

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 8<sup>th</sup> December, 2020



# Contents

<b>I Meloni's Lectures</b>	<b>1</b>
<b>1 Basics Definitions and Compartmental Models</b>	<b>3</b>
1.1 Compartmental models . . . . .	3
1.2 Basic models . . . . .	6
1.2.1 SI model . . . . .	6
1.2.2 SIS model . . . . .	7
1.2.3 SIR model . . . . .	10
1.3 Extensions of the SIR model . . . . .	12
1.3.1 SIR with Demography . . . . .	12
1.3.2 SIRS Model . . . . .	13
1.3.3 SEIR Model . . . . .	14
1.4 Summary of compartmental models in well-mixed populations . . . . .	16
<b>2 Network Science - Basics</b>	<b>17</b>
2.1 Main definitions . . . . .	17
2.2 Degree distribution over networks . . . . .	19
2.2.1 Erdős and Rényi Model: random graphs . . . . .	19
2.2.2 Scale-free networks . . . . .	21
2.2.3 Barabási-Albert Model . . . . .	23
<b>3 Epidemic Spreading on Networks</b>	<b>25</b>
3.1 SIS model in a network . . . . .	25
3.1.1 Homogeneous Networks . . . . .	26
3.1.2 Heterogeneous Networks . . . . .	28
3.2 SIR model in a network . . . . .	31
3.2.1 Degree-based mean-field theories (DBMF) . . . . .	31
3.2.2 Individual-based mean-field theories (IBMF) . . . . .	32
3.2.3 DBMF vs IBMF: Epidemic threshold . . . . .	34
3.2.4 IBMF and Pair approximation . . . . .	36
<b>4 Epidemic spreading on networks: advanced models</b>	<b>39</b>
4.1 Markovian Models . . . . .	39
4.2 Non-Markovian Epidemic Spreading . . . . .	40
4.2.1 SIR Model with Multiple Infectious Stages . . . . .	41
4.2.2 Generalized SIS Model . . . . .	42
4.3 Interacting diseases . . . . .	45
<b>5 Spreading in social systems</b>	<b>53</b>
5.1 Complex contagion . . . . .	54
5.2 Applications to Online Social Networks . . . . .	56

<b>II Poletto's Lectures</b>	<b>61</b>
<b>6 Introduction to metapopulation models</b>	<b>63</b>
6.1 Spatial spread of epidemics . . . . .	63
6.1.1 Human mobility . . . . .	63
6.1.2 Modelling Human Mobility . . . . .	66
6.2 Integrating Human Mobility in Epidemic Models . . . . .	67
6.2.1 SIR metapopulation model . . . . .	67
6.3 Application of metapopulation models . . . . .	69
6.3.1 Spatial propagation dynamics . . . . .	69
6.3.2 SIR metapopulation model with memory . . . . .	73
6.4 Global Invasion Threshold . . . . .	76
6.4.1 Homogeneous networks . . . . .	76
6.4.2 Heterogeneous networks . . . . .	78
6.5 Spatial spread of competing diseases . . . . .	80
<b>7 Temporal Networks</b>	<b>85</b>
7.1 Temporal networks dynamics . . . . .	88
7.1.1 Activity driven model . . . . .	89
7.1.2 Randomised Reference Models . . . . .	91

# Part I

## Meloni's Lectures



# 1

## Basics Definitions and Compartmental Models

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

– Unknown author

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  $\beta$  and  $\mu$ .

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

**Lecture 2.**  
Friday 2<sup>nd</sup>  
October, 2020.  
Compiled: Tuesday  
8<sup>th</sup> December,  
2020.

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  $\beta$  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 (**Susceptible**), the individual can contract the disease (**Exposed**) and then, only after some time, becomes infectious (**Infectious**) until he recovers (**Recovered**) (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.

