University of Padova

Lecture Notes
of
Life Data Epidemiology

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

Authors:
Alice Pagano
Andrea Nicolai

Tuesday 8th December, 2020

Contents

Part I

Meloni's Lectures

1

Basics Definitions and Compartmental Models

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

– Unknown author

Lecture 2. *Friday 2nd October, 2020. Compiled: Tuesday 8th December, 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

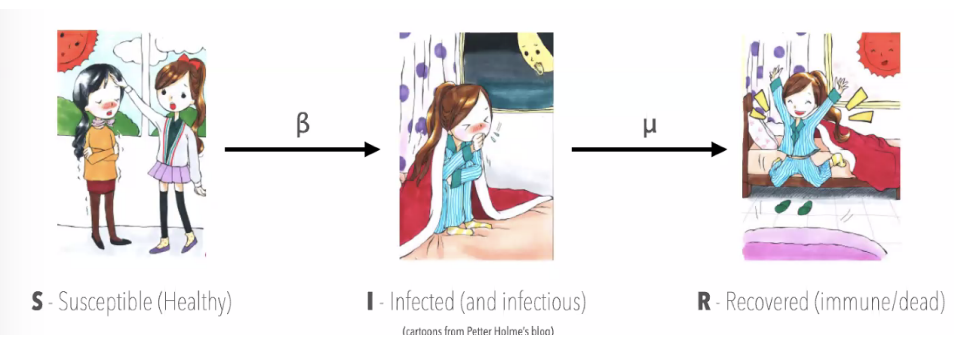


Figure 1.1: Classification of infected population in three different stages of the disease.

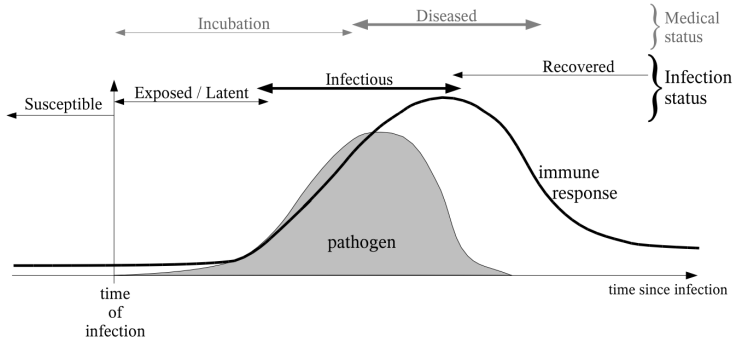


Figure 1.2: A sketch of the time-line of infection, showing the dynamics of the pathogen (grey area) and the host immune response (black line) with the labeling for the various infection classes: **S**usceptible, **E**xposed, **I**nfectious, and **R**ecovered. Note that the period when symptoms are experienced (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.

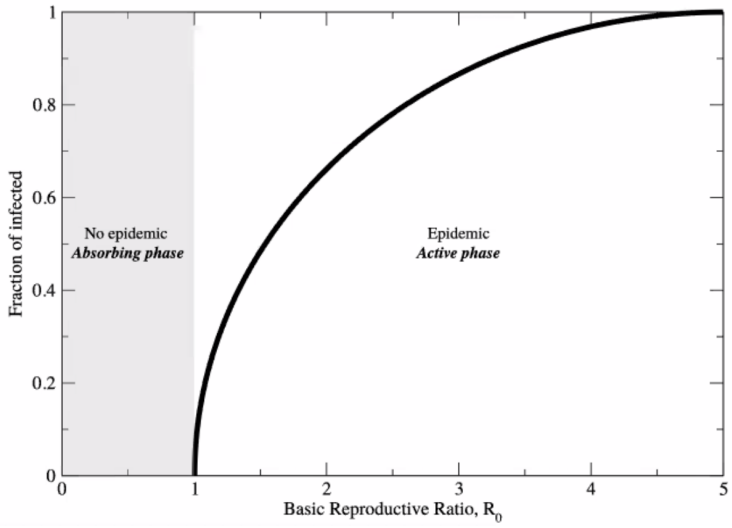


Figure 1.3: Fraction of infected vs basic Reproductive Ratio, R_0.

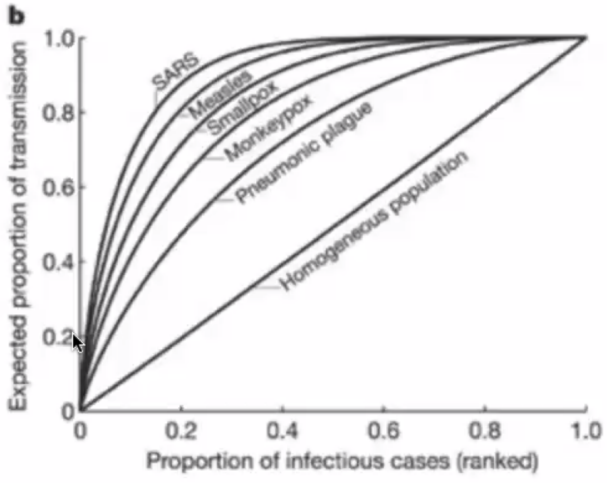


Figure 1.4: Figure from: Lloyd-Smith et al. Nature 438, 355–359 (2005).

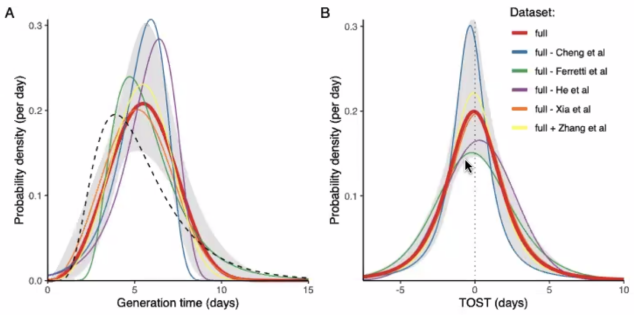


Figure 1.5: Figure from: Ferretti et al. `https://www.medrxiv.org/content/10.1101/2020.09.04.20188516v1`

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: Tuesday 8^{th} December, 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.

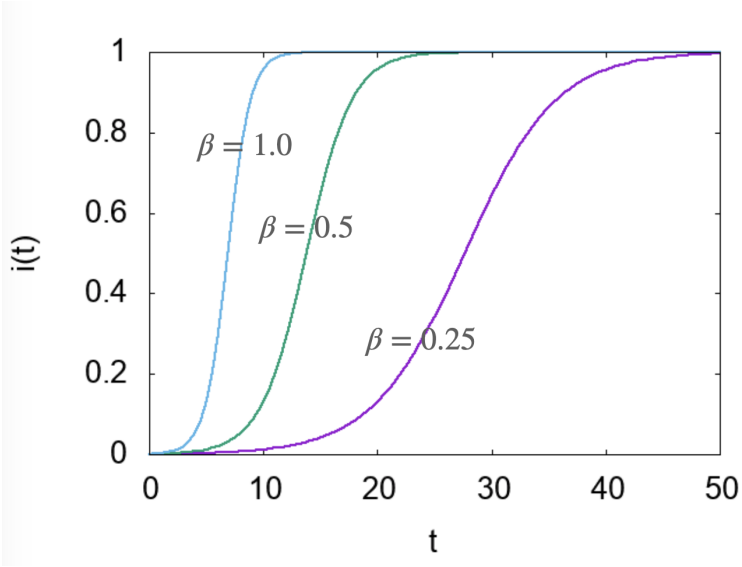


Figure 1.6: Plot of the solution of the SI model for different β.

