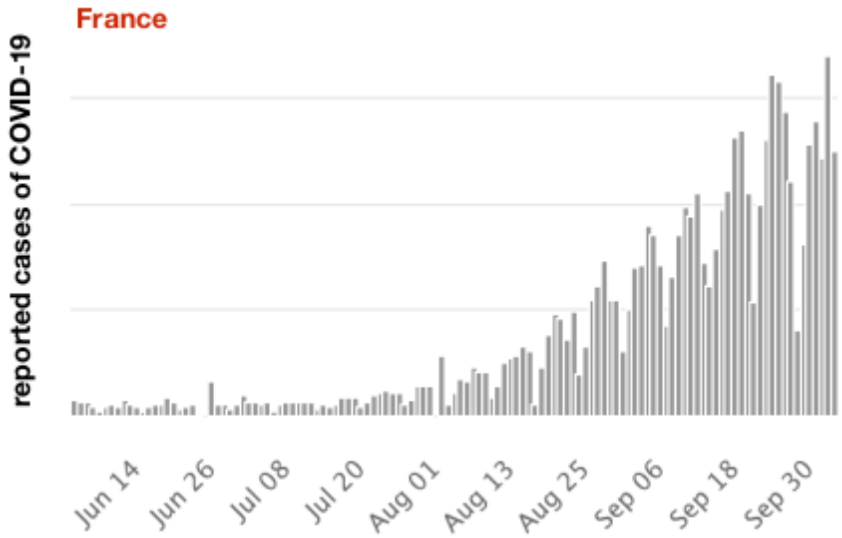


Figure 8.1: COVID-19 curve incidence data in France. For each day, number of reported new cases are shown.

We want to **model** these data, so in other words we want to make **make sense** out of it by interpreting raw numbers. In order to do it, we need first to understand how it was collected and, recalling that data actually gives **partial information**, this must be *completed* by the means of a **model**.

The main **goals** one wants to pursue by using models are:

- **nowcasting**: provide assessment regarding the *present* and *close future* of the epidemics. Therefore, one may need to understand what is and what will be the extension/distribution by groups/regions of the epidemic that is present at that moment, given the partial information returned by data.

- **forecasting**: prediction in *longer term*. For instance, one would like to predict hospitals occupancy, when and how high will be the epidemic peak, how many people will be infected over the next weeks or months, when epidemic will end and what will be its final size.

- **medical and biological understanding**: at the very beginning we have no medical/biological knowledge about the epidemic. Hence, we want to study for instance how it propagates and through what vectors (human-to-human, zoonotic, vector-borne, direct transmission, fomite, aerosol, droplets, etc...), the role of asymptomatic/pre-symptomatic in transmission, susceptibility and rate of symptoms by age group.

- **exploration of counterfactuals and hypothetical scenarios**: we run our model to perform *scenario analysis* and understand what is the best strategy to use in the future. It may be related to vaccination, pharmacological interventions, lockdown, travel restrictions and their impact on the future spreading. Indeed, these are *long-term* projections, despite one may want to explore the case and what would have occurred if a decision had not been made. Hence these arguments are valid also for the past: for instance we want to quantify the impact of lockdown in spreading.

Let us discuss **what** is the *data* we usually work with. When we speak about **incidence in a given area and at time** t, we refer to the **fraction** of population resident in that area that has contracted the disease at time t. Hence, formalizing[1]:

$$\textbf{incidence} = \frac{\text{number of people hit by flu}}{\text{population at risk}} \tag{8.1}$$

Let us take a look closer to the **numerator**. Obviously it is impossible to have a complete information about how many people have the flu at this moment. But, first, one has to face the first problem of **case definition**: set of criteria used in making a decision as to whether an individual has a disease or any other health event of interest. Some possible *criteria* may involve: clinical (e.g. symptoms), laboratory characteristics (e.g. exams, test results). Moreover personal information are taken into account, such as whether this individual travelled to regions at risk/had contact with people at risk can be classified using three levels: confirmed, probable, possible. Cases definition can be either more *sensitive* or *specific*, and it has to be tuned according to the risk assessment. A **sensitive** case definition will detect many cases but may also count as cases individuals who do not have the disease (*possible overestimation*). On the other hand, **specific** case definition is more likely to include only persons who truly have the disease under investigation but also more likely to miss some cases (*possible underestimation*). This is summarized in Fig. 8.2.

	Disease is truly present	Disease is truly absent	Total
complies to case definition	a	b	all cases
does not comply to case definition	c	d	all non-cases
	all 'diseased'	all 'non-diseased'	all people in the study sample

Sensitivity = [a / (a+c)]

Specificity = [d / (b+d)]

Figure 8.2: Case definition can be either more sensitive or specific. In the first case we may end up overestimating the number of cases, whereas when a test is more specific this might lead to underestimation.

[1] We will use *flu* as example, since we have much data of it

Let us see **how** data is **collected**: in France there is a network of *General Practitioners* (see Fig. 8.3) that are volunteers and daily send a report to Health Minister concerning all the cases they have visited during a workday. However, for some diseases, for instance measles, every family doctor is obliged to report the case. Let us continue analyzing the *flu* case. The number of cases reported is indeed the number of cases seen by General Practitioners defined under the basis of some clinical criteria. These are *possible cases*: the guarantee can be returned only after a laboratory confirmation which is available only for a small proportion of cases.

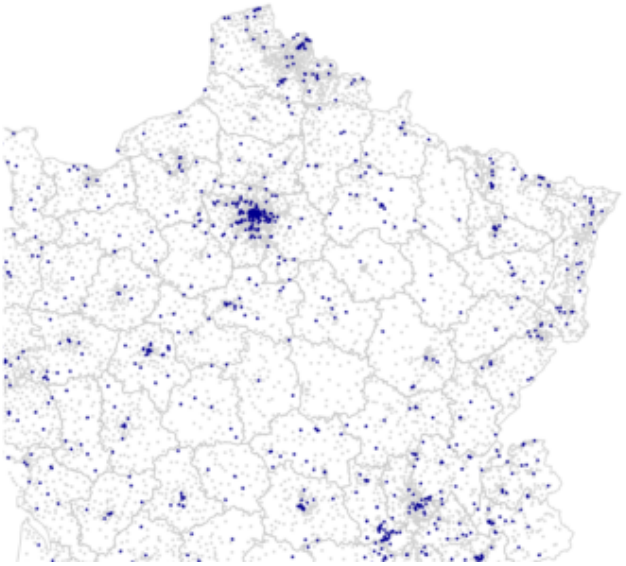


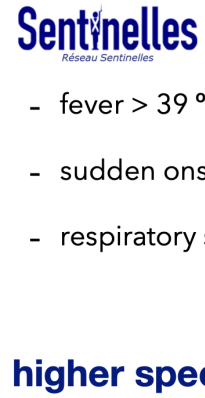
Figure 8.3: The Surveillance Network in france (SN) is based on a fraction of General Practitioner ($\sim 1\%$), who are volunteers.

Given the symptoms of flu:

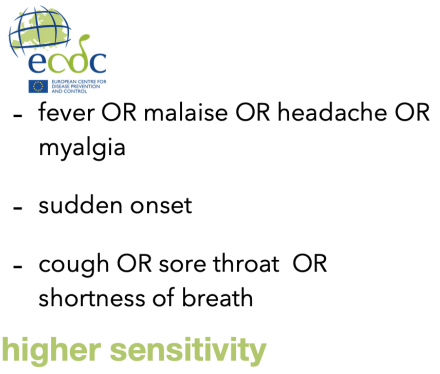
- no symptoms ($\sim 30\%$);
- upper respiratory symptoms, e.g. nasal stuffiness, runny nose, sore throat, sneezing, hoarseness, ear pressure, or earache ($\sim 60\%$);
- lower respiratory symptoms, e.g. cough, breathing difficulty, and chest discomfort ($\sim 2\%$);
- fever ($\sim 35\%$);

The main concept is that it is important to understand according to what criteria data is collected in order to deal better with observables, indeed observables are a *proxy* for the real data. One should note that they may be different even if they regard the same quantity (look Fig. 8.4): for $ECDC$ incidence data we might observe a peak in autumn because of respiratory infections, whereas this might not be present in *Sentinelles* reports, given they do not classify it as case of flu. The $ECDC$ case definition, as one can imagine, has higher sensitivity and this can return overestimation of the number of cases. Conversely, according to *Sentinelles* cases definition we might underestimate their number.

Let us take a look closer to the **denominator**. It refers to the *catchment population*, i.e. all the people living in the catchment area of the General Practitioner reporting the cases, who would usually seek healthcare at the site when they get sick. Therefore, the denominator for the area a that can be computed at a first



(a) "Sentinelles" clinical case definition for flu.



(b) "ECDC" clinical case definition for flu.

Figure 8.4

approximation is[2]:

$$\text{denominator}_a = \text{Population}_a \frac{GP_{SN,a}}{GP_a}$$

where the ratio $\frac{GP_{SN,a}}{GP_a}$ is the proportion of General Practitioners that contribute to the *Surveillance Network* ($\sim 1\%$ according to Fig. 8.3). Moreover, another **problem** that biases our observable is given by the **consultancy rate**. Since many people are asymptomatic or paucisymptomatic, the rate of people going to be examinated by family doctor is highly variable by age: young people, except very little children, are more likely to not go, whereas adults need to go in order to have permission to stay home from work. Raw numbers depend also on family doctors density, on the health-care system (how expensive is going to the GP), and on the period of the year that brings specific diseases. In conclusion, even though data might look simple at a first glance, dealing with it needs to take into account many variables all together and some assumptions are more likely to be made if data is not available. Another important point is that the **confirmed flu cases** are a very small subset among *I*nfluenza *L*ike *I*lness (ILI) people (symptomatic), people that go to General Practitioner, Infected people which either can be detectable or not and that, obviously, need to be recorded as infected.