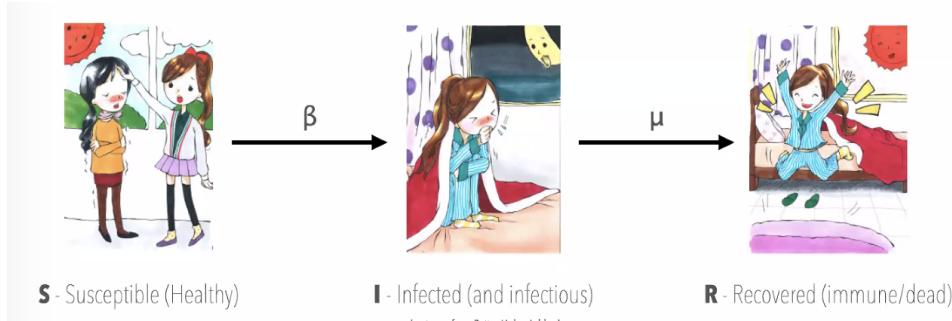
Now, let us introduce the **Basic Reproductive Number  $R_0$**  (pr. “*R naught*”) which is a measure of the infection in the population. If we wanted to empirically determine it: we put one guy inside a group for an arbitrarily long time period and, at the end, we count the number of secondary cases 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:

$$\begin{cases} R_0 < 1 \\ R_0 = 1 \\ R_0 > 1 \end{cases} \quad (1.1)$$

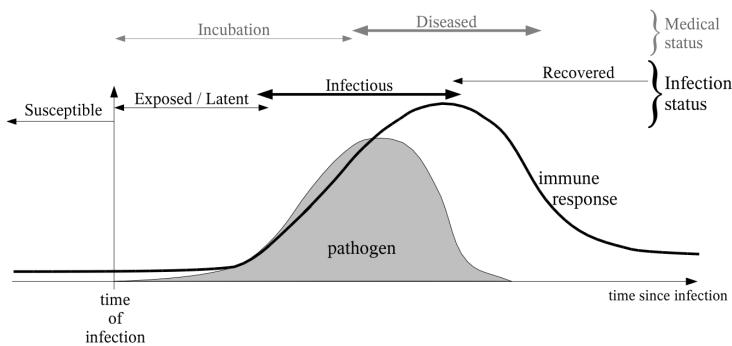
Let us consider the plot of Fig. 1.3, we have a sort of **second order phase transition** at the point  $R_0 = 1$ . Note that  $R_0$  for the SARS is higher than the one of COVID-19. However, we did not experienced an outbreak of this disease, so that is not the only parameters to be taken into account 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