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:

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

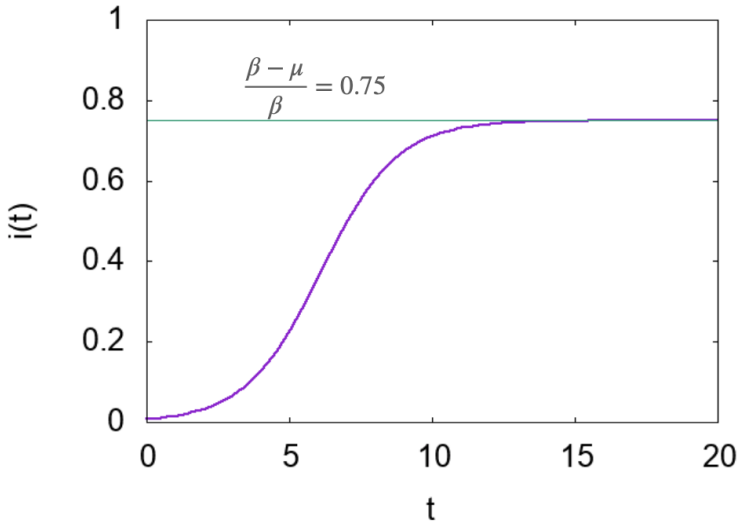


Figure 1.7: Plot of the solution of the SIS model.

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).

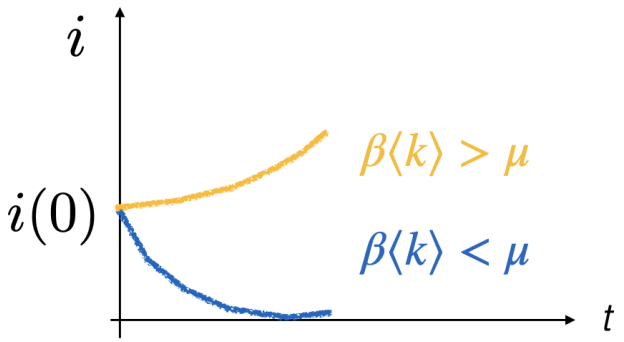


Figure 1.8: Initial transient for the SIS model.

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.

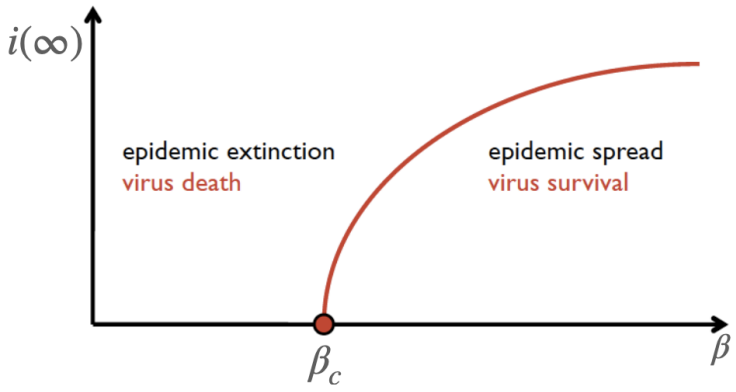


Figure 1.9: Epidemic diagram.

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:

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

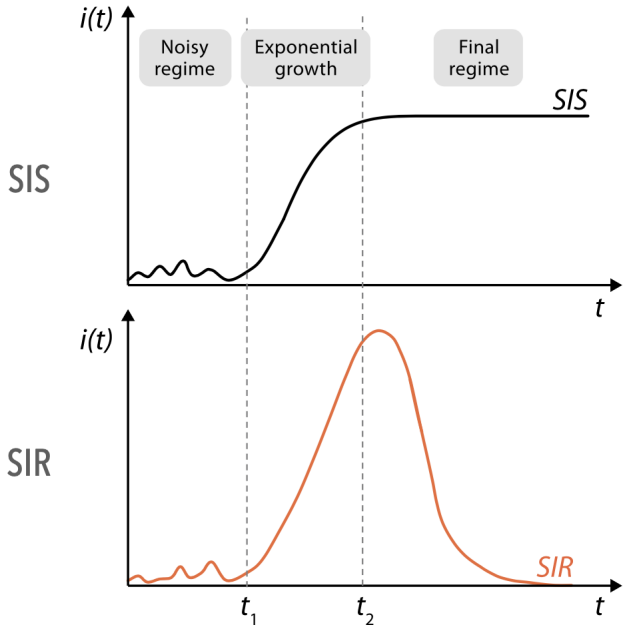


Figure 1.10: Epidemic regimes.

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: Tuesday 8th December, 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:

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

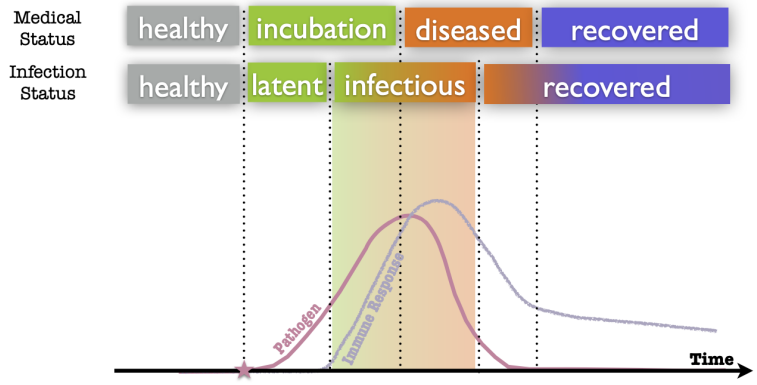


Figure 1.11: Difference between infection status and medical status.

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).

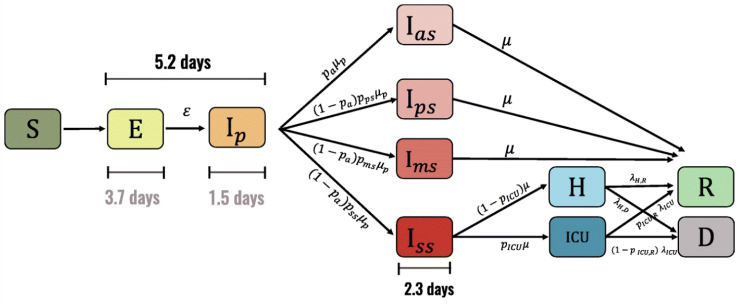


Figure 1.12: Model for Covid-19.

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: Tuesday 8th December, 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**.

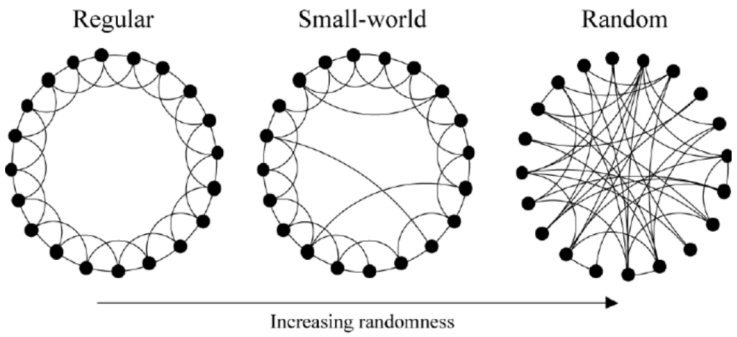


Figure 2.1: Idea of Watts and Strogatz model.

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:

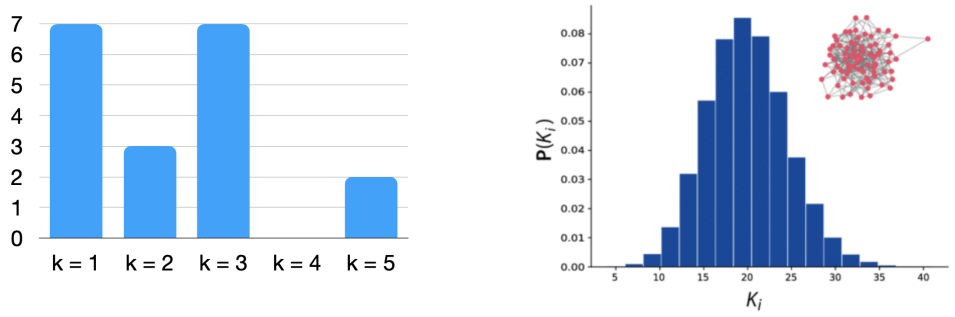


Figure 2.2: **Left:** degree distribution in a small network. **Right:** degree distribution in a network with random connections.

- 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.