Another *characteristic* of the **case definition** is that it might be **variable in time**, specially when the range of symptoms is still unknown. It was the case of **COVID-19** at the very beginning. In addition, case definition is **matter of authority**: once cases are reported in hospitals, some papers concerning viral loads, symptoms, evidences are published. Health authority needs to collect and through them in order to define better the case definition, keeping in mind a sort of trade off: if case definition is *high sensitive*, there might be false positives and also cause panic among people. Conversely, if the definition is *too much specific*, we risk to let infectious people go around and spread the infection. This tuning depends on the goals one may want to pursue.

With regards to **COVID-19**, case definition was therefore varying in time being the range of symptoms unknown. Moreover, more problems arose since the disease at the *very beginning* was not wide spread and it was still unclear the region where it was spreading: the denominator related to the catchment population was kind of difficult to estimate at that time. The *reporting rate* was highly variable in time: at the beginning, the tracing system is able to intercept all the case, but the surveillance

[2]Horvitz DG, Thompson DJ. A JASA. 1952;47:663–85

system might saturate even though the case definition remains unchanged. Moreover, due to the change in time of case definition, number of cases can be always retrospectively corrected: real time analysis numbers are indeed biased and this is why sometimes we observe spikes in incidence curves.

8.2 Epidemic Modeling and Bayesian Inference

Typically, the steps are the following ones:

- **Model design/implementation**: decide the model ingredients that synthesise available medical, biological, information etc. We can consider different models, ingredients that describe our **hypotheses**. These can be for instance helpful to simplify our problem and must be *clearly* stated at the beginning.
- **Model calibration**: estimate model parameters from available data, a.k.a. model *fitting*
- **Model validation**: confirming that model output is sufficiently accurate in reproducing the data. It is done as the result of *model calibration*, when we can tell whether our model well represents data also taking into account secondary aspects.

Let us introduce now some concepts about **bayesian inference** and **Maximum Likelihood** that will help us throughout this process. Once we observe data, one wants to introduce an **Observation model** $\mathcal{O}$ according to which we assume to have acquired our data. In our case, we assume it to be a *Binomial* process with probability p: every case, according to a probability p goes to the doctor. In addition we need to define an **Epidemic Model** (e.g. SIR, or alternatively $SEIR$, $SIRS$...) and a **vector of parameters** $\vec{\theta} = (\beta, \mu, ...)$ where some are to be inferred, while others are assumed to kept fixed. This obviously depends on our data and on our **Initial Conditions** I_0, namely the number of infectious. All these quantities help us to define the curve represented in Fig.8.5.

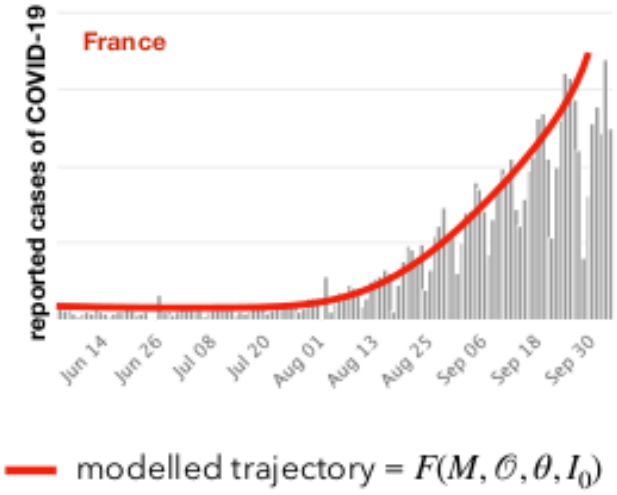


Figure 8.5: Curve fitting the incidence curve presented before (see Fig. 8.1).

We therefore use a **Maximum Likelihood approach** that returns from a probabilistic formulation: indeed the relation between model and data is probabilistic, and our aim is to identify the trajectory and thus the parameters, $\vec{\theta}$, that are **more**

probable given the data we have. One should note that, in the *bayesian framework* probability is used as a **measure of uncertainty**.

When we are dealing with single random variable A we talk about *univariate probability*: the probability that A takes value a is defined as $p(A = a) = p(a)$ and the normalization condition holds $\sum_a p(a) = 1$. For *continuous variables* it can be rewritten as $\int p(a)\,\mathrm{d}a = 1$. When random variables are more than one, e.g. A, B, the joint probability that A takes value a and B takes value b is written as $p(A = a, B = b) = p(a, b)$. The marginal probability $p(a) = \sum_b p(a, b)$ and is the probability that A takes value a regardless b: we indeed summed over all the possible $p(b)$. For *continuous variables* it can be rewritten as $p(a) = \int p(a, b,)db$

Some of the **basic properties** we will deal are the following:

- **Conditional probability** of a from random variable A, given that the outcome of a random variable B was b is $p(A = a|B = b) = p(a|b)$.

- **Bayes Theorem** allows us to rewrite the conditional probability as follows $p(a|b) = \frac{p(a,b)}{p(b)}$.

- **Chain rule**: $p(a, b, c) = p(a|b, c)p(b|c)p(c)$.

We are going now to introduce some variables and a short overview over statistical inference. As said, the latter helps us in drawing conclusions from numerical data about quantities that are not observed: for instance we see that a disease is more frequent in adults and we want to infer its prevalence in children population. Some of these unobserved quantities can be $\tilde{\mathbf{y}}$: potentially observable quantities such as future observations of a process (e.g. predictions), and $\tilde{\theta}$ that are quantities not directly observable such as *parameters* that govern hypothetical process. **Bayesian statistical conclusions** about a parameter θ or unobserved data $\tilde{y}$ are made in terms of *probability statements*. These are expressed as **conditional** probabilities on the observed values of y: $p(\theta|y)$.

In other words, we want to obtain a distribution for θ conditioned to y: $p(\theta|y)$. In order to pursue our goal:

- we need a model (M) that provides us the joint probability distribution of θ and y: $p(\theta, y)$;

- given the model M, thanks to Bayes' Theorem we can write $p(\theta, y) = p(y|\theta)p(\theta)$, with $p(\theta)$ that is the **prior distribution** and $p(y|\theta)$ that is the sampling distribution;

- we use the Bayes rule to *condition* on the known value of the data y, namely:

$$p(\theta, y) = p(\theta|y)p(y) = p(y|\theta)p(\theta)$$

The **unnormalized posterior density**, namely the expression that helps us inferring the parameters we need, is:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{8.2}$$

It is unnormalized since we do not care about the normalization term: it is constant.

One should note that data affects the posterior **only** through $p(y|\theta)$. If we keep fixed y and let θ vary, this is the **Likelihood function** $\mathcal{L}(\theta) = p(y|\theta)$

Example 1: Hemophilia

Let us consider now an **example**. *Hemophilia* is a hereditary disease associated to a gene of the chromosome X. This is recessive inheritance: a man who inherits the gene is affected, a woman who inherits the gene on only one X is not affected. Let us recall, for the sake of completeness, that a man has chromosomes XY, while woman XX.
We want to deal with the following **problem**: given that a woman has an affected brother and a father not affected, she can be a carrier of the gene on either one X. We want to estimate whether she is a carrier, and we define $\theta = 1$ as the situation where she actually is, while $\theta = 0$ describes a situation where she is not. Since we do not have *any* other information, **a priori** one should not introduce any bias, hence $p(\theta = 1) = p(\theta = 0) = 0.5$.
Our empirical **data** consists on the fact that she has got two sons, and neither of the two is affected: $y_1 = 0$ and $y_2 = 0$. The **Likelihood** is therefore:

$$p(y_1 = 0, y_2 = 0|\theta = 1) = 0.5^2$$
$$p(y_1 = 0, y_2 = 0|\theta = 0) = 1$$

Multiplying these terms, we can obtain the **posterior**, namely:

$$p(\theta = 1|y_1, y_2) = \frac{p(y_1, y_2|\theta = 1)p(\theta = 1)}{p(y_1, y_2|\theta = 1)p(\theta = 1) + p(y_1, y_2|\theta = 0)p(\theta = 0)}$$
$$= \frac{0.25 \cdot 0.5}{0.25 \cdot 0.5 + 0.5} = 0.2$$

So the probability that she is a carrier, given our observation is quite low.
However, most of the times, it might happen that **new data is available**: for example the same woman has a third son, which is not affected $y_3 = 0$. Obviously, one does not want to lose all the information obtained so far, hence we **update the prior**. The **prior** becomes:

$$p(\theta = 1) = 0.2, \qquad p(\theta = 0) = 0.8$$

which is the posterior of before. The **Likelihood** follows the same argument as before $P(y_3 = 0|\theta = 1) = 0.5$. Out of these expression we can compute the **posterior**, which is:

$$p(\theta = 1|y_3) = \frac{p(y_3|\theta = 1)p(\theta = 1)}{p(y_3|\theta = 1)p(\theta = 1) + p(y_3|\theta = 0)p(\theta = 0)} = \frac{0.5 \cdot 0.2}{0.5 \cdot 0.2 + 0.8} = 0.111$$

Indeed, the probability for the woman to be a carrier is even lower. We want to stress once again that, updating the posterior, we have not lost any information that was previously obtained.

Example 2: Bernoulli trial

Let us consider **another example**, with *Bernoulli trials*. Recalling that for a **Binomial distribution** we have n independent trials with two possible complementary outcomes (either failure or success) and we observe y successes. The probability of a success is θ, consequently for a failure is $1 - \theta$ and this is the parameter one may want to estimate. For instance, given a number n of observations with y successes, we want to infer whether a coin is fair $\theta = 1/2$.
In the case, we do not have any information so we can use a **uniform prior**

for the paramterθ: $\mathcal{U}[0,1]$. The **Likelihood** is a *Binomial distribution* with parameters:

$$P(y|\theta) = \texttt{Bin}(y|n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

Note as we do not write the dependence on n on the left side because is part of the experimental design and considered fixed. All probabilities will be conditional on n. The **posterior** becomes:

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}$$

that is nothing more that a *Beta* distribution Beta$(y+1, n-y+1)$. Note that $\binom{n}{y}$ does not depend on θ therefore can be disregarded.