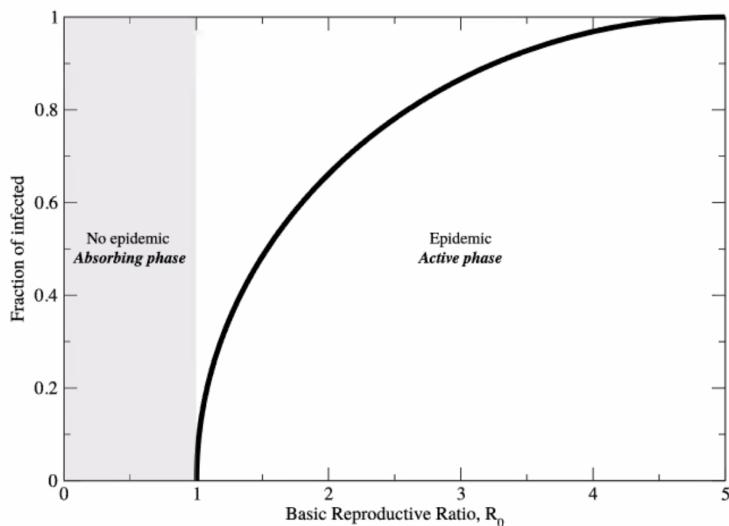


**Figure 1.1:** Classification of infected population in three different stages of the disease.

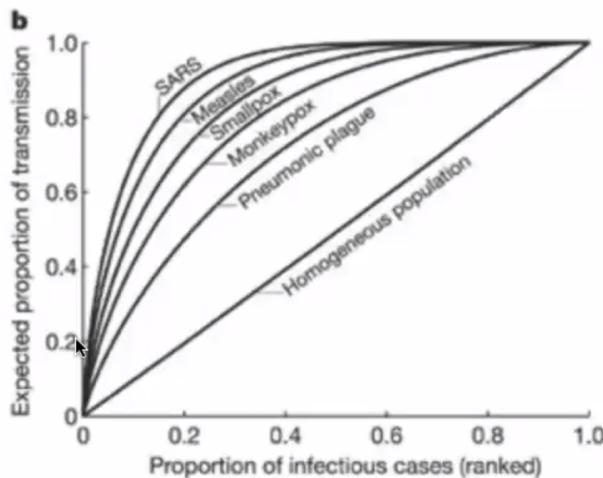


**Figure 1.2:** A sketch of the time-line of infection, showing the dynamics of the pathogen (grey area) and the host immune response (black line) with the labeling for the various infection classes: **Susceptible**, **Exposed**, **Infectious**, and **Recovered**. Note that the period when symptoms are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

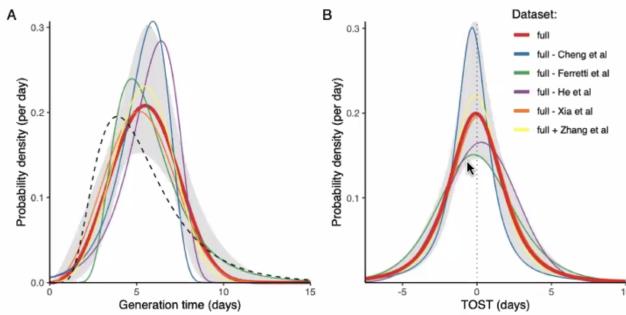
same of  $R_0$  but varying wrt time. Hence, it is the average number of secondary cases that a single case produces in a population at time  $t$ .



**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  $\alpha$  depends on the model ( $\alpha$ : 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

**Lecture 3.**  
Thursday 8<sup>th</sup>  
October, 2020.  
Compiled: Tuesday  
8<sup>th</sup> December,  
2020.

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).

## 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** (**Susceptible Infected**). 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:



where  $\beta$  is the “*per contact*” *infection rate* and dictates the speed of the spreading. We can write down the **equation** that can be solved exactly:

$$\begin{aligned} \frac{ds}{dt} &= -\beta \langle k \rangle si \\ \frac{di}{dt} &= \beta \langle k \rangle si \end{aligned} \quad (1.3)$$

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  $\beta$ . The product  $si$  is the probability of having a contact between an infected and a susceptible, and  $\beta si$  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:

$$\frac{di}{dt} = \beta i(1 - i) \rightarrow \frac{1}{\beta i(1 - i)} di = dt \rightarrow \frac{1}{\beta(1 - i)} di + \frac{1}{\beta i} di = dt$$

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:

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

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 <sup>1</sup>, 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  $\beta$  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.

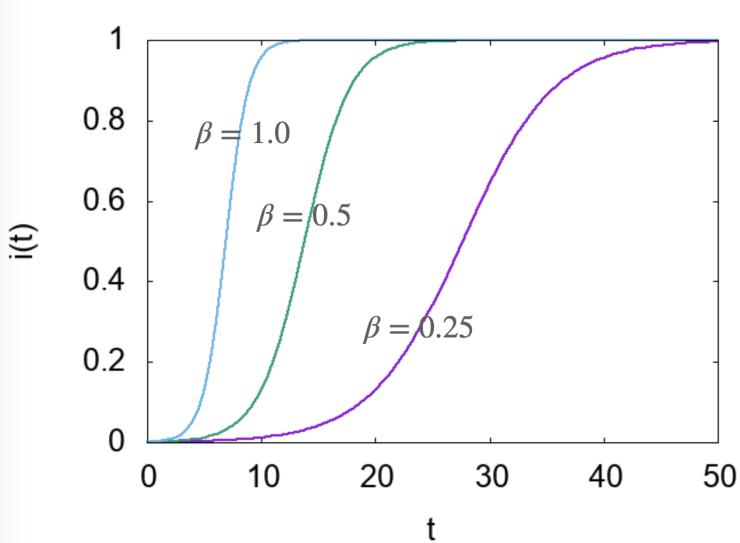
### 1.2.2 SIS model

Now, let us introduce a slightly more complicated model, that is the **SIS** model, where compartments are **Susceptible**, **Infected**, **Susceptible**. Transitions now are two:




---

<sup>1</sup>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  $\beta$ .