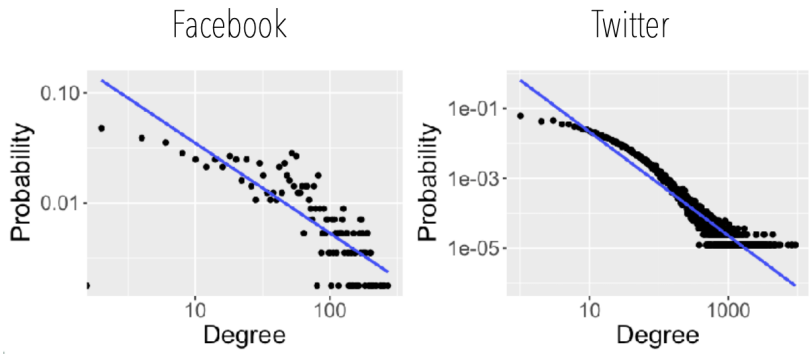


Figure 2.3: Real network of Facebook and Twitter.

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

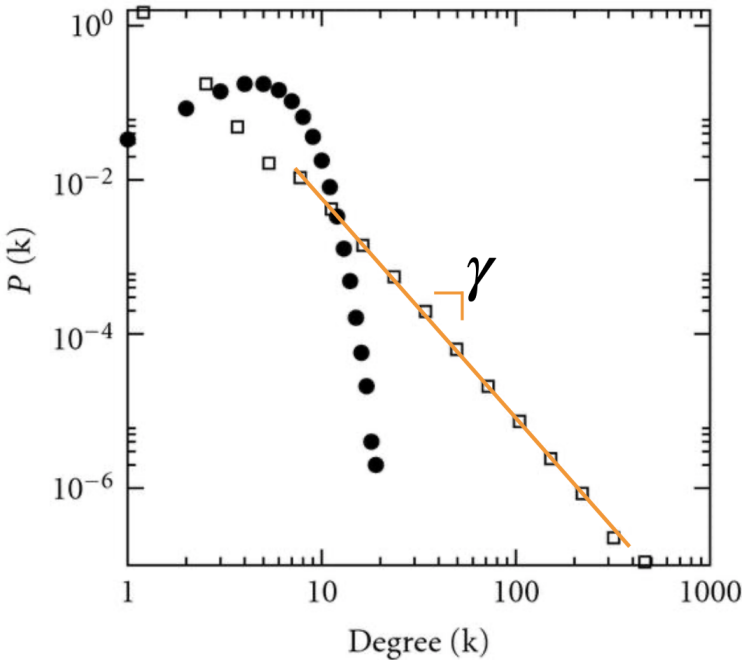


Figure 2.4: Difference between random networks and scale-free networks.

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

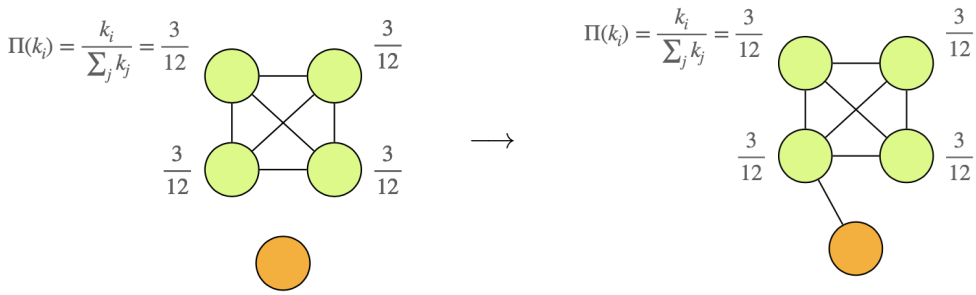


Figure 2.5: Example of Barabási-Albert algorithm.

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: Tuesday 8th December, 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.

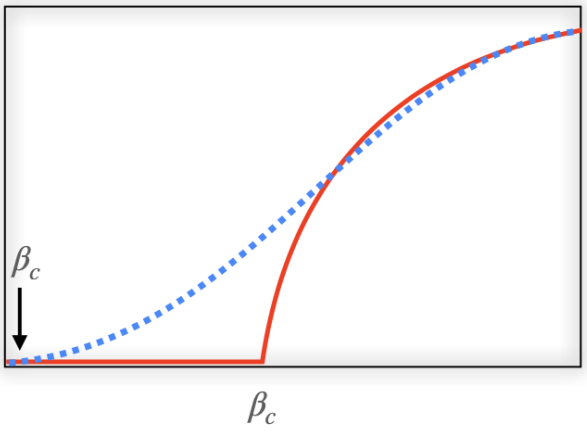


Figure 3.1: In scale-free networks (and many heavy-tailed distributions) the epidemic threshold vanishes in the thermodynamic limit.

[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.

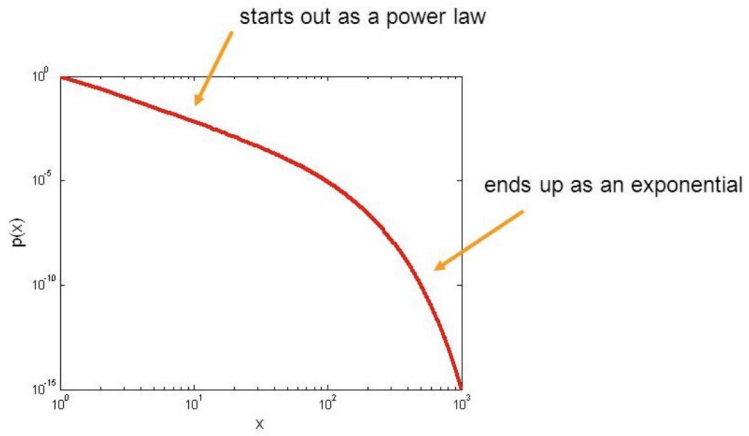


Figure 3.2: Power-law with an exponential cut-off.

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:

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

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

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

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

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

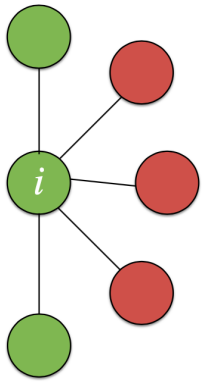


Figure 3.3: In green suscpetible nodes, while in red the infected neighbours.

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:

$$q_i(t) = 1 - \prod_{j=1}^{N} [1 - \beta A_{ij}\rho_j(t)] \tag{3.17}$$

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:

$$\mu\rho_i^* = (1 - \rho_i^*)q_i^* \quad \rightarrow \quad q_i^* = 1 - \prod_{j=1}^{N} \left[1 - \beta A_{ij}\rho_j^*\right] \tag{3.18}$$

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:

$$q_i^* = 1 - \prod_{j=1}^{N} \left[1 - \beta A_{ij}\varepsilon_j^*\right] \simeq \beta \sum_{j=1}^{N} A_{ij}\varepsilon_j^* \tag{3.19}$$

Substituting what we have just found in the lhs of 3.18 we obtain:

$$\mu\varepsilon_i^* = \beta(1-\varepsilon_i^*)\sum_{j=1}^{N} A_{ij}\varepsilon_j^* \tag{3.20}$$

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:

$$\frac{\mu}{\beta}\varepsilon_i^* = \sum_{j=1}^{N} A_{ij}\varepsilon_j^* \tag{3.21}$$

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:

$$\beta = \frac{\mu}{\Lambda_i} \tag{3.22}$$

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:

$$\beta_c = \frac{\mu}{\Lambda_{max}} \tag{3.23}$$

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

3.2.3 DBMF vs IBMF: Epidemic treshold

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

- for **DBMF**:
$$\beta_c^{DBMF} = \frac{\mu\langle k\rangle}{\langle k^2\rangle}$$
- for **IBMF**:
$$\beta_c^{IBMF} = \frac{\mu}{\Lambda_{max}}$$

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

$$\Lambda_{max} \sim \max\left(\sqrt{k_{\max}}, \frac{\langle k^2\rangle}{\langle k\rangle}\right) \tag{3.24}$$

And in particular:

$$\beta_c \sim \begin{cases} \mu/\sqrt{k_{\max}} & \gamma > 5/2 \\ \mu\langle k\rangle / \langle k^2\rangle & 2 < \gamma < 5/2 \end{cases} \tag{3.25}$$

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.

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:

$$\beta_c \xrightarrow{N\to\infty} 0 \tag{3.26}$$

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 DBMF, 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.

Lecture 9.
Thursday 29th October, 2020.
Compiled: Tuesday 8th December, 2020.

