LECTURE NOTES OF LIFE DATA EPIDEMIOLOGY

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

Edited By

Alice Pagano

The University of Padua

Contents

Part I

Meloni's Lectures

1

Basics Definitions and Compartmental Models

Some models are wrong, but most of them are useful.

– Unknown author

Lecture 2.
Friday 2nd October, 2020.
Compiled: Wednesday 11th November, 2020.

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 deal with 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 that we have. This is the main idea behind the computation of R_0. This parameters 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, we did not experienced an outbreak of this disease, so that is not the only parameters 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

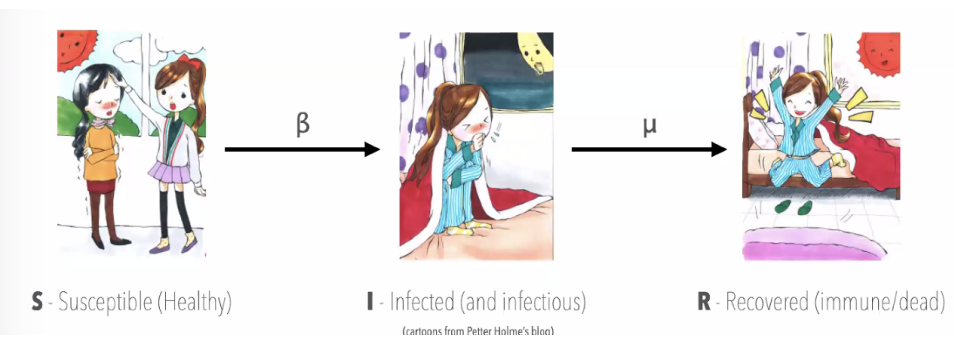


Figure 1.1: Classification of infected population in three different stages of the disease.

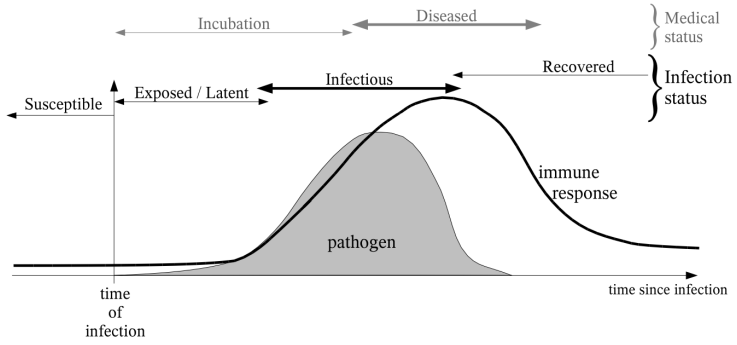


Figure 1.2: A sketch of the time-line of infection, showing the dynamics of the pathogen (grey area) and the host immune response (black line) with the labeling for the various infection classes: **S**usceptible, **E**xposed, **I**nfectious, and **R**ecovered. Note that the period when symptoms are experienced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

same of R_0 but varying wrt time. Hence, it is the average number of secondary cases that a single case produces in a population at time t.

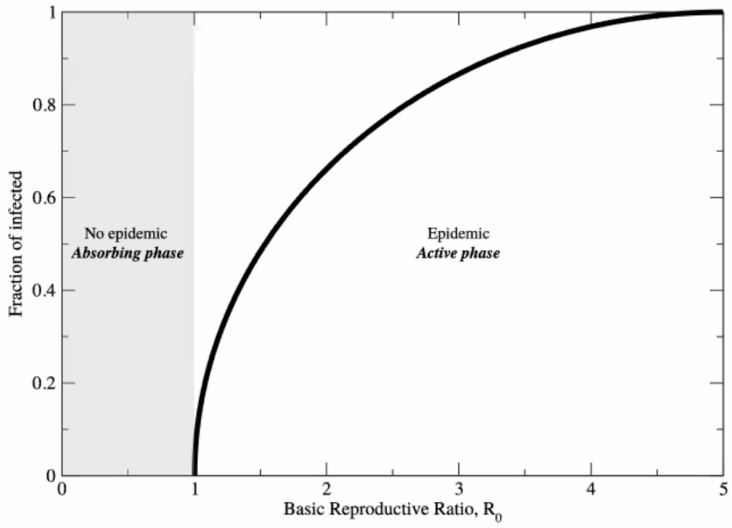


Figure 1.3: Fraction of infected vs basic Reproductive Ratio, R_0.