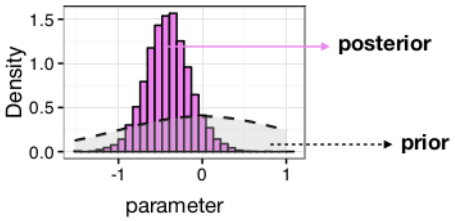


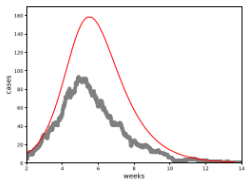
Figure 8.6: When we work with Bayesian inference we usually want to infer the most likely value for a parameter thanks to the *posterior*, given the observation (*data*) and our previous knowledge (*prior*).

Shortly, the **prior distribution** summarises my a priori knowledge about parameters. It might be defined based on the literature, for instance if we are analysing an outbreak of flu and our goal is to estimate R_0, we may want to look at previous R_0 estimates. If we have no prior knowledge on the problem, however, the best idea is to use a vague, or flat, *noninformative* prior. Instead, the **posterior distribution** is a compromise between data and prior information. Such compromise is increasingly controlled by data as the sample size increases. Posterior *variance* on average is smaller than prior variance: if it occurs, then this denotes either a conflict or an inconsistency between sampling model (i.e. data we obtain) and the prior distribution. The **main information** one wants to obtain from the posterior are:

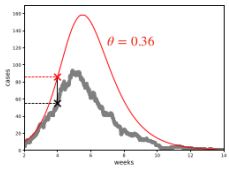
- **Mode** of the posterior, i.e. the most likely parameter given the data.
- **Uncertainty** associated to our estimate, i.e. the $C.I.$ (credibility interval) and usually it is given by the 2.5% and 97.5% quantiles. It is really a relevant quantity: the range according to which the mode spans can lead to really **different** and **opposite outcomes**.

Let us discuss now a more practical problem. We want now to understand how to **fit an incidence curve**, such as the one in Fig. 8.7a where all symptomatic cases that go to the doctor are detected, to estimate R_0. In other words, we have an incidence curve and want to fit a SIR model. The steps to follow are:

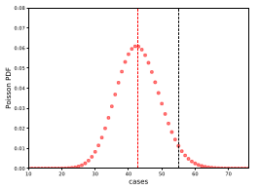
- **Data**. We know that the infection causes *symptoms* for the 50% of cases. We assume that all symptomatic cases go to the doctor and are detected.
- **Observation model**: weekly cases are independently detected with probability $d = 0.5$. Observations y_t are independent. This process is binomial: each case has the 50% to be detected. However, for large numbers, it can be approximated as a Poisson.



(a) Raw data: incidence curve.



(b) Prediction of *total* number of cases for $\theta = 0.36$, i.e. $I(t)$.



(c) Prediction of *detected* number of cases for $\theta = 0.36$ at fourth week, i.e. $y_{t=4}$.

Figure 8.7

- **Model**: SIR, from the literature and peer review journals we know that the average infection duration is $\mu^{-1} = 5.5$ days. We can assume this parameters to be fixed.
- **Parameter** we want to fit is $\theta = \beta$
- **Prior** uniform in $[0, 1]$, since we do not have any information on β.
- **Likelihood** $\mathcal{L}(\theta) = p(y_1, , ...y_t, ..., y_{t_M}|\theta) = \prod_t p(y_t|\theta)$, cases at t-th week are denoted by y_t

Then, for each value of β (our θ to be inferred) we run a simulation of the trajectory of the SIR, fixing μ and I_0 based on available knowledge. Practically, we can approximate observed data to be distributed as a Poisson(λ), where $\lambda = I(t)d$. This is the number of cases we see at t-th week times the detection probability. The total number of cases is distributed as the red line in fig. 8.7b. This returns us the model projection related to the observation.

For instance, at week $t = 4$, we observe $y_{t=4} = 55$ cases. If we assume data is distributed as a Poisson, the sampling distribution is $y_{t=4}|\theta \sim \text{Poisson}(\lambda)$, then $\lambda = I(t)d = 86 \cdot 0.5 = 43$ we expect 43 detected cases. Keeping the parameter θ fixed, we have to compute the likelihood as $\mathcal{L}(\theta = 0.36) = \text{Poisson}(55|\lambda)$. This procedure has to be done for every value of the parameter θ. Moreover, one should take into account that since dealing with products is uncomfortable because of really small numbers and for computational simplicity, we take the logarithm[3] of the likelihood, therefore considering the sum: $\log \mathcal{L}(\theta) = \sum_t \log\ p(y_t|\theta)$.

Let us summarize the **basic idea** behind likelihood computation: we want to evaluate the probability of the data given the model and the parameters. In order to **estimate** θ we keep the model M and x_0 fixed and vary θ to compute the probability $p(y|\theta)$. The Likelihood function is $\mathcal{L}(\theta) = p(y|\theta)$, and generally it can span a wide range of orders of magnitude, which can lead to numerical problems. In practice it is better to work with the log-likelihood: $\log \mathcal{L}(\theta) = \log\ p(y_1, ..., y_n|\theta) = \sum_i \log\ p(y_i|\theta)$. The best estimate for θ is actually the one that maximizes the posterior, i.e. the **mode**.

In general the **posterior distribution** is *difficult* to obtain analytically, therefore **numerical integration** is required.

Lecture 21. *Thursday 10th December, 2020. Compiled: Saturday 14th August, 2021.*

8.3 Monte Carlo approaches

In order to **compute the likelihood** as a starting point, if our knowledge is none, we can use a **grid of points** in order to compute $p(\vec{\theta}|y)$ for $\theta_1, \theta_2, ...\theta_n$ equally

[3]It can be done since $\mathcal{L}(\theta)$ is monotone, without any losing of generality.

spaced. In this way, we *approximate* the continuous density function $p(\vec{\theta}|y)$ with the **discrete density function**:

$$\frac{p(\theta_i|y)}{\sum_i p(\theta_i|y)}$$

However, this approach is to become rapidly *unfeasibile* as the dimensionality of the parameter space becomes larger.

Another approximation one might want to take is the **Trapezoidal approximation**: after computing $p(\theta|y)$ for a discrete set of points $\theta_1, ...\theta_n$ we can approximate $p(\theta|y)$ with a piecewise-linear function, connecting the $p(\theta_i|y)$ points with linear segments.

One may ask now how to **numerically sample** $p(\theta|y)$. In order to do so, we use a *positive function* $g(\theta)$ defined for all θ such that $p(\theta|y) > 0$. Moreover it must hold that:

- we are able to draw samples from $g(\theta)$;
- for some *constant* M the ratio $p(\theta|y)/g(\theta) \leqslant M$ is defined;
- $g(\theta)$ must have a finite integral.

The **rejection sampling** algorithm hence is constituted by these two steps (Fig. 8.8):

- sampling θ at random from the pdf of $g(\theta)$;
- with probability that is given by the ratio $p(\theta|y)/Mg(\theta)$ we accept θ pretending it is drawn from the real $p(\theta|y)$ we wanted to approximate.

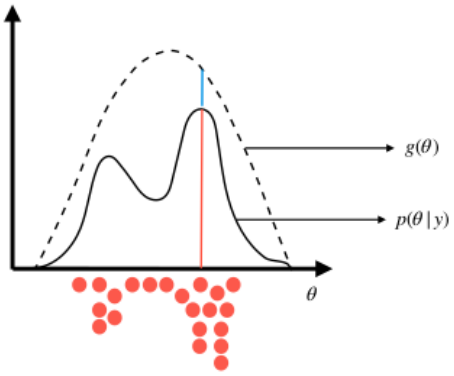
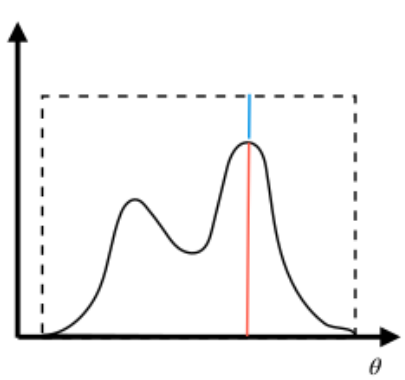


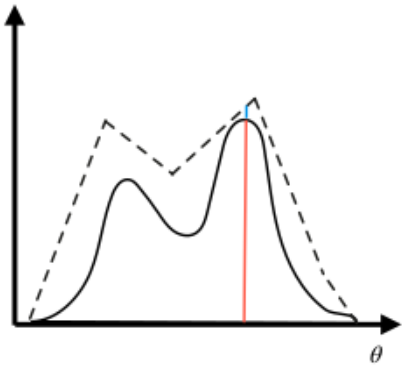
Figure 8.8: Rejection sampling method in order to sample from a distribution

The choice $g(\theta)$ plays an important role, specially wrt computational effort: the more different it is from $p(\theta|y)$, the higher is the rejection rate. However, if we have no information about the shape of $p(\theta|y)$ we should use a flat $g(\theta)$ (see Fig. 8.9a). On the other hand using the *trapezoidal approximation* to define $g(\theta)$ we can reduce this problem (see Fig. 8.9b).