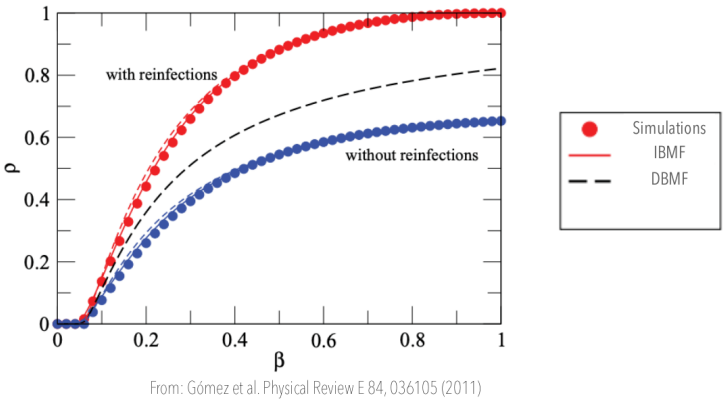


Figure 3.4: The quenched mean field (IBMF) curve follows exactly the simulation, while DBMF the is precise only around the epidemic threshold.

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. We start from:

$$\dot{\rho}_i = -\mu\rho_i + (1-\rho_i)q_i \qquad \text{where} \qquad q_i = 1 - \prod_{j=1}^{N}[1 - \beta A_{ij}\rho_j]$$

In this case, the **adjacency matrix** is replaced by an **Annealed Adjacency Matrix** (**AAM**) whose general form is:

$$\bar{A}_{ij} = \frac{k_j P(k_i|k_j)}{N P(k_i)} \tag{3.27}$$

which 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 becomes k_j. This is the number of trials we have available in order to create this specific connection i, j, over

all the possible connections that the network can return. If we substitute the explicit form of the adjiacency matrix in the expression 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 $\Theta_k = \sum_{k'} P(k'|k) \rho_{k'}$:

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

which is the formula obtained for DBMF. Hence, in the DBMF we are implicitly assuming that $\beta \rho_k \ll 1$. This is the reason why DBMF is accurate only around the epidemic threshold. At the end, we are able to pass from IBMF to DBMF and actually we are explaining the 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* which means that $\text{Prob}[\sigma_i(t) = 0]$ and $\text{Prob}[\sigma_j(t) = 0]$ are statistically independent. 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:

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

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 $\dot{\rho}$ (this involves three nodes terms and so on). 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 of being 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:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} E[X_i(t) X_j(t)] = &-2\mu \mathbb{E}[X_i(t) X_j(t)] + \beta \sum_k A_{ik} \mathbb{E}[X_j(t) X_k(t)] \\ &+ \beta \sum_k A_{jk} \mathbb{E}[X_i(t) X_k(t)] - \beta \sum_k (A_{ik} + A_{jk}) \mathbb{E}[X_i(t) X_j(t) X_k(t)] \end{aligned} \tag{3.28}$$

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:

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

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

4

Epidemic spreading on networks: advanced models

In this chapter, we are gonna study non-Markovian epidemic spreading. In the literature, it is not seen, but if you want to implement a realistic model it is very important.

4.1 Markovian Models

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.

All the models we have seen till now assume that β (infection process) and μ (recovery process) are constant rates. This means that movement between compartments takes place at constant rate, or equivalent the probability of jumping from one compartment to another does not depend on the time spent in the compartment. Essentially, we are considering a **memoryless process**. The jump are memoryless, i.e. we are running a Markov chain.

The fundamental property of a Markov chain is called **Markov property**: *the jump probability at time $t + \Delta t$ does not depend on elapsed time, but only on t*, i.e. we have not to take into account the time that we spent in that compartment. This property is very useful for mathematical treatments.

Jumps made at a constant rate (β and μ) imply that the time spent inside each compartment τ follows a **exponential distribution** (Fig. 4.1):

$$P(x) = \tau e^{-\tau x} \tag{4.1}$$

with mean $\tau = \frac{1}{\mu}$, i.e. τ is the infectious period (average time spent in I).

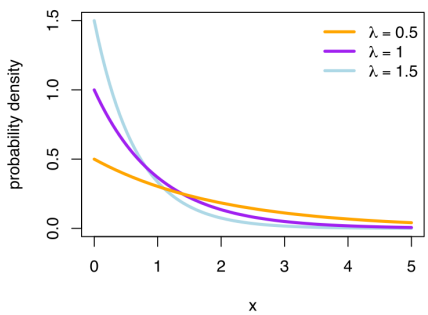


Figure 4.1: Exponential probability density function for different $\lambda = \tau$. Note as the most probable value is when we start our observation, namely at the moment in which $x = 0$.

What are the implication of this property? The most important one is that an exponentially distributed infectious period implies that the most probable duration of the disease is 0. Indeed, the probability decreases with time. More particularly, it depends on the mean, but in any case the most probable jump (time in which I am making the jump) is at the beginning. It is something that is **not realistic**, indeed if you got influenza at least you will spend some time infected. And actually, if you are looking how infectius period are distributed in real life, it is something which is quite different. For a disease, you know exactly when it starts but do not you know when it ends. For instance, let us consider the left plot in Fig. 4.2 for 2009 H1N1 Influenza. For this type of diesease the plot show the distribtuion of the infectious period, which has as probable value 2 days and an half. The most important things is that it is not 0. On the right, we have also the estimates for Covid-19. One process we use to measure the infectious period is the **serial interval**, i.e. from symptoms to symptom. Or we have also the **generation time**, i.e. from infection to infection. Obviously, these are approximations but can give you some means.

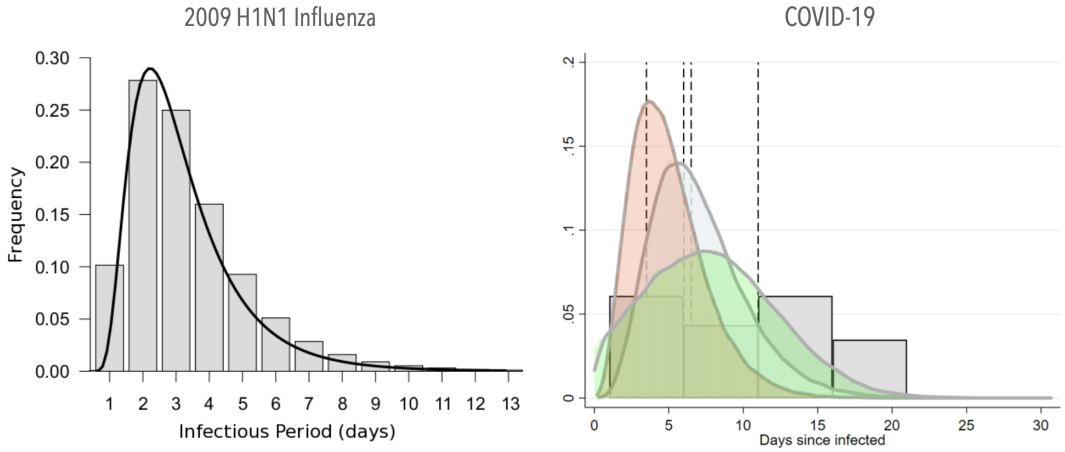


Figure 4.2: **Left:** Histogram of 1009 H1N1 Influenza. **Right:** distribution of Covid-19.

The problem is that, all these results demonstrate that these kind of diseases are *non-markovian*: **transition probability depends on the time spent in a compartment**. Indeed, patients usually spend some time infected before starting to recover.

4.2 Non-Markovian Epidemic Spreading

How are we gonna model this non markovianity? What are the distribution that better describe what we saw in the data?

Patients usually spend some time infected before starting to recover. This situation is better approximated by a **Gamma distribution**:

$$P(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \tag{4.2}$$

where k is the *shape*, θ the *scale* and $\Gamma(x)$ is the *gamma function*:

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, \mathrm{d}t$$

The gamma distribution has mean $k\theta$ and variance $k\theta^2$. This shape start to be somehow what we saw in the data.

What is similar to the gamma distribution is the **Erlang distribution**, where we have the factorial instead of the Gamma function:

$$P(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \tag{4.3}$$

or the **Weibull distribution**:

$$P(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & <\geq 0 \\ 0 & x < 0 \end{cases} \tag{4.4}$$

All of these distribution are able to reproduce real histograms.