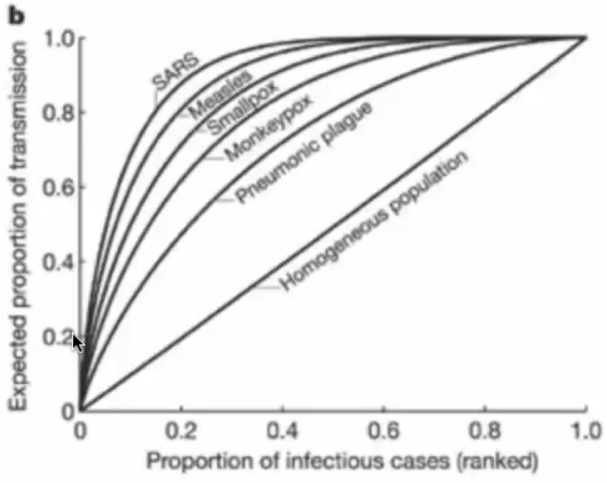


Figure 1.4: Figure from: Lloyd-Smith et al. Nature 438, 355–359 (2005).

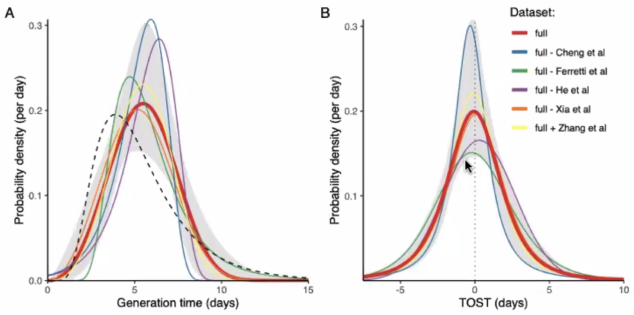


Figure 1.5: Figure from: Ferretti et al. `https://www.medrxiv.org/content/10.1101/2020.09.04.20188516v1`

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: Wednesday 11^{th} November, 2020.*

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 **equation** 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.

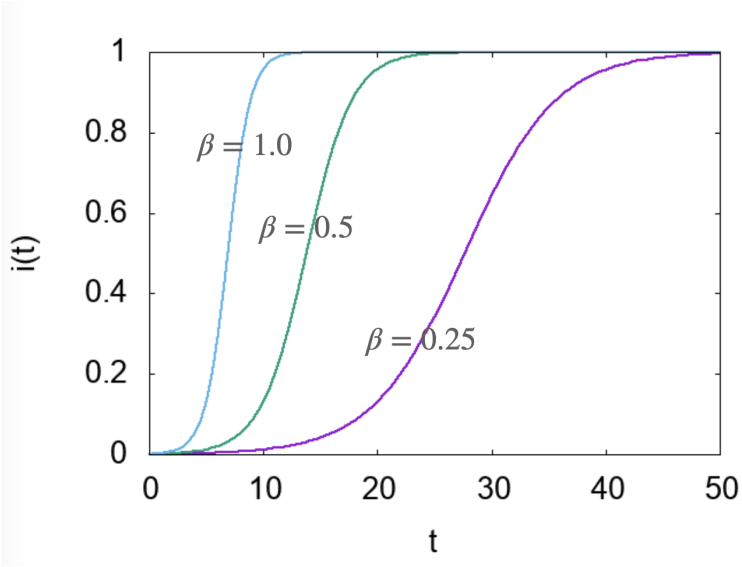


Figure 1.6: Plot of the solution of the SI model for different β.

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:

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

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:

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

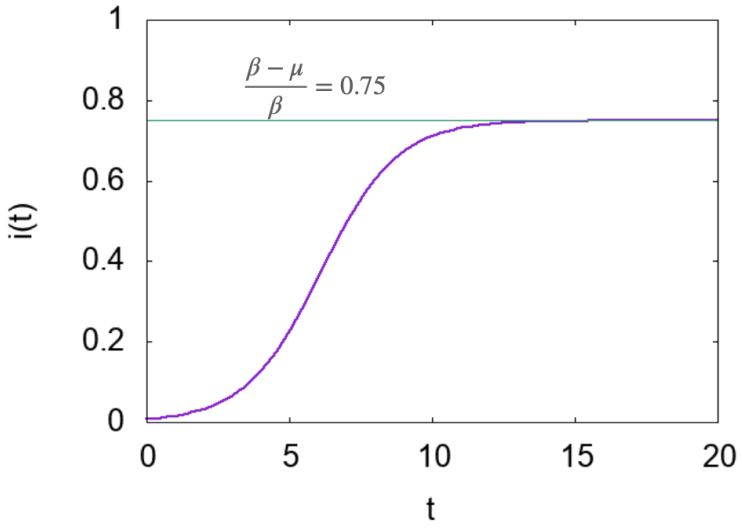


Figure 1.7: Plot of the solution of the SIS model.