We will now introduce the most efficient way to sample from distribution: **Markov Chain Monte Carlo** (MCMC). The general idea behind them is to start from an initial point in the parameter space θ_0, and starting from it to create a *random walk*. It is a *sequence* $\theta_0, \theta_1, ...\theta_t$ where each θ_i is draw from a given *transition distribution*, built such that the random walk converges to $p(\theta|y)$. That is to say we need to run the simulation long enough such that the distribution of the current draws becomes close enough to the stationary distribution $p(\theta|y)$. An example of MCMC is the so called **Metropolis algorithm** (see Fig. 8.10a, 8.10b). As in the general case we



(a) Rejection sampling using a "box" distribution. Efficiency is very low.



(b) Rejection sampling using $g(\theta)$ as a trapezoidal approximation for $p(\theta|y)$. Indeed we reach a high efficiency.

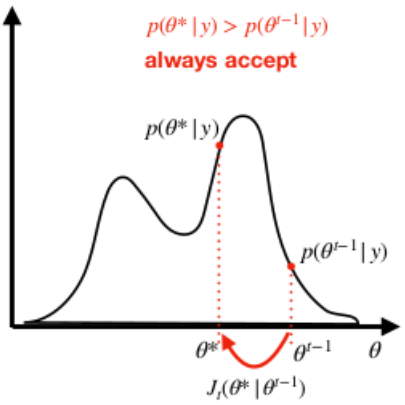
Figure 8.9

draw samples from a starting point θ_0 and then we iterate the following procedure for $t = 1, 2...$:

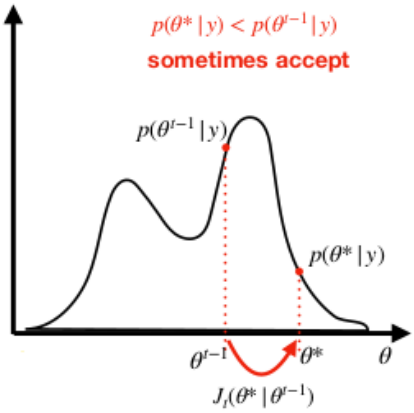
- sample a candidate point θ^* from a jumping distribution $J_t(\theta^*|\theta^{t-1})$. This distribution *must* be **symmetric**, i.e. such that $J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a) \quad \forall a, b, t$;
- compute the ratio of the densities:

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$$

- update $\theta^t = \theta^*$ with probability $\min(r, 1)$ (Fig. 8.10b), otherwise $\theta^* = \theta^{t-1}$ (Fig. 8.10a).



(a) If we jump from a less probable θ for to a more probable one, we *always accept* the drawn candidate θ^*.



(b) If we jump from a more probable θ for $p(\theta)$ to a less probable one, we *sometimes accept* the move according to a probability given by the ratio between the candidate $p(\theta^*|y)$ and $p(\theta^{t-1}|y)$.

Figure 8.10: Metropolis Algorithm.

Note that it is possible to show that for a *Metropolis Algorithm*, the Markov chain converges to a stationary distribution, which is $p(\theta|y)$. One may want to apply what we discussed so far and simulate *multiple chains* simultaneously with starting points dispersed in the parameter space. We keep monitoring the quantity (i.e. the parameter) of our interest and measuring variation between and within the different

sequences, until all of them **converge** to a **unique distribution**. Therefore, the initial part of the random walk has to be neglected since samples are strongly correlated, we refer to this fact as **burn-in**.

Let us now consider the following example: we want to sample from a posterior density that is a **bivariate unit normal**. The probability density function is therefore:

$$p(\theta_1, \theta_2|y) = \mathcal{N}(\theta_1, \theta_2|0, \mathbb{1}_2) = \frac{1}{2\pi} e^{-\frac{\theta_1^2+\theta_2^2}{2}}$$

where $\mathbb{1}_2$ is the 2×2 identity matrix.

As for the **jumping distribution** of the kind $J_t(\theta^*|\theta^{t-1}) = \mathcal{N}(\theta^*|\theta^{t-1}, \sigma^2\mathbb{1}_1)$, that we recall *must* be **symmetric**, we can choose a *bivariate normal* as well:

$$J(\theta_1^*, \theta_2^*|\theta_1^{t-1}, \theta_2^{t-1}) = \frac{1}{2\pi\sigma^2} e^{-\frac{(\theta_1^*-\theta_1^{t-1})^2+(\theta_2^*-\theta_2^{t-1})^2}{2\sigma^2}}$$

The only **parameter** we need to tune as for the *jumping distribution* is σ:

- **large** σ: the larger it is, the larger jumps will be allowed such that we will be able to explore a larger space of parameters. On the other hand, almost all proposal will be rejected. Hence, despite the large inefficiency, we will end up being stuck;
- **small** σ: the chain will be too slow and the algorithm will be inefficient;
- **best** σ: we use as a *rule of thumb* a condition over the *acceptance rate* that has to be in the interval $0.23 - 0.44$.

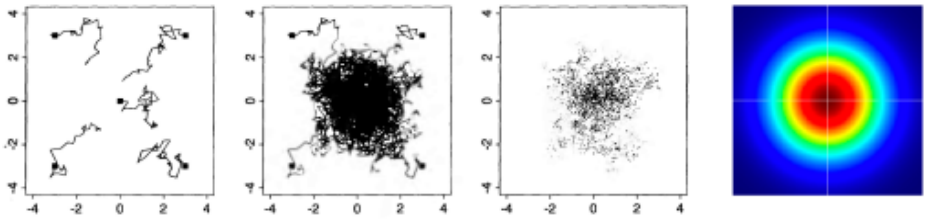


Figure 8.11: Markov Chain Monte Carlo at different stages of simulation. The distribution we use to sample from are bivariate Gaussians.

9

Outbreak Analysis

Let us apply all the knowledge we have acquired so far to a real world problem. We want to understand what we need to do when we observe an **outbreak** of an **emerging pathogen** about which we know *nothing*. Let us immerse ourselves in the following situation: to us all epidemiological characteristic, namely μ, β, generation time, are unknown. We will use COVID-19 as a *paradigmatic* case. For Fig. 9.1 one can tell that we achieved very good results in studying the disease as fast as possible: indeed all relevant quantities were computed in a couple of months.

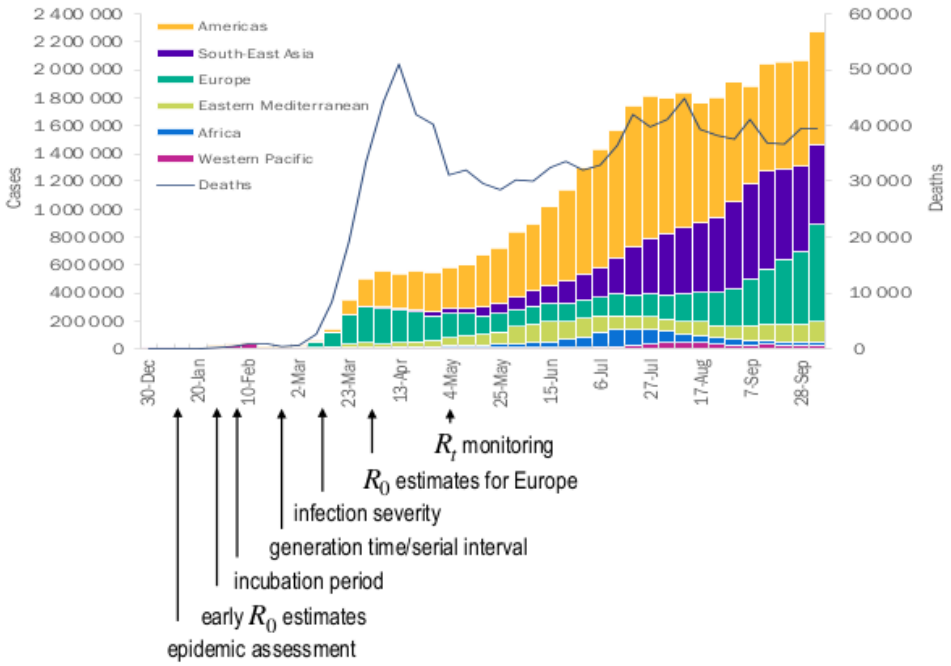


Figure 9.1: COVID-19 outbreak analysis Number of COVID-19 cases reported weekly by WHO Region, and global deaths, 30 December 2019. Arrows describe the moment at which the quantity reported has been computed. The epidemic assessment started from the letter written by Imperial College.

9.1 Basic Reproductive Ratio R_0 estimation

We will discuss now how to **process** the raw data in order to obtain, for instance, the **basic reproductive ratio** R_0. It is a really relevant quantity since we can retrieve the *final attack rate* if the virus spread: this is helpful in policy making to understand how much serious the disease can be, as well as how far we are from the epidemic threshold and how many infections/deaths we might expect.

9.1.1 R_0 from the early exponential growth

We will try to obtain R_0 using the *early exponential growth* typical of epidemic spread. This is indeed useful for **real time analyses** of highly transmissible infections (i.e. you see an exponential growth of cases).

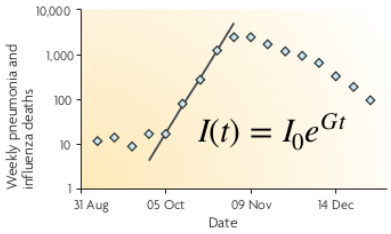
The estimate of R_0 will therefore computed by the means of the **exponential growth**, and obviously it is valid in the case we observe an epidemic. We will start to tackle this problem first introducing a simple version of the *Kermack and the McKendrick model*[1] according to which the **incidence** at the **early stages is** (Fig. 9.2a):

$$I(t) = I_0 e^{Gt} = I_0 e^{\mu(R_0-1)t} \tag{9.1}$$

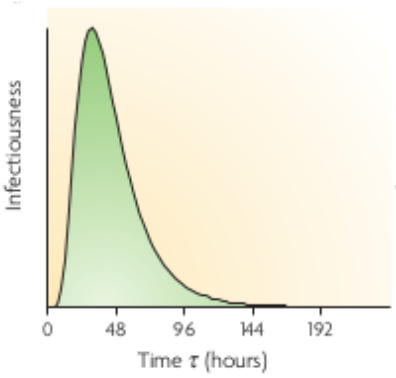
We must observe indeed a region where cases grow exponentially: here we fit using the curve just introduced. Reverting this last formula we are able to **estimate** R_0 from the *exponential growth factor* G:

$$R_0 = \mu^{-1} G + 1 \tag{9.2}$$

In this way we are **implicitly assuming** that *infectivity* β remains constant for the whole period of infection. It is indeed a strong, as well as unrealistic, assumption. Moreover, we assume that *recovery* is a Poisson process, whose waiting times are exponentially distributed with average μ^{-1}.