However, the question still remains: how to include non-markovian elements in classical epidemiological models (with the assumption of markovian...)? We use a trick for the indectious period: *the sum of exponential random variables obeys a Gamma distribution.* How we are gonna incorporate that in our model? Instead of having just one transition at a coastant rate (expoential distribution), what we need to have to have a gamma distribution?

The trick is the fact that instead of having just one transition, we are gonna include more and more transitions. Instead of having just one single infectious state, individuals move from one compartment to the other, such that they spend at least some time infectious before starting the recovery. We obtain again a Markovian model.

4.2.1 SIR Model with Multiple Infectious Stages

To repeat, the solution is using **multiple infectious compartment**, i.e. SIR in well mixed populations becomes $SI_1I_2 \dots I_kR$. For instance, we are imposing that these transitions are sequential:

$$\begin{aligned} S + I &\xrightarrow{\beta} I + I_1 \\ I_1 &\xrightarrow{K\mu} I_2 \\ I_2 &\xrightarrow{K\mu} I_3 \\ &\vdots \\ I_k &\xrightarrow{K\mu} R \end{aligned} \tag{4.5}$$

Hence, if I want to get recovered I need to spent some times infectious, but the model is still markovian! More precisely, the equations are:

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= -\beta si \\ \frac{\mathrm{d}i_1}{\mathrm{d}t} &= \beta si - K\mu i_1 \\ \frac{\mathrm{d}i_2}{\mathrm{d}t} &= K\mu i_1 - K\mu i_2 \\ &\vdots \\ \frac{\mathrm{d}r}{\mathrm{d}t} &= K\mu i_k \end{aligned} \tag{4.6}$$

where the rate of each I transition is $K\mu$ and with $i = \sum_{k=1}^{K} i_k$. We got that this is the infectious period distribution:

$$P(\tau) = \frac{(\mu K)^K}{\Gamma(K)} \tau^{K-1} e^{-\mu K \tau} \tag{4.7}$$

where the mean is still $1/\mu$, but the shape is totally different. which is the gamma function. We have two special cases:

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

Other quantities are:

$$R_0 = \frac{\beta}{\mu}$$

and the **final epidemic size**:

$$r_\infty = 1 - e^{-R_0 r_\infty}$$

Moreover, we have an early growth:

$$i(t) \simeq i_0 e^{\lambda t}$$

instead of $i(t) \simeq i_0 e^{(\beta-\mu)t}$. Hence, the disease is growing faster and has a shorter duration.

With λ as the solution of:

$$R_0 = \frac{\lambda}{\mu\left(1 - \left(\frac{\lambda}{K\mu} + 1\right)^{-K}\right)}$$

The $SI_1I_2 \dots I_kR$ model has several limitaionts:

- it is defined only for well-mixed populations;
- focus only on the infectious period distribution. We have that infections are still Markovian and this model only reproduces as Gamma distribution.

4.2.2 Generalized SIS Model

Now, we are going to present something which is more general where we can include non-markovian both in recovery and infections. Is it possible to write down a general model on networks? The answer is yes, it is a bit more complicated and we still needs some kind of approximation at some point. In particular, we need a mean-field approximation.

We have to change our point of view. We are gonna use a slightly different approach, i.e. instead of probabilities we are gonna talk about events. The idea is that we are gonna modelling in this case the infections and recoveries with two random numbers which we extract from distribution and are as general as possible.

The ingredients are:

- a random number $R_i(t)$: recovery time of node i when infected;
- a random number $M_{ij}(t)$: infection times at which node i tries to infect node j.

In order:

1. node i get the infection at time t;
2. we extract the random number $R_i(t)$ which represents the time in which node i is gonna recovery (or, the time for which it stay infected);

3. then, we extract the random number $M_{ij}(t)$ which reprsenets the number of trials that i try to infect node j while infected;

4. we generate a sequence of times

$$T_{ij}^{(1)} \leq \cdots \leq T_{ij}^{(M_{ij}(t))} \leq R_i(t)$$

in which node i try to infect node j. For instance, $T_{ij}^{(1)}$ is the first time that node i try to infect node j then we have the second time and so on;

5. I am gonna repeat the last step for all my neighbours.

Hence, the trasmissibility of the disease is seen as how many trials I am gonna make to infect. One important thins is that $R_i(t)$ and $M_{ij}(t)$ can be drawn from any distribution and not only from the exponential one. How do we extract the T_{ij} is not important at this point, because we are only gonna focus on the distribution of $R_i(t)$ and $M_{ij}(t)$.

Now, let us make some assumptions to make the model more reasonable and then treat it analytically. We assume that:

- $R_i(t)$ and $M_{ij}(t)$ do not depend on time, i.e. $R_i(t) \equiv R_i$ and $M_{ij}(t) \equiv M_{ij}$;
- $R_i(t)$ and $M_{ij}(t)$ do not depend on i and j, i.e. same distribution for all the nodes $R_i \equiv R$ and $M_{ij} \equiv M$.

hence, we are assuming that these numbers should not depend on time and for instance they are typical for the disease. It is valid both for the recovery and for the infections. However, if we consider restrictions as lockdown and so on these number should change, but for the let us consider the simplest model without such restrictions. Indeed, with these assumptions we are reducing the complexity.

We call:

- $\mathbb{E}[R]$, the expected value of R;
- $\mathbb{E}[M]$, the expected value of M;
- v_i, the probability that node i is infected in the steady state.

Now, let us build the model:

1. let us suppose that we are in the steady state of the system (all the transient are passed). In a large time interval $[0, S]$ the number of times node j has been infected is proportional to S (it is linear). Since the length of each infected period is $E[R]$, the **number of infected periods** experienced by a node j (number of times a node has been infected in the interval $[0, S]$) can be written as:

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

2. during each infected period, node j will try to infect i an average $E[M]$ number of times. So, the **total number of infection attempts** from node j to i in a large period of time are:

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

the number of times j has been infectious multiplied by the number of infection attempts per each time;

3. then, we make the **mean-field assumption**: the conjunct probability that j is infected and i is susceptible is:

$$\text{Prob}[\sigma_i(t)=0,\sigma_j(t)=1] \rightarrow \underbrace{\text{Prob}[\sigma_i(t)=0]}_{(1-v_i)}\underbrace{\text{Prob}[\sigma_j(t)=1]}_{v_j}$$

4. we can write the **total number of successful infection attempts** for j to i as:

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

that is the total number of attempts multiplied by the probability that i was not infected.

Summing over all the neighbors of i we have the the **total number of successful infections** i will receive during interval $[0,S]$ is:

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

5. in the *steady state* the number of successful infections ($S\sum_{j=1}^{N}a_{ij}\frac{\mathbb{E}[M]}{\mathbb{E}[R]}v_j(1-v_i)$) is asymptotically equal to the number of infected periods experienced by i: $v_iS/E[R]$. Thus we have:

$$\cancel{S}\sum_{j=1}^{N}a_{ij}\frac{\mathbb{E}[M]}{\cancel{\mathbb{E}[R]}}v_j(1-v_i)=v_i\frac{\cancel{S}}{\cancel{\mathbb{E}[R]}} \qquad \rightarrow \qquad v_i=\mathbb{E}[M](1-v_i)\sum_{j=1}^{N}a_{ij}v_j \quad (4.8)$$

hence the probability of i being infected depends by the sum over all its neighbours, times the term $E[M]$ which is the average infection attempts that it is gonna experience during the evolution.

Do the last expression sound familiar? This is exactly the same expression of the IBMF but with a generic infection term $E[M]$:

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

The implications are:

- the definition of M implicitly includes the recovery term: $T_{ij}^{(1)}\leq\cdots\leq T_{ij}^{(M_{ij}(t))}\leq R_i(t)$;
- exponential case: $E[M]$ is the expected number of infection events in a Poisson preocess with intensity β within an exponential recovery time with expectation $1/\mu$. Thus: $E[M]=\beta/\mu$.
- epidemic treshold (lower bound):

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

Hence, this is the form in which we can generate a generic model from.