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  $\mu$ .

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  $\mu$  is the **recovery rate** which determines the *time-scale of the infection*. Dividing  $\beta$  by  $\mu$  you can **rescale** all the **dynamics**. The **equations** are exactly the same as before, except for a term:

$$\begin{aligned} \frac{ds}{dt} &= -\beta \langle k \rangle si + \mu i \\ \frac{di}{dt} &= \beta \langle k \rangle si - \mu i \end{aligned} \tag{1.6}$$

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

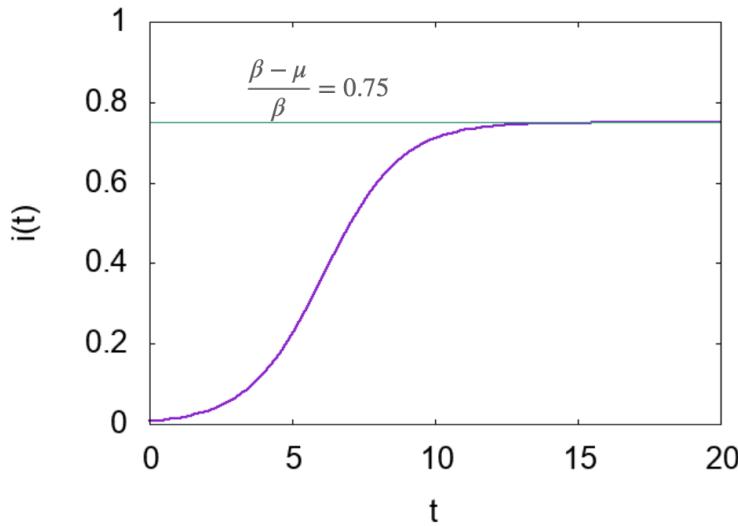
Also, the shape of the **solution** is a sigmoid as before:

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

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  $\mu$ , we can obtain larger fluctuations (Fig. 1.7).

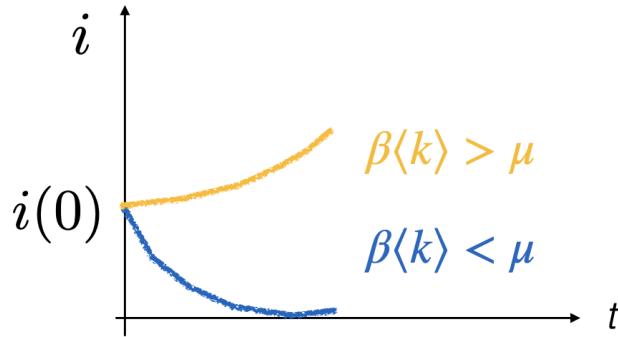
It can be instructive to study what happens according to this model at the **transient**. At the beginning, one can assume that almost the entire population is composed by susceptible people ( $s \sim 1$ ), while the number of infected is very small ( $i \ll 1$ ). Hence, the differential equations can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

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



**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:

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

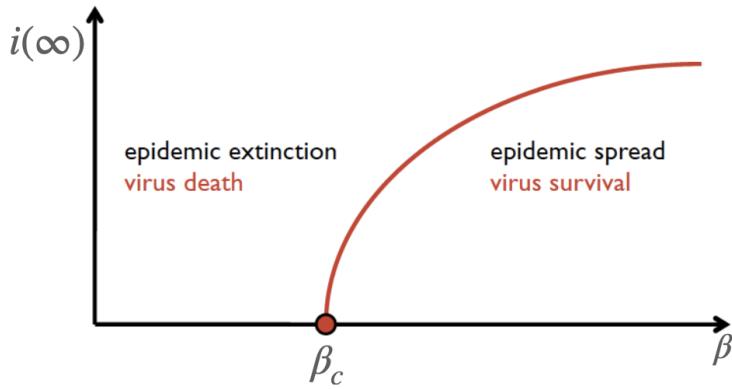
and we have that:

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

where  $\beta_c$  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.  $\beta_c$  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:

$$R_0 = \frac{\beta \langle k \rangle}{\mu} = 1 \quad (1.9)$$

### 1.2.3 SIR model

We now discuss the so called *SIR* model, whose compartments are **Susceptible**, **Infected** and **Recovered**. 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.