(a) Exponential early growth for SIR model.



(b) More general distribution for infectiousness, which might vary on time.

Figure 9.2

We define the **Generation time** T_c as the average time from infection of the infector to the infection of the infected. In this case, it holds that the average generation time is the average infectious duration $T_c = \mu^{-1}$, hence we can rewrite:

$$R_0 = T_c G + 1 \tag{9.3}$$

However, this means that **infectious periods** are **exponentially distributed**, which is quite *unrealistic*[2]. Another *unrealistic* aspect is the **lack** of **exposed period**, which is quite always present.

[1] The Kermack-McKendrick model is an SIR model for the number of people infected with a contagious illness in a closed population over time. It also assumes a completely homogeneous population with no age, spatial, or social structure.

[2] Wallinga, Lipsitch, How generation intervals shape the relationship between growth rates and reproductive numbers Proc. R. Soc B (2007) 274, 599

Let us consider the more general situation as possible where the *generation times* distribution is $w(\tau)$. We recall now that the **transmissibility generic function** $\beta(\tau)$ can be rewritten as function of the generating time as:

$$w(\tau) = \frac{\beta(\tau)}{R_0} \tag{9.4}$$

This means that there is a period in which we are more likely to infect someone once infected, i.e. we are more infectious. Obviously, this is the peak in the distribution $w(\tau)$ (see Fig. 9.2b). The other way around to see it is the following: the time needed in order to generate a secondary case, which might be in number more than one, distributes as in Fig. 9.2b.

Let us imagine to have a number $I(u)$ of newly infected individuals at time u. The **number of individuals they will infect** at time $t = u + \tau$ in a time interval $\delta\tau$ is a random process, and follows a **Poisson** distribution with mean:

$$I(u)\beta(\tau)\delta\tau = I(t-\tau)\beta(\tau)\delta\tau \tag{9.5}$$

One can easily see that the number of **newly generated** infected people *does* depend on the number of *previously infected* individuals. Formally, this is a **renewal equation** (Lotka-Euler equation)[3]:

$$I(t) = \int_0^\infty I(t-\tau)\beta(\tau)d\tau \tag{9.6}$$

Simple SIR model actually accounts for these last considerations, but it assumes that generation times distribution is exponential. We wanted to drop this assumption: we are indeed considering non-markovian epidemics. For this reason we had to introduce the aforementioned **integral equation**, which takes into account the exponential case too, and we could not write an exponential equation any more. Only for now, we are assuming well-mixed and therefore infinite population.

We want now to use the *renewal equation* in order to estimate R_0. This can be done however *only* when we observe an **exponential growth** of cases, according to which we can derive the *exponential growing factor* G. Hence we can write the infectious people at time $I(t)$ in function of infectious people at time $t-\tau$ as follows:

$$I(t) = I(t-\tau)e^{G\tau} \tag{9.7}$$

By using last equality, the renewal equation becomes:

$$I(t) = \int_0^\infty I(t)e^{-G\tau}\beta(\tau)d\tau \tag{9.8}$$

Dividing both sides by $I(t)$ and replacing 9.4 in the last equation, we end up with:

$$\frac{1}{R_0} = \int_0^\infty e^{-G\tau}w(\tau)d\tau \tag{9.9}$$

From which we can provide an estimate for R_0. One should have noted that it is the **Laplace Transform** of $w(\tau)$: therefore we are able to relate directly R_0 to the generation times $w(\tau)$ distribution, specially to its moments generating function via $\mathcal{M}(z) = \int_0^\infty e^{z\tau}w(\tau)d\tau$:

$$R_0 = \frac{1}{\mathcal{M}(-G)} \tag{9.10}$$

We recall now the **properties** of the *Laplace Transform* and its **epidemiological implications**:

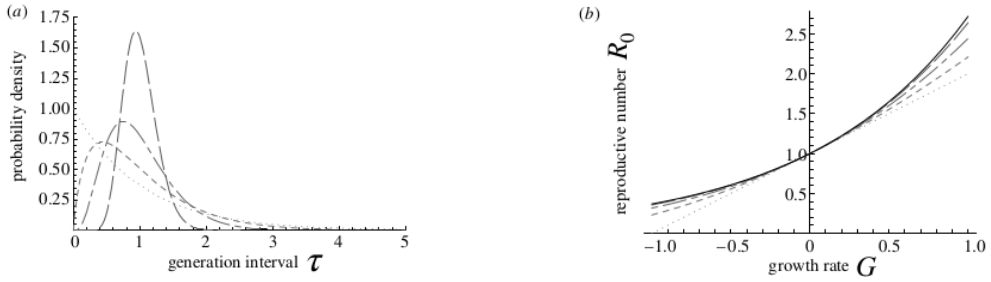
[3] a.k.a., borrowed from population dynamics. New births depend on people that were born 20-30 years ago: for instance their number (larger it is, the more births we expect), fecundity rate and how this parameter evolves during the whole lifespan of a person. Many models from ecology and demography can be applied to the epidemiological framework.

- $\mathcal{M}_{w(\tau)}(z) \geqslant 0 \implies R_0 \geqslant 0$ because the Laplace transform is *always* greater or equal than zero.
- $\mathcal{M}'_{w(\tau)}(z) = \int_0^\infty \tau e^{z\tau} w(\tau) d\tau > 0 \implies R_0$ is an increasing function of G, since the derivative of the Laplace transform is always greater than zero. Bigger the exponential growth, bigger is R_0.
- $\mathcal{M}_{w(\tau)}(0) = \int_0^\infty w(\tau) d\tau = 1 \implies$ if $G = 0 \implies R_0 = 1$ from normalization condition. If we observe a flat curve, i.e. no exponential growth, then $R_0 = 1$.

Let us assume that $w(\tau)$ is distributed as a Gaussian, that we recall is the least informative distribution[4]. We know from theory that its *Laplace Transform* is:

$$w(\tau) = \mathcal{N}(T_c, \sigma^2) \implies R_0 = e^{GT_c - (1/2)G^2\sigma^2}$$

This means that R_0 depends also on the variance of generation times distribution. Let is assume that the generation time is like the infectious period ($\langle T_c \rangle \sim \mu^{-1}$). For a given $\langle T_c \rangle$, if the variability σ^2 of the generation times is higher, then the exponential term becomes lower, hence leading to lower R_0 (see Fig. 9.3). For instance: if an individual remains infected for a shorter period, it is going to produce quicker more infections hence the exponential growth will become steeper. This fast chain will dominate over the other chain of transmissions.



(a) **(b)**

Figure 9.3: Different distributions to of generation times (9.3a) lead to different R_0 once we observe an exponential growth G (9.3b).

Whereas, if we consider a **Delta Distribution**, i.e. a well peaked distribution with no variance at all, its transform is:

$$w(\tau) = \delta(T_c) \implies R_0 = e^{GT_c}$$

that sets an **Upper bound** of R_0. This is the case where the infection duration is more or less equal for everybody.

In other words, adding some variability onto a fixed $\langle T_c \rangle$ generation time, a given R_0 will correspond to higher G. This can be formalized thanks to **Jensen's inequality**: the average of transformed stochastic variables is at least equal to the transformed average of those variables when the transformation is *convex*:

$$\int_0^\infty e^{z\tau} w(\tau) d\tau \geqslant e^{z\int_0^\infty \tau w(\tau) d\tau} \implies \mathcal{M}_{w(\tau)}(z) \geqslant e^{zT_c} \implies R_0 \leqslant e^{GT_c} \tag{9.11}$$

Recalling that τ is distributed according to some distribution $w(\tau)$, we assume that τ is the sum of two stochastic variables:

$$\tau = \tau_E + \tau_I$$

[4]It means that we know only the mean μ and the variance σ^2 of a distribution and nothing more.

where τ_E is distributed according to $g(\tau_E)$ and τ_I follows another distribution, namely $h(\tau_I)$. One can prove that the generation times distribution $w(\tau)$ can be rewritten as the *convolution* between g and h:

$$w(\tau) = g(\tau_E) * h(\tau_I) \tag{9.12}$$

and, its Laplace transform is:

$$\mathcal{M}_{w(\tau)} = \mathcal{M}_{g(\tau_E)} \times \mathcal{M}_{h(\tau_I)} \tag{9.13}$$

This can turn out to be useful if we want to add some complexity to our model[5] and consider the generation time as the sum of a **latent period**, during which *no infections* can be generated and that distributes as $g(\tau_E)$, and a second term that is an **infection period** τ_I where infections can be generated and distributes as $h(\tau_I)$. For instance, let as assume that $\tau_E \sim Exp(\varepsilon)$ $\tau_I \sim Exp(\mu)$. The basic reproductive ratio is defined as $R_0 = \frac{1}{\mathcal{M}(-\mathcal{G})}$: let us compute it analytically. We know that:

$$\frac{1}{R_0} = \mathcal{M}(-G) = \int_0^\infty e^{-G\tau}\omega(\tau)d\tau \tag{9.14}$$