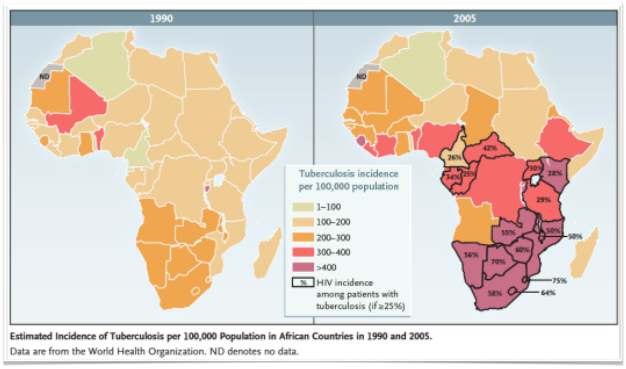
4.3 Interacting diseases

Lecture 11. *Thursday 5^{th} November, 2020. Compiled: Tuesday 8^{th} December, 2020.*

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

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

Let us now consider the simplest case, where we have only two different diseases that **cooperate**. That is to say that one diseases **boost** the spreading of the other one. The map fig.4.3 represents the spreading of Tuberculosis (TB) in 1990. The latter is a disease which has a very long latent period: it can stay in a latent state actually for many many years or, sometimes, we may have contracted it without it never showing up. Let us compare the same the same map some years later, in 2005. One should note that the disease has exploded, especially in southern regions: numbers almost have doubled. The reason is that, during that time window, HIV reached the African continent. HIV is a disease which compromises our immune system and makes it less efficient: the probability of getting any other disease is higher than the normal.



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

Figure 4.3: Map of the prevalence of TB for different years (1990 and 2005).

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

The downside, when **competition** between diseases may arise, is shown as an example by seasonal influenza. From fig. 4.4 we can see the distribution of the different influenza viruses that are out during a single season. Individuals can be infected by different strains of influenza, therefore the distribution of the prevalence of total infected people is the sum of the values regarding the three strains.

It can be shown that our immune system has a sort of **memory** which provides a **long lasting** immunity. The main consequence is that if we accidentally get a disease, there is a possibility that we end up not contracting it once again for some time in the future. Moreover, it can provide also a sort of **cross-immunity** for some other diseases that are similar to the one we contracted, even though our immune systems had never faced them. As one can imagine, this implies that the actual

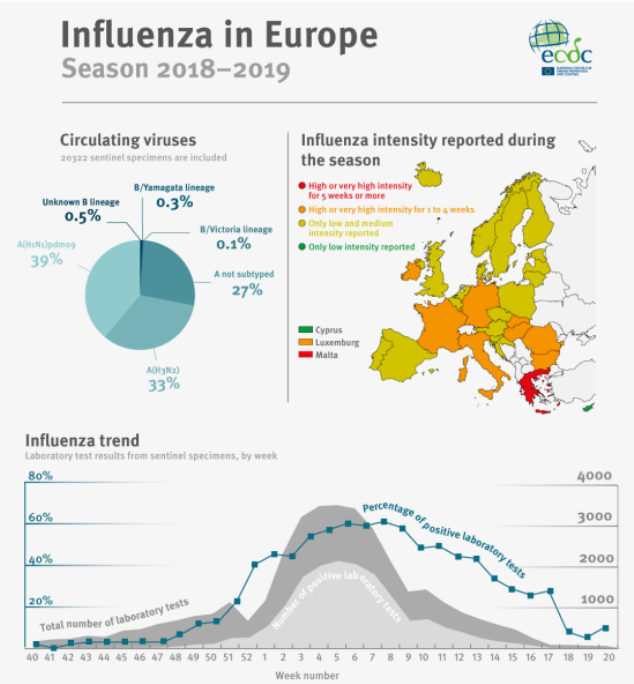


Figure 4.4: Distribution of different strains of influenza for a single season.

susceptible population is just a reduced fraction of the whole one, and should explain why we do not usually observe influenza pandemics.

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

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

For instance, the simplest case one can think of is to couple two different diseases whose dynamics are respectively **SIS** and **SIS**. We end up having twice the number of states (see fig 4.5), since we take into account all the possible combinations between compartments. A single node i at time t can be be either one of the following states $\{SS, SI, IS, II, \}$ with its own probability density $\{[\rho^{SS}]^i_t, [\rho^{SI}]^i_t, [\rho^{IS}]^i_t, [\rho^{II}]^i_t\}$. Therefore, every disease will be defined thanks to its own **parameters** $\{\beta_1, \beta_2, \mu_1, \mu_2\}$. However, some more are to be introduced, namely the ones that **encode** the **interaction** between diseases.

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

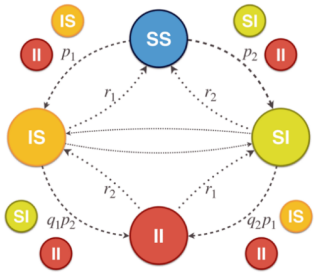


Figure 4.5: Coupled SIS model and all possible combinations and different transition probabilities between each compartment.

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

At this point it comes to take into account, and therefore model, the interaction effects between the two diseases (see fig. 4.6). These interactions may result in three effects:

- **modified susceptibility** $\lambda_\mathbf{a}$: being infected of one diseases makes an individual more ($\lambda_a > 1$) or less ($\lambda_a < 1$) probable to be infected by a second one. It is the case for the HIV that increases the probability of contracting TB, or for a strain of a seasonal influenza that inhibits individuals to get influenza of another kind
- **modified infectivity** $\lambda_\mathbf{b}$ once we get infected by one disease, we are less infectious wrt second one
- **modified infectious period** η if we are infected of one disease, it can favor/hinder the recovery from the other disease. As for the second case, it may happen when our immune system is compromised.

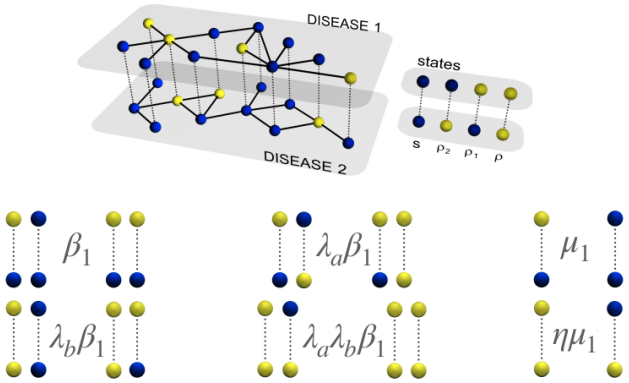


Figure 4.6: Coupled SIS networks and their effects on each other.

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

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

$$\dot{SS}(k,l) = -(k\sigma_1 + l\sigma_2)SS(k,l) + \mu_1 IS(k,l) + \mu_2 SI(k,l) \tag{4.9}$$

$$\dot{IS}(k,l) = k\sigma_1 SS(k,l) - l\lambda_a\sigma_2 IS(k,l) - \mu_1 IS(k,l) + \eta\mu_2 II(k,l) \tag{4.10}$$

$$\dot{SI}(k,l) = l\sigma_2 SS(k,l) - k\lambda_a\sigma_1 SI(k,l) - \mu_2 SI(k,l) + \eta\mu_1 II(k,l) \tag{4.11}$$

$$\dot{II}(k,l) = k\lambda_a\sigma_1 SI(k,l) + l\lambda_a\sigma_2 IS(k,l) - (\eta\mu_1 + \eta\mu_2)II(k,l) \tag{4.12}$$

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

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

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

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

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

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

While for network 2:

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

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

In order to solve these equations, firstly we need to assume that we are in the **steady state**. Later we are able to write down a couple of self-consistent equations for σ_1 and σ_2, and then solve them by finding the intersection. Finally, the **epidemic threshold**:

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

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

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

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

For instance, let us consider the 3D epidemic diagram shown for **cooperating diseases** (see fig. 4.7) with $\lambda > 1$ and $\eta < 1$.