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).

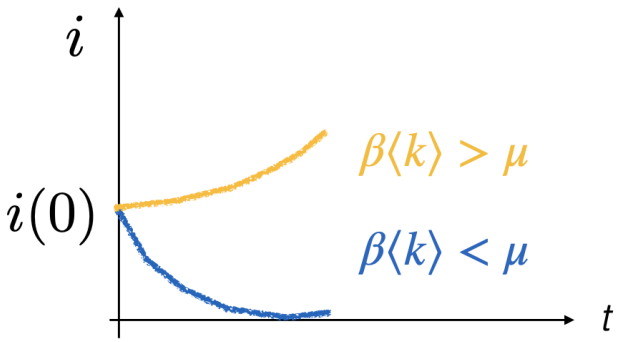


Figure 1.8: Initial transient for the SIS model.

One very important thing is that considering the **steady state** we can have two possible behaviors:

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

and we have that:

$$i > 0 \iff \beta > \beta_c = \frac{\mu}{\langle k \rangle} \tag{1.8}$$

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.

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.

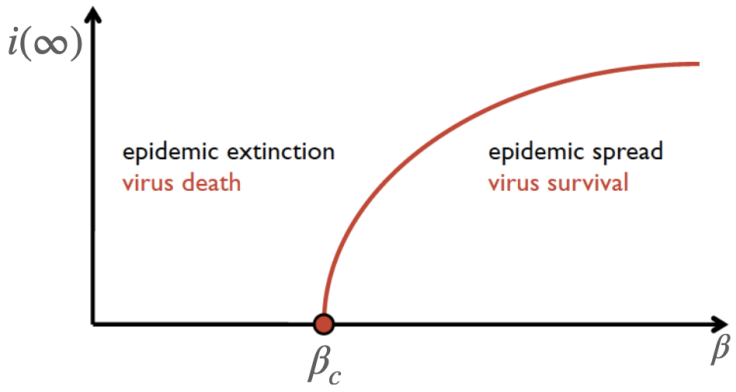


Figure 1.9: Epidemic diagram.

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 need a minimum of infected people and the disease is going to spread. Considering for instance the case of the SIS model:

$$R_0 = \frac{\beta \langle k \rangle}{\mu} = 1 \tag{1.9}$$

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:

$$\begin{aligned} S + I &\xrightarrow{\beta} I + I \\ I &\xrightarrow{\mu} R \end{aligned} \tag{1.10}$$

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:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= -\beta \langle k \rangle si \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \overbrace{\beta \langle k \rangle si}^{\text{New infections}} - \overbrace{\mu i}^{\text{Recovery}} \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= \mu i \end{aligned} \tag{1.11}$$

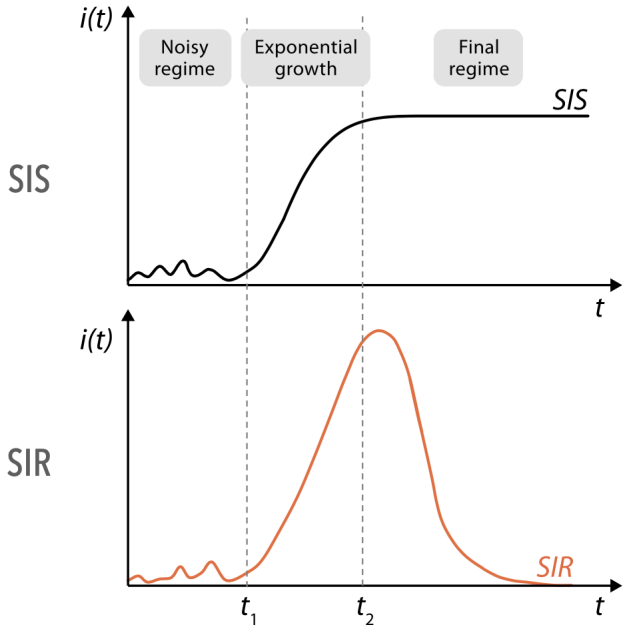


Figure 1.10: Epidemic regimes.