Given $g(\tau_E) = \varepsilon e^{-\varepsilon\tau_E}$ and $g(\tau_I) = \mu e^{-\mu\tau_I}$

$$\begin{aligned}
\mathcal{M}(-G) = \mathcal{M}_{g(\tau_E)} \times \mathcal{M}_{h(\tau_I)} &= \int_0^\infty e^{-G\tau_E} g(\tau_E)d\tau_E \times \int_0^\infty e^{-G\tau_I} g(\tau_I)d\tau_I \\
&= \int_0^\infty e^{-G\tau_E}\varepsilon e^{-\varepsilon\tau_E}d\tau_E \times \int_0^\infty e^{-G\tau_I}\mu e^{-\mu\tau_I}d\tau_I \\
&= \varepsilon\int_0^\infty e^{-(G+\varepsilon)\tau_E}d\tau_E \times \mu\int_0^\infty e^{-(G+\mu)\tau_I}d\tau_I \\
&= \frac{\varepsilon}{G+\varepsilon} \times \frac{\mu}{G+\mu} = \frac{1}{R_0}
\end{aligned}$$

The **basic reproductive ratio**:

$$R_0 = \left(1+\frac{G}{\varepsilon}\right)\cdot\left(1+\frac{G}{\mu}\right) = 1 + \frac{G}{\varepsilon} + \frac{G}{\mu} + \frac{G^2}{\varepsilon\cdot\mu} \cong 1 + G(\varepsilon^{-1} + \mu^{-1}) \tag{9.15}$$

One should note that we have exactly got back to the result of the **SEIR** model.

One should note however that if we observe an **exponential growth**, we cannot conclude *anything* on the *Basic Reproductive Ratio* R_0, since we need both average and variance information of **generation time distribution** as well: this indeed helps to interpret the exponential growth (Fig. 9.3b).

In practice, when we have an **incidence time series** $\{y_t\}$ we need to:

- *Define a time window* where the **growth** is **exponential**.
- We make a **Poisson regression** of the incidence points, assuming that our process is Poisson and with expected value:

$$\log\left(\mathbb{E}(y_t|t)\right) = \alpha_0 + Gt$$

- We estimate through the *likelihood* both **initial conditions** and the **exponential growth**.

$$\mathcal{L}(\alpha_0, G|\{y_t\}) = \prod_{t=0}^{t_M} \frac{e^{y_t(\alpha_0+Gt)}e^{-e^{\alpha_0+Gt}}}{y_t!}$$

[5] What we observe in reality is that infectivity continuously evolves throughout time: sometimes it is less and sometimes is high. We can simplify and divide this behavior into two discrete stages: Exposed and Infected.

Initial conditions are fundamental, since we need them to properly understand and interpret G. Even though we want to infer a single parameter and α_0 has no specific meaning taken by itself, those are needed.

Clearly, the proper distribution of generation times is difficult to obtain analytically; therefore, we use only available information on the generation time distribution and make some assumptions in order to compute R_0. Many times the only information we have is its *average* T_c and *dispersion* σ, hence we can exploit the **Gaussian approximation**:

$$R_0 \simeq e^{GT_c - (1/2)G^2\sigma^2}$$

However, often the **distribution of serial intervals** is used as a *proxy* of the distribution of generation time.

Lecture 22. Friday 11th December, 2020. Compiled: Saturday 14th August, 2021.

9.1.2 R_0 from the cluster size

We will now provide another tool to estimate R_0: we will try to infer it from the **cluster size**. This is useful specially when *infections are not highly transmissible*, that is to say when we do **not** observe an **exponential growth** of cases, but only sporadic infections.

Let us discuss which are the possible infections with $R_0 < 1$, in particular what are their main features. These are for example *zoonotic infections*, such as $MERS$[6] that was transmitted through dromedaries and camels and whose mortality rate was high, or infections that are *close* to *eradication threshold*, such as measles. It is important to analyze such diseases and to understand how far we are from the epidemic threshold, since the higher the number of infected individuals, the higher the risk of a mutation in the pathogen or a decrease in vaccine uptake.

We want to understand how to deal with this problem: a first approach is to *assume* that the *few cases* we observe are **linked** to a **unique cluster**.

We define as a **cluster** all cases generated by the so called *index case*. A cluster is an entity like the one in Fig. 9.4, whose most important quantities are:

- **Index case**: infection caused by an external source.
- **Offsprings**: cases infected by the index case.

In the case where we deal with many clusters, the distribution of their size s depends on R_0 according to $P(s|R_0)$.

Figure 9.4: Estimation of R_0 using a cluster approach when observing a small amount of cases.

Some examples of $P(s|R_0)$ probability density might be, depending on the assumptions we make on the heterogeneity of the network:

[6] Middle East Respiratory Sindrome (MERS) coronavirus, 2013. Breban, et al. The Lancet 2013

- All infectious individuals behave equally and generate on average R_0 transmissions. The number of *offsprings* is k and are distributed as a Poisson:

$$k \sim \text{Pois}(k|R_0) \implies P(s|R_0) = \frac{(sR_0)^{s-1}e^{-sR_0}}{s!}$$

- Introducing a *continuous-time SIR* dynamics, since R_0 is below 1 stochastic effects are important. The size of the cluster therefore follows a *Markovian birth and death process* of the kind:

$$P(s|R_0) = \frac{(2s-2)!}{s!(s-1)!}\frac{R_0^{s-1}}{(R_0+1)^{2s-1}}$$

However, it might happen, specially when cases are few, that many of them are not reported: data we have is almost for sure biased in this sense. Hence we need to account also for **under-reporting**. Each case may go unobserved with probability p_{miss}: as a consequence if a cluster has real size s, we may observe a cluster of size $o \leqslant s$.

The **probability of observing a cluster** of size $o \geqslant 1$, given R_0, is Binomial wrt missed cases:

$$P(o|R_0, p_{miss}, o \geqslant 1) = \frac{\sum_{s \geqslant 0} P(s|R_0)\binom{s}{o}p_{miss}^{s-o}(1-p_{miss}^{o})}{1 - o(o=0|R_0, p_{miss})}$$

where $P(s|R_0)$, can be any of the aforementioned probabilities. And the **likelihood**:

$$\mathcal{L}(o_i|R_0, p_{miss}) = \left(\sum_i o_i\right)! \prod_i \frac{1}{o_i!} P(o|R_0, p_{miss}, o \geqslant 1)$$

where, as an example, the first factor sums over all the possible sizes of the cluster that we may have observed, and all the possible permutations ("!").

Exploiting the tools we have just introduced epidemiologists were able to provide an estimate for R_0. It is interest to notice that at early spread, in February this was $R_0 \simeq 2$. More likely the R_0 for the Eastern strain of Coronavirus was computed. Indeed a couple of months later scientists were able to estimate R_0 for the European strain, and it resulted to be $R_0 \simeq 3$. This can explain why in UK the Chinese strain was overtaken by the European one, despite it appeared earlier and number of infected people was way larger.

9.2 Incubation period estimation

One other relevant quantity we need to estimate as soon as possible is the **incubation period**. It is really important since, under some assumptions, we can relate it to the *generation times* distribution from which to compute R_0. In order to do it, high quality data is needed. However, if there is none, we can still provide an estimation by using similar diseases one, even though obviously this analysis might lead to wrong outcomes.

The **incubation time** is defined as the time elapsed between the **onset** of symptoms and the *infection time*, namely when the **exposure** occurred (Fig. 9.5). The former is easy to collect, for instance when a person has been visited by a doctor and start showing any symptoms. In addition, incubation time is also a measure of the delay in the response of restriction policies in infection curve. However, exposure time is very difficult to retrieve, nonetheless is very important: pre-symptomatic phase relevance is even higher for COVID-19, since pre-symptomatic individuals turn

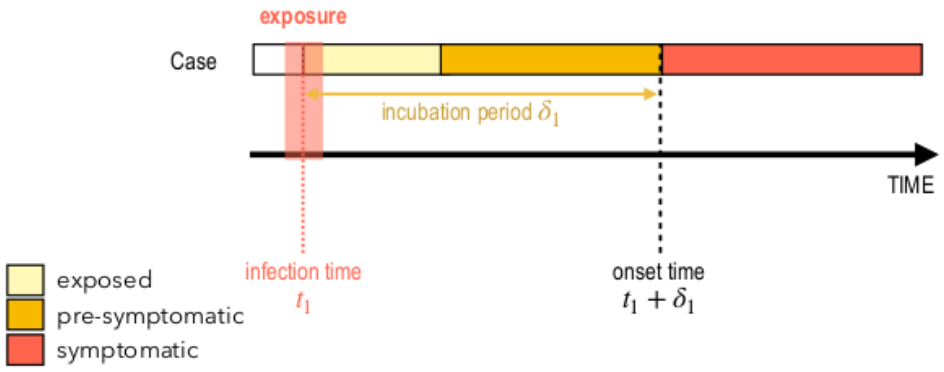


Figure 9.5: Incubation period analysis.

out to be really infectious. Some ways deal with this problem are through **contact tracing**, **case investigation**. Once a case is confirmed, contacts are investigated. They are contacted, isolated and go through clinical and virological assessment. This allows us to collect infector-infected pairs and more likely when contact occurred.