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

The opposite actually occurs when we consider **competing diseases** (see fig. 4.8), that is the case $\lambda < 1$ and $\eta > 1$. For either one of them it is more difficult to

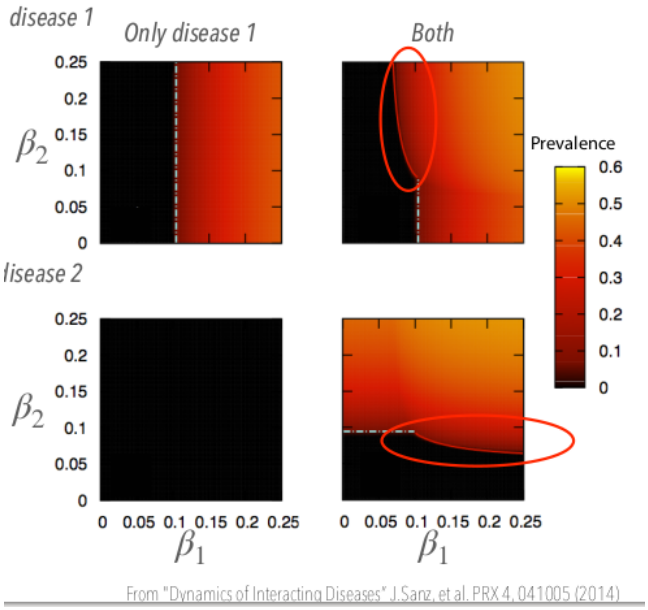


Figure 4.7: Cooperating diseases, $\lambda > 1$ and $\eta < 1$

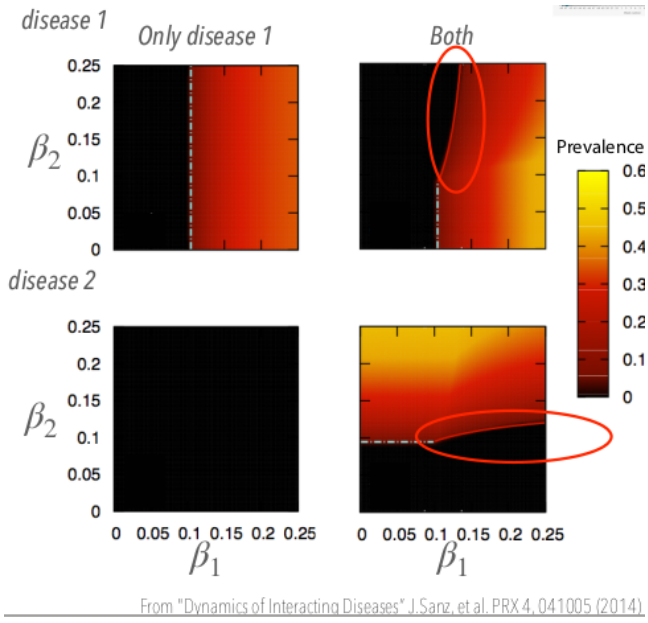


Figure 4.8: Competing diseases, $\lambda < 1$ and $\eta > 1$

spread once we have overcome the critical threshold for the other's. This concludes the discussion when taking for heterogeneous mean field degree assumption.

One may wonder now how **quenched mean-field** equation look like (individual based formulation). For the sake of simplicity we are going to discuss only a single network, and we will take into account only its main effect, that is the one on the **modified susceptibility**. Also under this assumptions equations $\{[\rho^{SS}]_i^{t+1}, [\rho^{SI}]_i^{t+1}, [\rho^{IS}]_i^{t+1}, [\rho^{II}]_i^{t+1}\}$ can be written, whose structure is exactly the same one as before for a single disease case. Each term contributes and plays a role in the probabilities $[\rho^{IS}]_i^{t+1}, [\rho^{SI}]_i^{t+1}$. Since these expressions are really long and complex, we are going to skip this part. Nonetheless we will briefly discuss an interesting **assumption** we make during our computations: we did insert the functions f_{SI}, f_{IS}. It was done to consider the fact that we cannot contract two diseases at the same time, it would be unrealistic. However, if during our simulations it happens, we pick only one disease at random. The

function f_{IS} is the following, for the $[\rho^{IS}]_i^{t+1}$ expression:

$$f_{IS} = \frac{q_{IS}(1 - 0.5q_{SI})}{q_{IS}(1 - 0.5q_{SI}) + q_{SI}(1 - 0.5q_{IS})} \tag{4.13}$$

where:

$$q_{IS} = 1 - \prod_j^N \left[1 - A_{ij}\beta_1 \left([\rho^{IS}]_j^{t+1} + [\rho^{II}]_j^{t+1}\right)\right] \tag{4.14}$$

and:

$$q_{SI} = 1 - \prod_j^N \left[1 - A_{ij}\beta_2 \left([\rho^{SI}]_j^{t+1} + [\rho^{II}]_j^{t+1}\right)\right] \tag{4.15}$$

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

$$\rho = \frac{1}{N}\sum_{i=1}^{N}(\rho_i^{IS} + \rho_i^{SI} + \rho_i^{II}) \tag{4.16}$$

If the probability of getting the disease is $\lambda = 2$ one can see in fig. 4.9 that the curve becomes steeper and, as λ increases more this behavior accentuates even more and looks like exploding. Moreover one should note that, despite the infectivity β is exactly the same, the prevalence ρ shows some discontinuities that becomes larger as more the two diseases cooperate more and more.

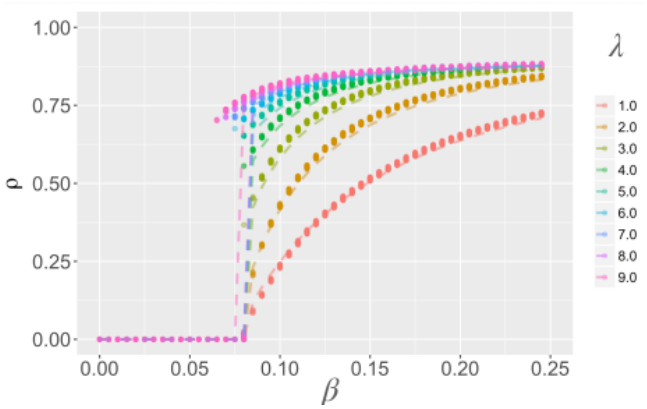


Figure 4.9: Cooperating diseases, numerical simulations results for different λ.

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

$$\rho = \frac{1}{N}\sum_{i}^{N}(\rho_i^{SI} + \rho_i^{IS}) \tag{4.17}$$

As expected, the prevalence shown in left figure 4.10 does not change at all and is the same as before. While, plotting the difference $|\rho^{IS} - \rho^{SI}|$ we observe some oscillations in values slightly after the critical threshold. This is explained by saying that only a disease can survive, and which is given by chance, being the two symmetrical (see right fig 4.10). As β increases, we see as the difference approaches zero: both of them here survive each with the same prevalence. That is to say that a half of

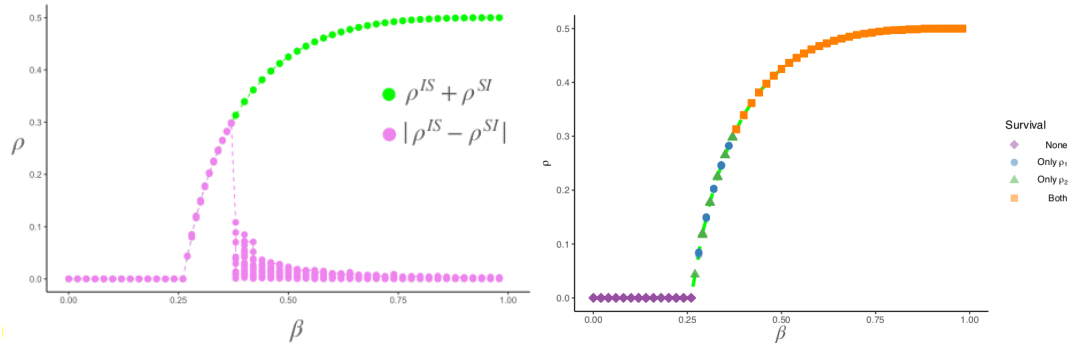


Figure 4.10: **Left:** Competing diseases, numerical simulations results. **Right:** Competing diseases, numerical simulations results with particular attentions to the **survival disease**. In the intermediate regime, the survivor is chosen by chance.

infected population has contracted a disease, while the other half the second one. For large β we see as the two diseases coexist.