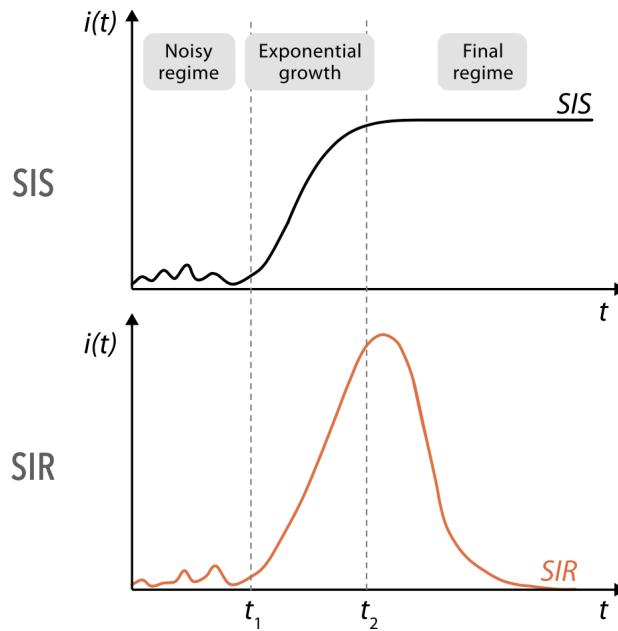
The transitions for this model are:



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:

$$\begin{aligned} \frac{ds}{dt} &= -\beta \langle k \rangle s i \\ \frac{di}{dt} &= \overbrace{\beta \langle k \rangle s i}^{\text{New infections}} - \overbrace{\mu i}^{\text{Recovery}} \\ \frac{dr}{dt} &= \mu i \end{aligned} \quad (1.11)$$



**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<sup>2</sup>. 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:

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

and the result we find is again:

$$\beta > \beta_c = \frac{\mu}{\langle k \rangle} \quad (1.12)$$

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:

$$\frac{ds}{dr} = \frac{-\beta \langle k \rangle s}{\mu}$$

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

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

---

<sup>2</sup>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:

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

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  $\alpha$  is the death rate, common to all classes. Moreover,  $\alpha$  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:

$$\frac{ds}{dt} + \frac{di}{dt} + \frac{dr}{dt} = 0 \quad (1.13)$$

Our equations become then:

$$\begin{aligned} \frac{ds}{dt} &= \alpha - \beta si - \alpha s \\ \frac{di}{dt} &= \beta si - \mu i - \alpha i \\ \frac{dr}{dt} &= \mu i - \alpha r \end{aligned} \quad (1.14)$$

where the **infectious period** is:

$$\tau = \frac{1}{\alpha + \mu} \quad (1.15)$$

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:

$$R_0 = \frac{\beta}{\alpha + \mu} \quad (1.16)$$

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

$$\frac{ds}{dt} = \frac{di}{dt} = \frac{dr}{dt} = 0$$

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

$$\frac{di}{dt} = 0 = \beta si - \mu i - \alpha i \rightarrow \beta s^* i^* - (\mu + \alpha) i^* = 0$$

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

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

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  $\frac{ds}{dt} = \alpha - \beta si - \alpha s$ , 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:

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

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.