One other approach one may want to follow to estimate the *incubation period* is to proceed with the **analysis of travelling cases**. With referral to Fig. 9.6 and COVID-19 outbreak: Backer et al.[7] analyzed 88 cases detection starting from January, 20th through January, 28th. Their travel history (to and) from Wuhan was known, as well as their symptom onset date. During this early stage of the epidemic, it is most likely that travellers were infected in Wuhan. Consequently, their time spent in Wuhan can be assumed to be the duration of exposure to infection without any contact tracing procedure. As said, we know *for sure* the date of the onset of symptoms, but we need to infer when infection occurred. One should note that the shorter the stay in a risky area, the more precision we have in inferring the duration of infection period. After providing these estimations, data was fitted and compared using 2-parameter continuous distributions supported on a semi-infinite intervals, such as Gamma, Weibull and Log-normal distribution. Later, they proceeded to maximize the Likelihood and, by the means of a strictly positive flat prior for the two parameters, since there was no guess about their value, they tried to infer them. The most credible estimation for the **incubation period** was:

$$\text{incubation period} \sim 6.4 \text{ days} \qquad \text{C.I. } 95\%[2.1, 11.1] \text{ days} \tag{9.16}$$

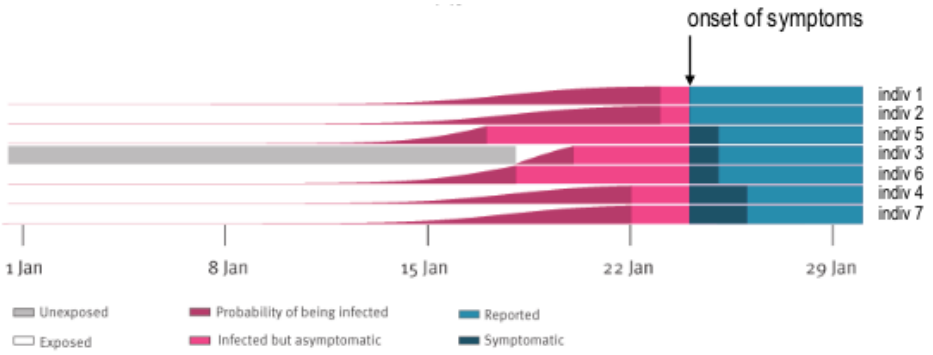


Figure 9.6: Pioneer analysis of travelling cases for COVID-19 outbreak.

[7]Backer, Klinkenberg, Wallinga. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, Euro Surveill. 2020;25(5).

9.3 Generation time and serial interval estimation

Let us recall the definition of the **generation time**: it is the time elapsed between the moment when the infector was infected and the moment when he infects someone as one can see in Fig. 9.7). It is indeed really difficult to obtain, nonetheless it can be computed thanks to the **serial interval** T_s that are easier to compute. It is the time *elapsed* between the *onset of symptoms* for two individuals, where we assume one to be the infector and the other one the infected. Therefore, the serial interval is a random variable that is linked to both generation times distribution (τ) and the incubation periods distribution (δ) in the following way:

$$\tau_S = \overbrace{(t_2+\delta_2)}^{\text{onset time case 2}} - \overbrace{(t_1+\delta_1)}^{\text{onset time case 1}} = \underbrace{(t_2-t_1)}_{\text{serial interval}} + (\delta_2-\delta_1) = \tau + (\delta_2-\delta_1) \quad (9.17)$$

where we sample $\tau \sim w(\tau|\theta_2)$, $\delta \sim g(\delta|\theta_1)$, $\tau_s \sim f(\tau_s|\theta_1,\theta_2)$. This indeed is true *on average*, and can be done since we assume that the average generation time and the average infectious duration are equal. Therefore, **serial interval** is often used as a **proxy** for the **generation time**, but with a warning! This argument is valid since $\langle\delta_1\rangle = \langle\delta_2\rangle$, therefore $\langle(\delta_2-\delta_1)\rangle = 0$, but we should recall that $\tau_S \neq \tau$: *on average* they are the same ($\langle\tau_s\rangle = \langle\tau\rangle$), but their *variance* is different ($\sigma_{\tau_S} > \sigma_\tau$) [8]!

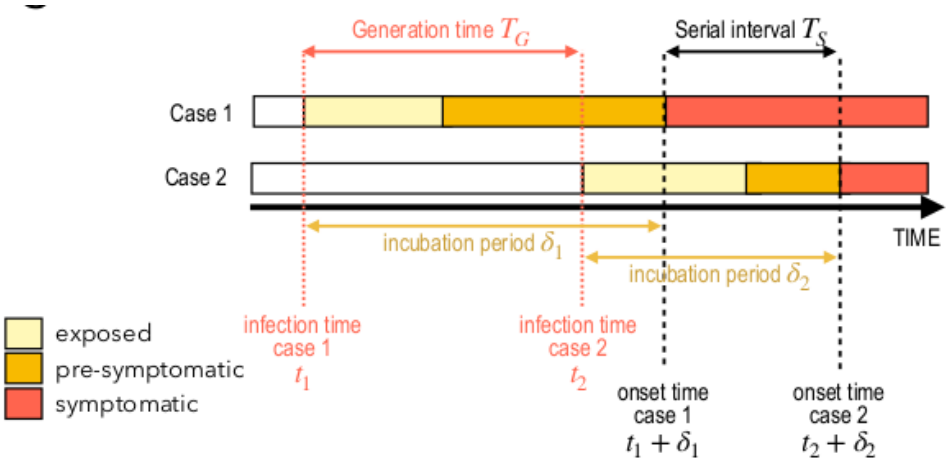


Figure 9.7: Generation time and serial interval relation.

Moreover, the variance of the generation times distribution f is greater than the one of w of infectious duration: indeed we can observe infections even caused by pre-symptomatic individuals: **pre-symptomatic phase** is important for transmission. In addition, variance of f is also greater than the variance of G and this might lead to an underestimation of R_0 since it holds that, relating the latter to the serial interval distribution, $R_0 \simeq e^{GT_G-(1/2)G^2\sigma^2}$.

As an example, let us consider the case of COVID-19 data and proceed with the estimation of the generation times distributions. Researchers[9] reported 91 confirmed cases in Singapore and 135 in Tianjin, and relied their paper on previous estimates of incubation period, whose distribution is:

$$\delta \sim g(\delta|\theta_1) = \text{Gamma}(\delta|\theta_1) \quad (9.18)$$

Data we are talking about consisted in information about infector and infected pairs. We know that **generation times** is $\tau_S = \tau + (\delta_2 - \delta_1)$, therefore distributes according

[8] Indeed, $Var(X+Y) = Var(X) + Var(Y)$.

[9] Ganyani, Kremer, Chen, Torneri, Faes, Wallinga, Hens. Estimating the generation interval for coronavirus disease (COVID-19) based on symptom onset data, Euro Surveill. 2020;25(17):pii=2000257

to the convolution between the two distributions h and w:

$$f(\tau_S|\theta_1,\theta_2) = \int_{-\infty}^{+\infty} w(\tau - x|\theta_2)h(x|\theta_1)dx = \mathcal{L}(\{\tau_{s,i}\}|,\theta_2\theta_1) \tag{9.19}$$

where $x = \delta_2 - \delta_1$, $x \sim h(x|\theta_1)$ and $\{\tau_{s,i}\}$ is the set of generation intervals observed.

Once we have proceeded to the estimation of θ_2 we can numerically compute:

$$P(t_2 < t_1 + \delta_1) = P(\tau_S < \delta_1) \tag{9.20}$$

that gives us the **proportion of cases generated by pre-symptomatic transmission**. The final results are the following for the COVID-19:

$$\text{generation time } = 5.20 \text{ days} \qquad \text{C.I. } 95\% \ [3.78 - 6.78] \text{ days} \tag{9.21}$$

and the *proportion of pre-symptomatic transmission* being around 50%, that is a high number. This implies that infected people tend to infect more if they are not reported, since this number is not negligible. A natural consequence is that, when we spot a case, it will have already infected almost half of the people it would infect during the whole infectious period.

The opposite case actually tells us why Ebola did not spread: despite a really high mortality rate $(60 - 70\%)$, the infectiousness was not constant in time and was proportional to the time individuals stay infected as long as with the severity of the symptoms. Those individuals could be successfully isolated in time as soon as any symptom shows up, thus avoiding the spread. In this case, **isolation** for infected individuals is a successful solution to stop the disease.

9.4 Infection severity

One of the most important quantities we may want to compute is the **infection severity**. Indeed, we need to know the full spectrum of symptoms including the ones of **pauci-symptomatics** and **asyntomatics**. It gives us a key to interpret data and observations, moreover allows us to estimate the **actual number of infections** and its **cumulated**: this might be helpful to understand how many people are susceptible to the disease at the moment. Another reason to explain why it is important is if, once recovered, we acquire some sort of immunity, even temporary, actual susceptible people become less in number. Moreover, it is important to compute the **infection fatality ratio** and to provide **projections on hospital needs**: we will know how many people would need ICU or hospitalization, based on the fraction of people that develop mild or more serious illness. These fractions are normalized wrt *total* number of infected people, including asymptomatic, so one should easily understand how much hard is to compute these ratios.

On the other hand this is a really difficult task to estimate these proportions, but this goal can be pursued through **contact tracing studies**, routine testing, or cohorts[10]. Indeed these are actually very difficult and expensive procedures, specially in terms of resources, and may require even formation for people which were assigned by these task. Moreover some natural experiments could be done. That is the case of the *Diamond princess*, a cruise ship in Japan where some people were tested positive. Therefore, others were quarantined on the boat for some time: eventual infections were registered, reported their eventual symptoms and finally led to the hospital. However, this brought someone to raise some ethical objections over this scientific non conventional approach. Another experiment was due to repatration flights, where

[10]We choose a sample of people/volunteers among a population, and proceed to follow their evolution in time in terms of infected, recovered, times distributions and so on.

people were tested in airports while returning home in addition to the reporting of their eventual symptoms.

In conclusion, the *true proportion* of asymptomatics of COVID-19 is still not certain ($\sim 20\% - 50\%$) and there is a strong dependence on age.