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, $r \ll 1$ so that:

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

and the result we find is again:

$$\beta > \beta_c = \frac{\mu}{\langle k \rangle} \tag{1.12}$$

Since we are able to obtain an analytic expression for S and I in this SIR model, we want to study what is the behavior for large times ($t = \infty$). 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:

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

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

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 $1/\alpha$ years^{-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:

$$\frac{\mathrm{d}s}{\mathrm{d}t} + \frac{\mathrm{d}i}{\mathrm{d}t} + \frac{\mathrm{d}r}{\mathrm{d}t} = 0 \tag{1.13}$$

Our equations become then:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= \alpha - \beta si - \alpha s \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \beta si - \mu i - \alpha i \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= \mu i - \alpha r \end{aligned} \tag{1.14}$$

where the **infectious period** is:

$$\tau = \frac{1}{\alpha + \mu} \tag{1.15}$$

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:

$$R_0 = \frac{\beta}{\alpha + \mu} \tag{1.16}$$

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:

$$i^*[\beta s^* - (\mu + \alpha)] = 0 \tag{1.17}$$

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:

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

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

$$(s^*, i^*, r^*) = \left(\frac{1}{R_0}, \frac{\alpha}{\beta}(R_0 - 1), 1 - \frac{1}{R_0} - \frac{\alpha}{\beta}(R_0 - 1)\right) \tag{1.18}$$

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: Wednesday 11th November, 2020.

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 ($w \rightarrow \infty$), 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:

$$\begin{aligned} S + I &\xrightarrow{\beta} I + I \\ I &\xrightarrow{\mu} R \\ R &\xrightarrow{w} S \end{aligned} \tag{1.19}$$

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

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= \alpha + wr - \beta si - \alpha s \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \beta si - \mu i - \alpha i \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= \mu i - wr - \alpha r \end{aligned} \tag{1.20}$$

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**:

$$\tau = \frac{1}{\alpha + \mu} \tag{1.21}$$

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

$$R_0 = \frac{\beta}{\alpha + \mu} \tag{1.22}$$

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 infectiuous 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.

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

$$\begin{aligned} S + I &\xrightarrow{\beta} I + E \\ E &\xrightarrow{\sigma} I \\ I &\xrightarrow{\mu} R \end{aligned} \tag{1.23}$$

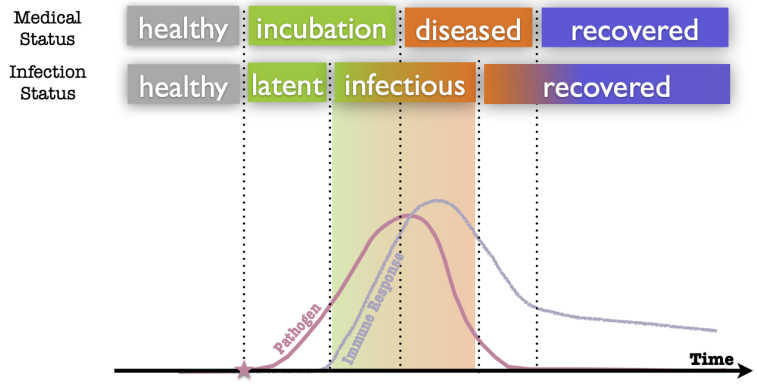


Figure 1.11: Difference between infection status and medical status.

with the equations:

$$\begin{aligned}\frac{\mathrm{d}s}{\mathrm{d}t} &= \alpha - \beta si - \alpha s \\ \frac{\mathrm{d}e}{\mathrm{d}t} &= \beta si - (\alpha+\sigma)e \\ \frac{\mathrm{d}i}{\mathrm{d}t} &= \sigma e - (\alpha+\mu)i \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= \mu i - \alpha r\end{aligned} \tag{1.24}$$

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

The **endemic state** is:

$$\begin{aligned}s^* &= \frac{(\alpha+\mu)(\alpha+\sigma)}{\beta\sigma} = \frac{1}{R_0} \\ e^* &= \frac{\alpha(\alpha+\mu)}{\beta\sigma}(R_0-1) \\ i^* &= \frac{\alpha}{\beta}(R_0-1)\end{aligned} \tag{1.25}$$

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

The $\mathbf{R_0}$ factor is:

$$R_0 = \frac{\beta\sigma}{(\alpha+\mu)(\alpha+\sigma)} \tag{1.26}$$

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:

$$\begin{aligned}i_{SEIR}(t) &\approx e^{\left(\sqrt{4(R_0-1)\sigma\mu+(\sigma+\mu)^2}-(\sigma+\mu)\right)t/2} \approx i_0 e^{\left(\sqrt{R_0}-1\right)\mu t} \\ i_{SIR}(t) &\approx i_0 e^{(R_0-1)\mu t}\end{aligned} \tag{1.27}$$

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).