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

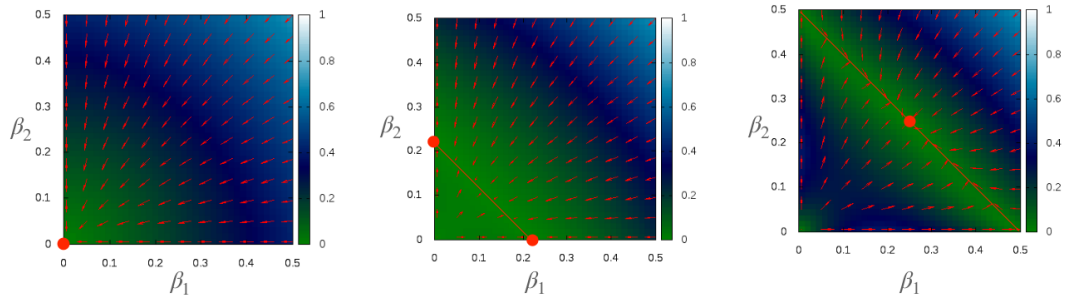


Figure 4.11: Competing diseases, manyfold that describes the dynamics of the prevalence according to different values of β_1 and β_2.

5

Spreading in social systems

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

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

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

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

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

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

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

5.1 Complex contagion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

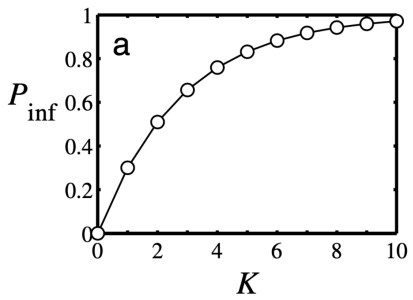


Figure 5.1: Generalized Complex Contagion model: epidemic spreading.

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

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

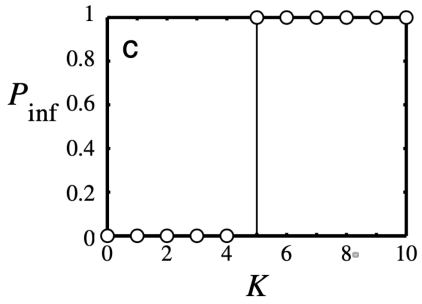


Figure 5.2: Generalized Complex Contagion model: deterministic threshold model.

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

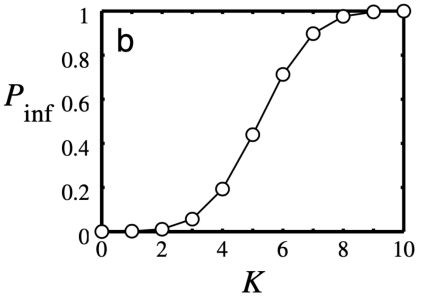


Figure 5.3: Generalized Complex Contagion model: stochastic threshold model.

5.2 Applications to Online Social Networks

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

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

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

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

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

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

And the same for users activity:

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

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

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

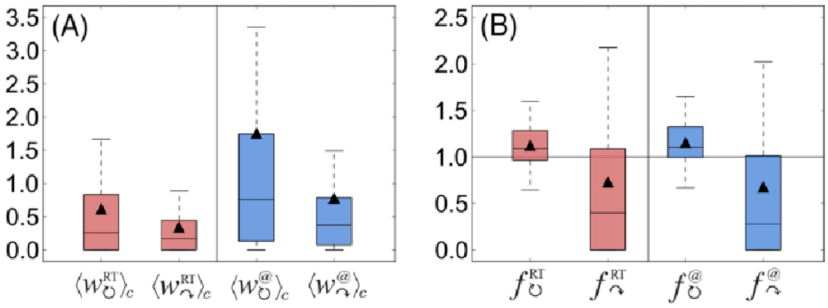


Figure 5.4: Spreading inside a community is favored (effects of homophily and reinforcement are noticeable).

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

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

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

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

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

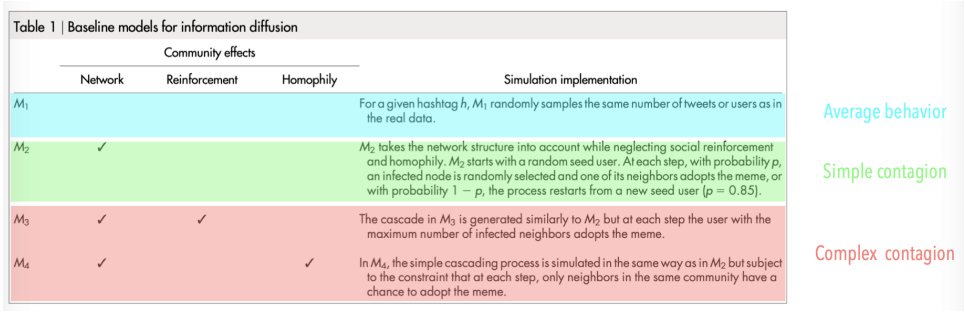


Table 1 | Baseline models for information diffusion

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

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

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

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

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

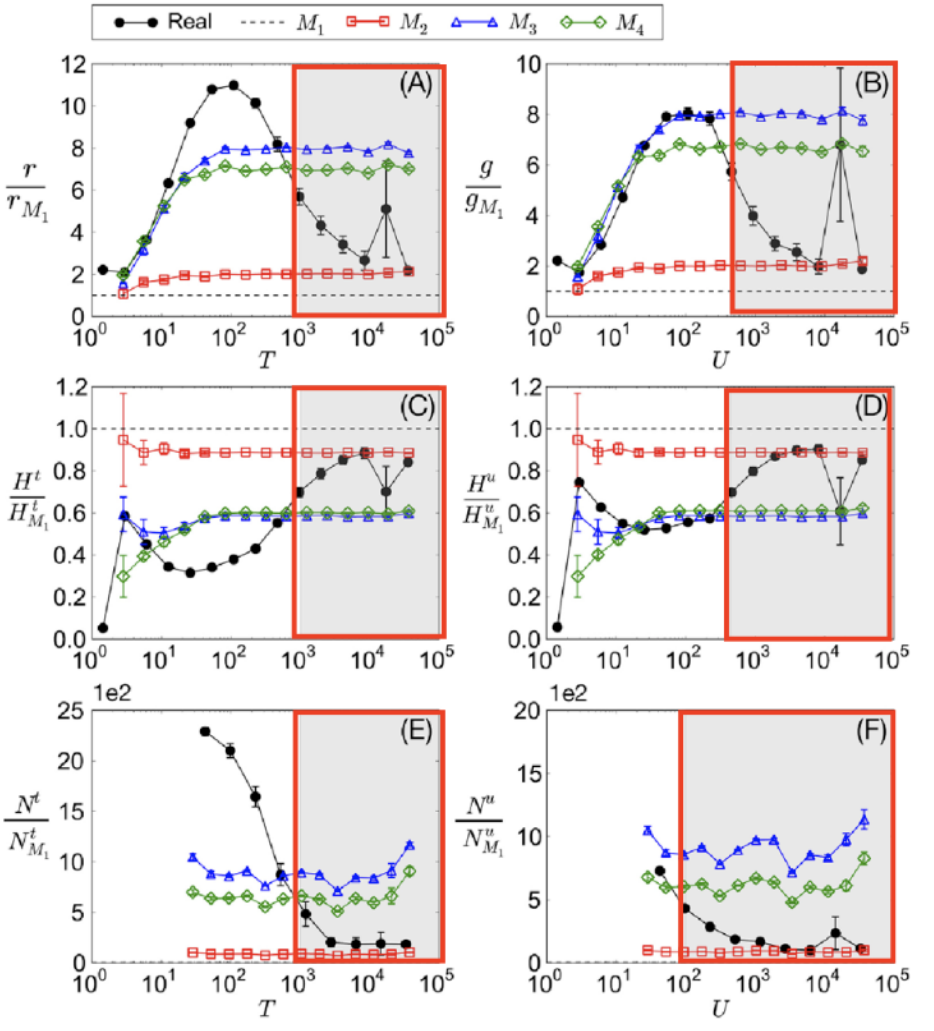


Figure 5.6: Results of information spreading in online social networks.

Part II

Poletto's Lectures

Bibliography