### 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:

$$\begin{aligned} \frac{ds}{dt} &= \alpha + wr - \beta si - \alpha s \\ \frac{di}{dt} &= \beta si - \mu i - \alpha i \\ \frac{dr}{dt} &= \mu i - wr - \alpha r \end{aligned} \tag{1.20}$$

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

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

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

while the  **$R_0$**  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

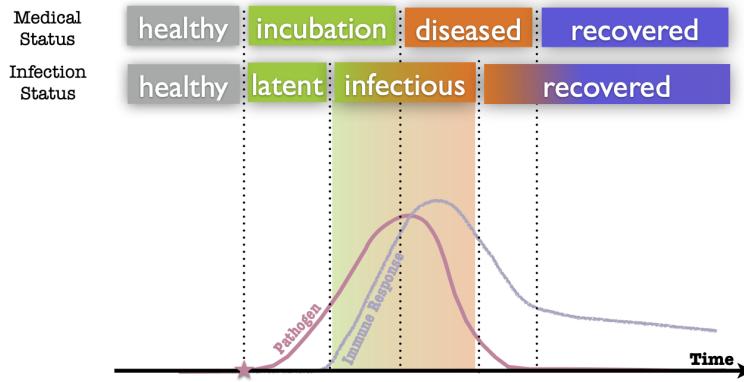
In reality people do not become instantaneously infectious, but there is a **latent period** which is the time between infection and becoming infectious. Indeed, the pathogen replication takes time, i.e. viral load is too low to be able to transmit the infection. This argument leads us to introduce the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

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

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

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:

$$\begin{aligned}\frac{ds}{dt} &= \alpha - \beta si - \alpha s \\ \frac{de}{dt} &= \beta si - (\alpha + \sigma)e \\ \frac{di}{dt} &= \sigma e - (\alpha + \mu)i \\ \frac{dr}{dt} &= \mu i - \alpha r\end{aligned}\tag{1.24}$$

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 ( $\sigma \rightarrow \infty$ ) we recover the endemic state of the SIR.

The **R<sub>0</sub>** factor is:

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

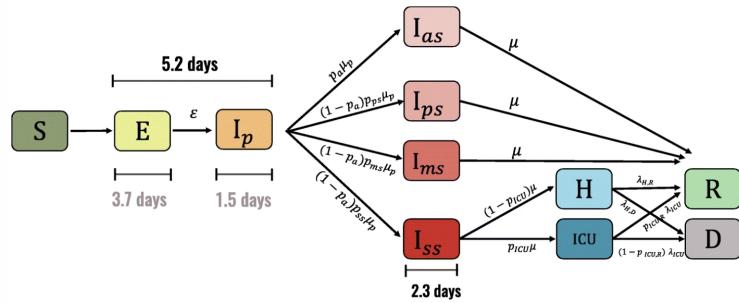
Since latent time is way shorter than demography one, usually  $\frac{\sigma}{\sigma+\alpha} \simeq 1$ , hence  $R_0 = \frac{\beta}{\alpha+\mu}$  as in the SIR with demography.

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:

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

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  $\beta$ , 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:

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

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

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

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

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

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

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

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

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

## Lecture 6.

Friday 16<sup>th</sup>

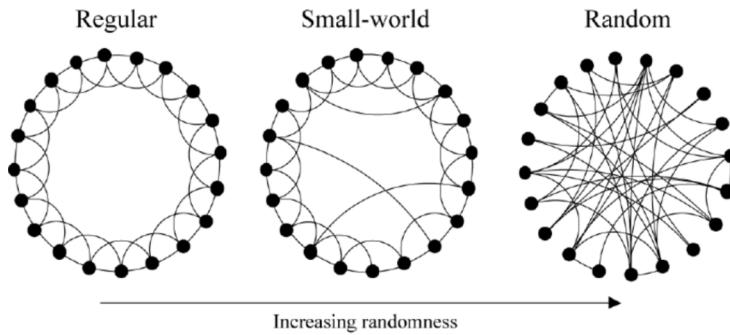
October, 2020.

Compiled: Tuesday

8<sup>th</sup> December,

2020.