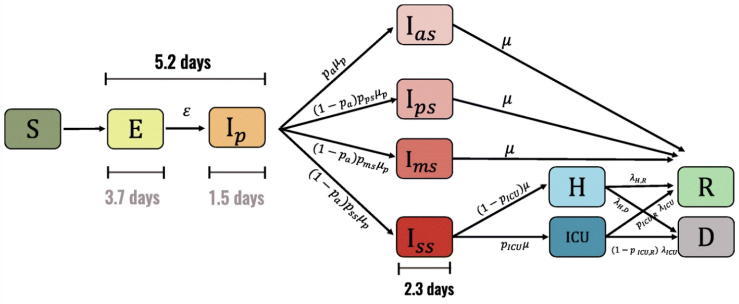


Figure 1.12: Model for Covid-19.

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 and **endemic** (meta-stable) **state**:
$$i = \frac{\beta - \mu}{\beta}$$
and we reach a sort of **dynamical equilibrium**. We can define an **epidemic threshold**:
$$\beta > \beta_c = \frac{\mu}{\langle k \rangle}$$

- 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:

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

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:

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

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 trough 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: Wednesday 11th November, 2020.

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**.

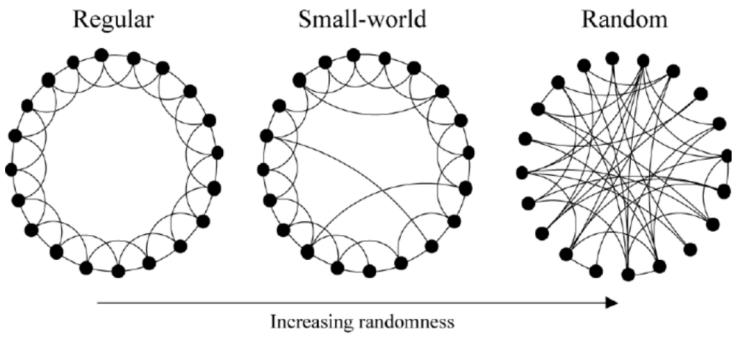


Figure 2.1: Idea of Watts and Strogatz model.

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:

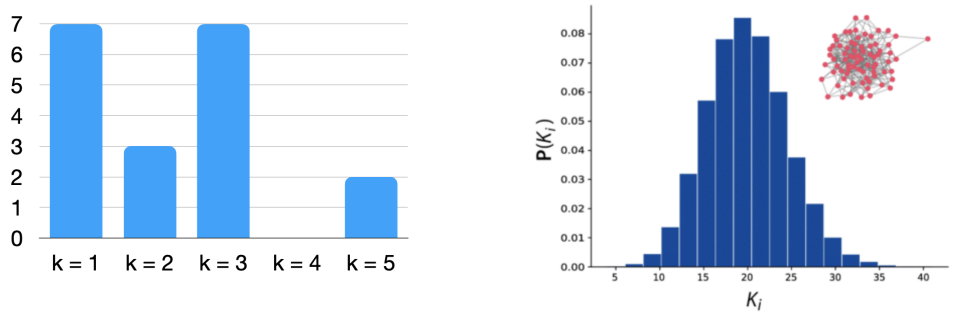


Figure 2.2: **Left:** degree distribution in a small network. **Right:** degree distribution in a network with random connections.

- 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 (N, p) distribution 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** we were looking for:

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

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.

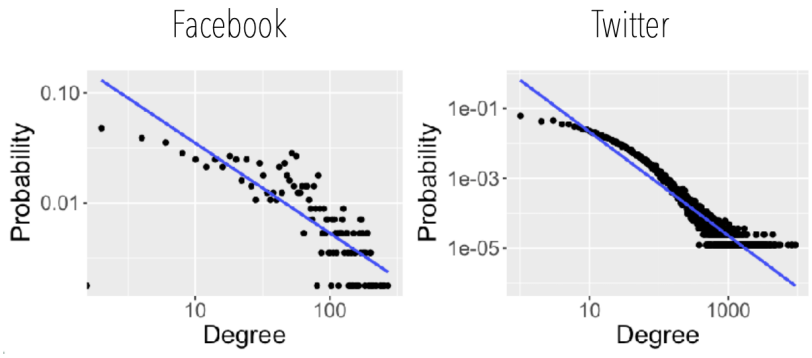


Figure 2.3: Real network of Facebook and Twitter.