9.5 Reproductive ratio R_t

Let us compute the most relevant quantity of the epidemiological process, namely the **Reproductive ratio** R_t. This is different from the **basic reproductive ratio** R_0 and we recall that the latter is average number of cases an infectious individual infects in a *fully susceptible population* during the course of his/her infectious period. Moreover, this is *computed at the initial stage of an outbreak*. However, this is not realistic, being the number of susceptible dependent on time, for instance due to acquired immunity.

Instead, the **Reproductive ratio** R_t is the **average number of cases** an infectious individual infects **at a given time** t during the course of his/her infectious period. This is the *natural extension* of R_0 to the later outbreak stage. It is an **indicator** of how the transmission potential of the epidemic evolves in time.

The **dependence in time** has to be introduced to take into account that as the outbreak unfolds the population is not fully susceptible anymore. We retrieve the simple version of the Kermack and McKendrick model, where the new infections:

$$I(t+\delta t) = I(t) e^{\left[\mu \int_t^{t+\delta t} \left(R_0 \frac{S(t')}{N(t')} - 1\right) dt'\right]} \tag{9.22}$$

That in turn can be locally approximated to:

$$I(t+\delta t) \simeq I(t) e^{\delta t \mu (R_t - 1)} \tag{9.23}$$

where $R_t = R_0 \frac{S(t)}{N(t)}$.

In reality things are more complex: R_t does not change only due to the depletion of susceptible (immunity building): it might change as effect of interventions, behavioural change of population ($\langle k_t \rangle$ term):

$$R_0 = \frac{\beta \langle k \rangle}{\mu} \longrightarrow R_t = \frac{\beta \langle k_t \rangle}{\mu} \frac{S(t)}{N(t)} \tag{9.24}$$

Now, we drop the assumption of the simple SIR and put ourselves in a more general framework. We will follow two different paths and interpretations to compute R_t.

Cori method

The first one is the method used almost in all Western countries, and was developed by **Cori** et al.[11]. The main pro is that it captures immediate changes in number of contacts thanks to lockdown or other restrictions, and this makes it really useful. It starts from a generic generation time distribution, namely $w(\tau)$, and it is based on the Lotka Euler equation:

$$I(t) = \int_0^\infty I(t-\tau) \beta(\tau) d\tau \tag{9.25}$$

where $\beta(\tau) = w(\tau) R_0$. One should note that the number of infected people at time t depends on the number of infected at time $t - \tau$ and on the model parameters at time τ!

[11] Cori, Ferguson, Fraser, Cauchemez, A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics, American Journal of Epidemiology, 178, 2013.

Let us generalize the last expression. In order to do so, we make the assumption that the reproductive ratio varies in time, being the infectiousness time dependent as well $\beta(\tau, t) = w(\tau)R_t$. Therefore introducing R_t:

$$I(t) = \int_0^\infty I(t-\tau)w(\tau)R_t d\tau \tag{9.26}$$

note as it was introduced the dependence over the absolute time t through the variable R_t.

Reverting the last equation and discretizing the time we obtain:

$$R_t = \frac{I_t}{\sum_{s=1}^{t} I_{t-s} w_s} \tag{9.27}$$

where the denominator indicates the total infectiousness of individuals that are infectious at the time t. According to this **interpretation**, R_t *is the average number of secondary cases that each infected individual would infect if the* **conditions remained as they were at time t**. This method is called "**real time method**", since it links the actual situation to the past one through the generation times distribution finally providing an estimation for R_t. Given the total number of newly infected people today we can assume that background situation will not change in the close future, finally trying to predict the future number of infections that is function of R_t. The R_t analysis in France is illustrated in Fig. 9.8.

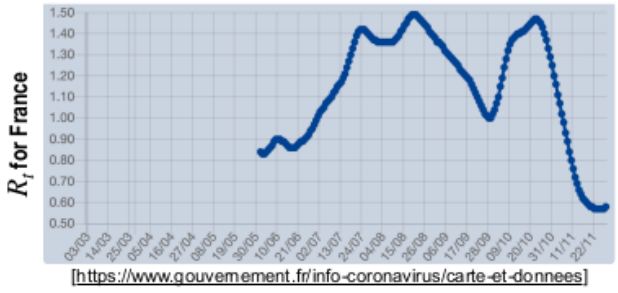


Figure 9.8: R_t analysis for France throughout 2020. Changes are mainly due to restriction and variation of $\langle k_t \rangle$ and most likely not due to immunity building.

Wallinga method

The second interpretation we introduce is the one made by **Wallinga** et al.[12]. It is used to infer *who infected whom* from available information. When we have an incidence curve, the **only information** regarding a case is the **date** in which a case was recorded. Hence, the relative **probability** p_{ij} that case i is infected by case j, given the time elapsed $t_i - t_j$ depends on the **generation interval** and assuming a case is registered the date in which it was infected is:

$$p_{ij} = \frac{w(t_i - t_j)}{\sum_{i \neq k} w(t_i - t_k)} \tag{9.28}$$

The denominator denotes all the case we have, and acts as a sort of normalization term. Sometimes, however, this is not realistic since there might be some delays in reporting the cases. We now introduce the **cohort reproduction number**:

$$R_j = \sum_i p_{ij} \tag{9.29}$$

[12]Wallinga, Teunis, Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures, American Journal of Epidemiology, 160, 6, 2004.

it counts the *average number of secondary transmissions caused by a cohort* that was infected at time step t. It is the infection potential of a cohort (might be even a single individual) at time t. We recall that a cohort is *not* a cluster, but is a group of cases that we follow from now on in the future, as a sample individuals. This method takes into account naturally the variability in the transmission potential of all individuals, since we are not making any assumption for it. We are trying to quantify **number of transmission generated by cases at time** t.

The **price to pay** for using this method is that it can be used **only retrospectively**: we are trying to compute when the secondary cases, generated by the infected at time t, have already been infected.

Method comparison

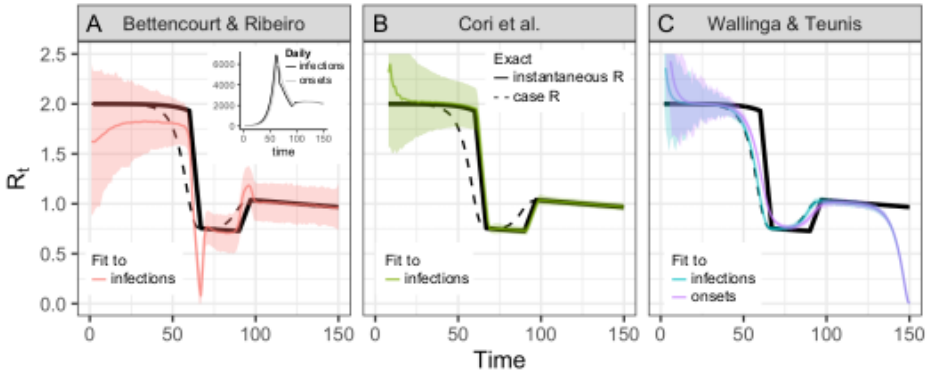


Figure 9.9: Difference between Bettencourt, Wallinga and Cori approaches for computing R_t. We see that Wallinga's has a cut-off. This behavior was expected since this approach analyzes future cases that are manually set to 0, nonetheless this kind of estimation is wrong. Cori's instead is more realistic.

We now want to make a little comparison between the two above described models (Fig. 9.9).

Cori describes the number of new infections at time t and link them to *past infections*. Hence, it **look backwards in time**. On the other hand, **Wallinga** relates the number of today's infections at time t to the *future cases* they will generate. Hence, it **looks forward in time**. The last model however cannot be used in real time analysis, but it sticks more to the R_t definition.

Let us use these approaches with a more practical task. Let us compute the actual life span of some individuals that was born in 2013. Using Wallinga's retrospective approach, we would need to wait until all individuals die out, and then proceed to estimation. Conversely, using Cori's approach, we could estimate the same quantity by assuming that death rates in the future will be similar to present ones. It is a more physical approach: indeed it takes into account actual conditions.

In reality, we do not have as data the time of infection, and generally we **report only delayed effects** (symptoms) of some events that had occurred some time in the past (Fig. 9.10). So, every report contains an **intrinsic delay** within itself, and even a delay of a single day can be risky for certain type of diseases. In this framework, hospital data is currently the best one since it relies on an uniformity of testing. In addition, outpatient testing data [13] is useful, but its availability depends on many variables such as "population testing" policies for a given country. In conclusion, if we want to compute R_t we either build a **compartmental model** and, sticking

[13] People that get tested under medical prescription or needs to travel and therefore have to be tested.

to it we estimate parameters using *maximum likelihood*, or alternatively, we build a **statistical model** and try to infer infection times and all related quantities by the means of *deconvolution*. Another last possibility is to build a model accounting for **latent states**, where the observables become *hospital admissions date* and the latent state is the one that immediately follows the infection. However, since we do not observe the time of infection, we can still use maximum likelihood to infer R_t.

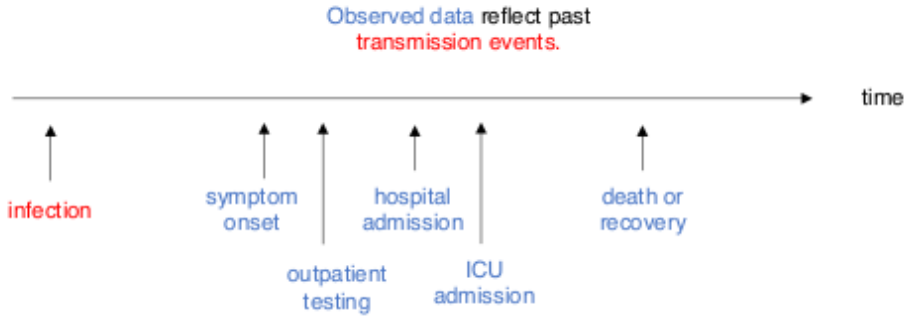


Figure 9.10: Generation time and serial interval relation.