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:

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

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:

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

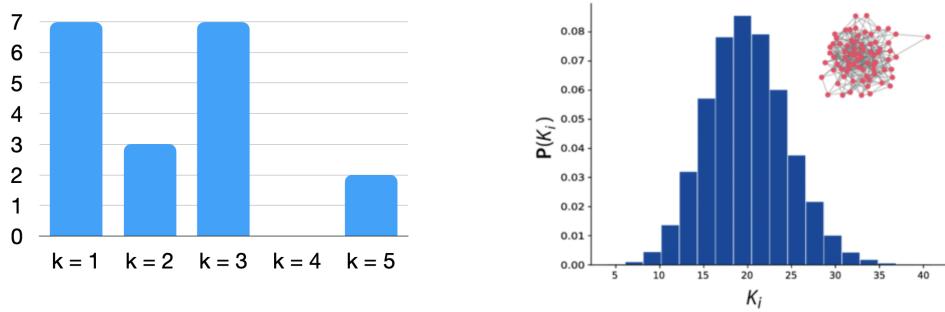
And it is represented by the random circle in Fig. 2.1.

## 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 Erdős and Rényi Model: random graphs

Let us consider the Erdős 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 Erdős 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:

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

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

$$\langle k \rangle = p(N-1), \quad \sigma_k^2 = p(1-p)(N-1) \quad (2.4)$$

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:

$$\frac{\sigma_k}{\langle k \rangle} = \sqrt{\frac{1-p}{p(N-1)}} \xrightarrow{N \rightarrow \infty} 0$$

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:

$$(1-p)^{N-1-k} \approx e^{(N-1-k) \log(1-\langle k \rangle/(N-1))} \xrightarrow{N \rightarrow \infty} e^{-\langle k \rangle}$$

and

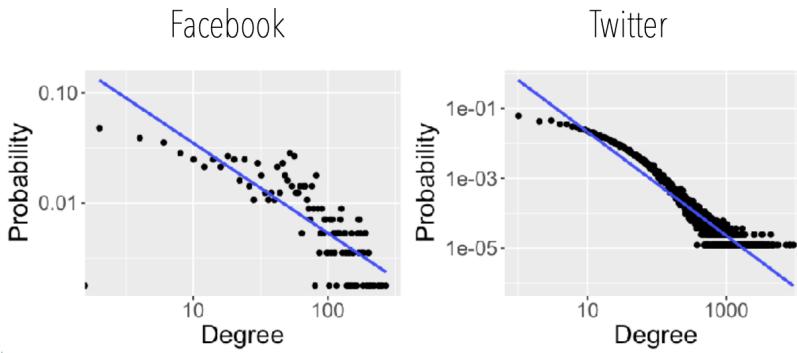
$$\binom{N-1}{k} \approx \frac{(N-1)^k}{k!}$$

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

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \quad (2.5)$$

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  $Np \rightarrow c > 1$ , 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.



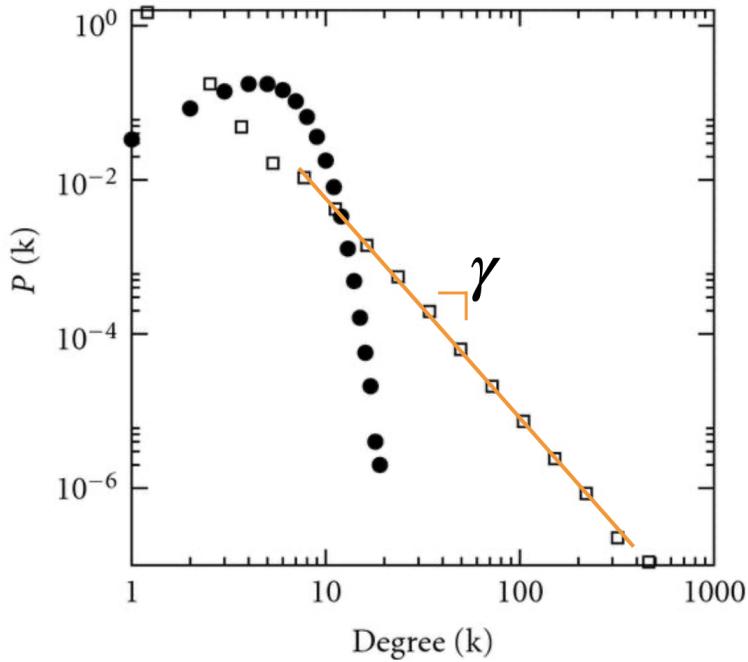
**Figure 2.3:** Real network of Facebook and Twitter.