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. 2.4, as an example, we shown 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

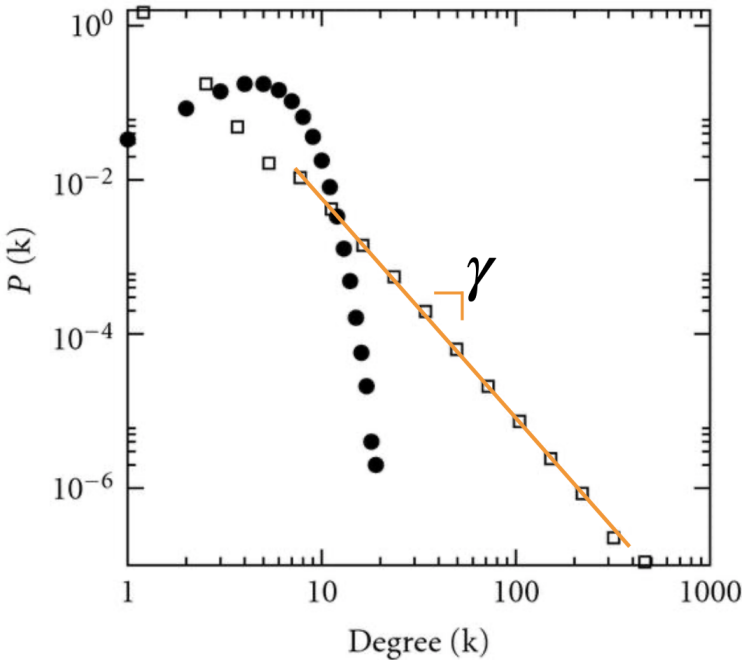


Figure 2.4: Difference between random networks and scale-free networks.

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:

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

Thus, when:

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

Since in most of networks $\gamma \sim 2-3$, so it is easily to understand that k_max scales **sub-linearly** with N, but still way faster than random graphs. 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:

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

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

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:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{2.9}$$

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

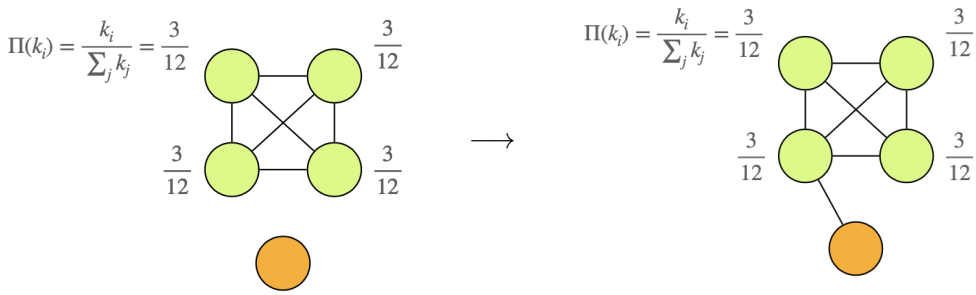


Figure 2.5: Example of Barabási-Albert algorithm.

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:

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

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 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. 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** where *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.

Remark. Let us recap 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: Wednesday 11th November, 2020.

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 **heteroegeneous 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 given by the sums:

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

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 distinct 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 of being infected $(1-\rho_k(t))$ 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.

For **random networks**, e.g. 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:

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

where we simply applied the definition of conditional probability. Note that $P(k')$ is the generic probability of getting a connection at random, times k', which is the number of connection that we pick up (namely the 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:

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

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 k', times the probability of being infected, Finally, we then we sum over all the possible degrees. While, expression on 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:

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

  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:

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

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

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 because,**if** our **network is big 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.

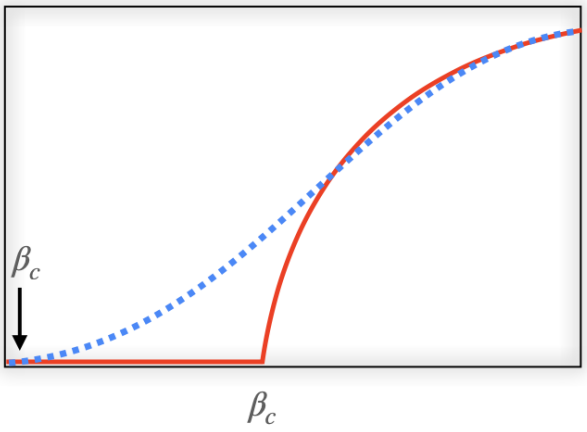


Figure 3.1: In scale-free networks (and many heavy-tailed distributions) the epidemic threshold vanishes in the thermodynamic limit.

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

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

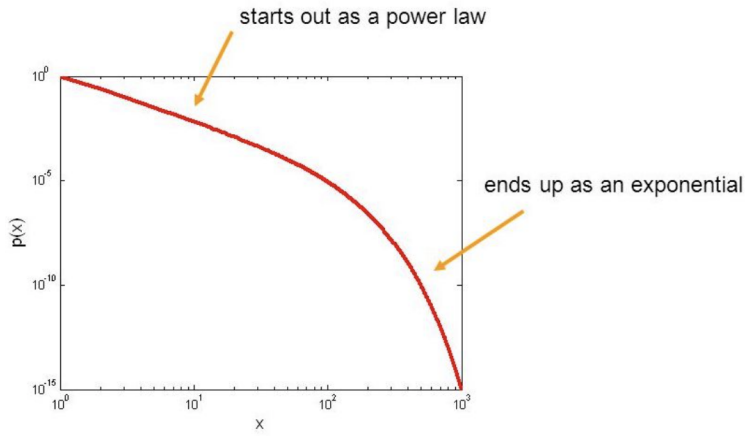


Figure 3.2: Power-law with an exponential cut-off.

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.

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).

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 continuos variable.

Let us consider $\rho_i(t)$ as the probability of a node 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**

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

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. Hence, the probability of getting infected the complementary to one probability of not getting infected by any neighbor:

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

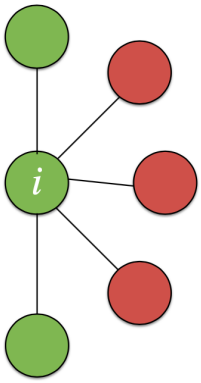


Figure 3.3: In green suscpetible nodes, while in red the infected neighbours.

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.

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:

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

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:

$$q_i^* = 1 - \prod_{j=1}^{N} \left[1 - \beta A_{ij}\varepsilon_j^*\right] \simeq \beta \sum_{j=1}^{N} A_{ij}\varepsilon_j^* \tag{3.22}$$

Substituting what we have just found in the lhs of 3.21 we obtain:

$$\mu\varepsilon_i^* = \beta(1-\varepsilon_i^*)\sum_{j=1}^{N} A_{ij}\varepsilon_j^* \tag{3.23}$$

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

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

Neglecting second order terms, we have that:

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

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:

$$\beta = \frac{\mu}{\Lambda_i} \tag{3.25}$$

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:

$$\beta_c = \frac{\mu}{\Lambda_{max}} \tag{3.26}$$

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**:
$$\beta_c^{DBMF} = \frac{\mu\langle k\rangle}{\langle k^*\rangle}$$
- for **IBMF**:
$$\beta_c^{IBMF} = \frac{\mu}{\Lambda_{max}}$$

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

$$\Lambda_{max} \sim \max\left(\sqrt{k_{\max}}, \frac{\langle k^*\rangle}{\langle k\rangle}\right) \tag{3.27}$$

And in particular:

$$\beta_c \sim \begin{cases} \mu/\sqrt{k_{\max}} & \gamma > 5/2 \\ \mu\langle k\rangle / \langle k^2\rangle & 2 < \gamma < 5/2 \end{cases} \tag{3.28}$$

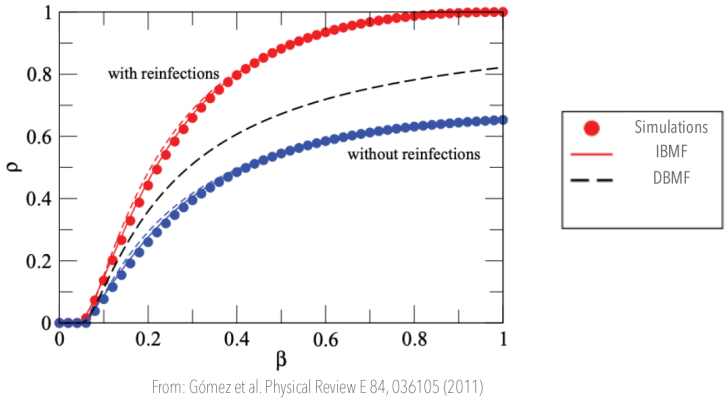
We can conclude that **IBMF** is **more accurate** than DBMF. Due to the approximation, indeed, the **DBMF** is **accurate only** in the **proximity of the epidemic threshold**, while IBMF is accurate for the entire epidemic diagram.

Lecture 9.
Thursday 29th October, 2020.
Compiled: Wednesday 11th November, 2020.

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:

$$\beta_c \xrightarrow{N\to\infty} 0 \tag{3.29}$$

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.



From: Gómez et al. Physical Review E 84, 036105 (2011)

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**, 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:

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

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. 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 3.22 of q_i:

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

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

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

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 we 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:

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

where $\mathbb{E}[X_i(t)X_j(t)]$ is the two nodes expectation probability to be infected.

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 all possible link we can have in the network. The main idea is:

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

$$= -2\mu\mathbb{E}[X_i(t)X_j(t)] + \beta\sum_k A_{jk}\mathbb{E}[X_j(t)X_k(t)] - \beta\sum_k (A_{ij}+A_{jk})\mathbb{E}[X_i(t)X_j(t)X_k(t)]$$

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 node is already infected and the susceptible term 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 we are considering the two extremes and the probability that j, 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:

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

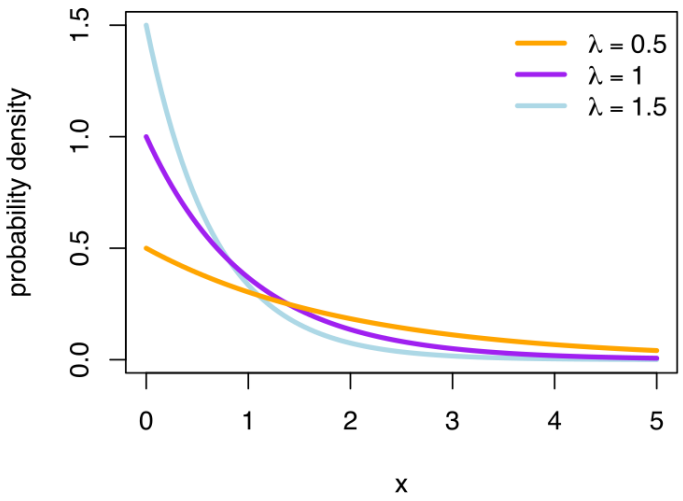
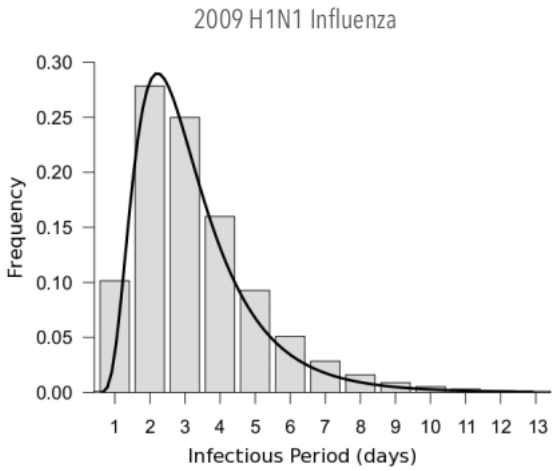


Figure 4.1: Exponential probability density function for different $\lambda = \tau$. Note as the most probable value is when we start our observation, namely at the moment in which $x = 0$.

We want now to discuss what are the implications 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.

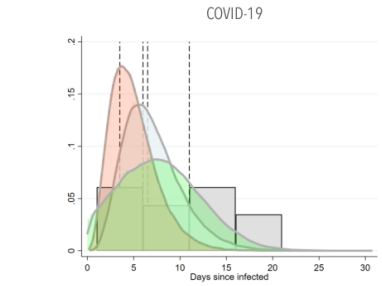
An observable easier to measure is the so called **serial interval**, namely the one that is between the onset of symptoms from an individual, to the onset of symptoms for another individual. Despite this is just an **approximation**, it still can tell us some useful information.

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**:

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

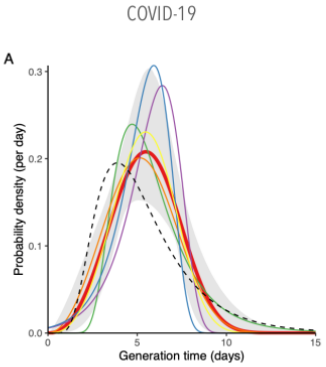
the curve in fig. 4.4 starts to resemble somehow our observations. A similar family of distributions is the so called **Erlang distributions**, where the factorial replaces the Gamma function in the denominator.

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



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)

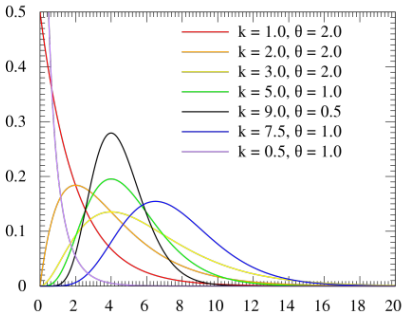
(a) Some of the inferred recovery distributions for the Covid-19 disease.



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



By Gamma_distribution_pdf.png: MarkSweep and Cburnettderivative work: Autopilot (talk) - Gamma_distribution_pdf.png, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=10734916

Figure 4.4: Gamma distribution for different changes of parameter.

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 $i = \sum_{k=1}^{K} i_k$. 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. 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 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 that $R_i(t)$ 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 i 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 neighbor 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 *do 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 "bound" for it:

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

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:

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

Part II

Poletto's Lectures

Bibliography