### 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 show two real social networks we know pretty well: Facebook and Twitter. Note that both plots are in log-log scale. Generalizing, we can say that most of the real networks scales in the same way.

We now want to understand how the **degree distribution** looks like. Let us consider Fig. 2.4: black dots follow the Poissonian distribution that we were mentioning before, while the squares follow a power-law  $P(k) \sim k^{-\gamma}$ , which is **heavy tailed distribution**, in the sense that possibility for large degrees is not null. One should note that the Poissonian distribution is not able to reproduce the heterogeneity we can see in the data, while the power-law is. Hence, in most contexts real networks are **highly heterogeneous** and degrees can span **several orders of magnitude**. In particular, the  $\gamma$  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  $\gamma$  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:

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

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

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

Thus, when:

$$k_{max} = k_{min} N^{\frac{1}{\gamma-1}} \quad (2.7)$$

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:

$$\langle k^n \rangle = \int_{k_{min}}^{\infty} k^n P(k) dk = \int_{k_{min}}^{\infty} C_0 k^{n-\gamma} dk \quad (2.8)$$

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  $t$ :

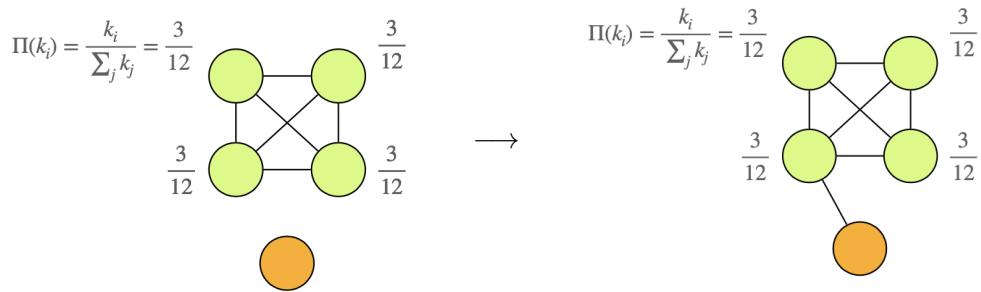
$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \quad (2.9)$$

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 to 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  $\langle k \rangle \rightarrow c$ , but  $\langle k^2 \rangle \rightarrow \infty$  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:

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

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  $\frac{\sigma_k}{\langle k \rangle} = \sqrt{\frac{1-p}{p(N-1)}} \xrightarrow{N \rightarrow \infty} 0$ .

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  $\text{Prob}[\sigma_i(t) = 0, \sigma_j(t) = 1] = 0, \sigma_j(t) = 1]$  are **statistically independent**. This implies that the joint probability can be factorized as follows:

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

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:

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

Actually, this last term is the degree of the network:

$$\sum_j A_{ij} = k_i \simeq \langle k \rangle \quad (3.2)$$

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

$$\frac{d\rho}{dt} = \beta \langle k \rangle (1 - \rho)\rho - \mu\rho \quad (3.3)$$

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 ( $\rho_i$ ). 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.

### 3.1.2 Heterogeneous Networks

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

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

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

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

where  $s_k$  and  $\rho_k$  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  $(s_k, \rho_k)$  in the system, but in this case depending on each degree  $k$ . Obviously, the total fraction of  $\rho$  and  $s$  in the system is given by the sums:

$$\rho = \sum_k P(k)\rho_k, \quad s = \sum_k P(k)s_k \quad (3.4)$$

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

$$\frac{d}{dt}\rho_k(t) = -\mu\rho_k(t) + \beta k (1 - \rho_k(t))\Theta_k(t) \quad (3.5)$$

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:

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

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:

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

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<sup>1</sup>. 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  $\frac{d}{dt}\rho_k(t)$  is similar to the ones previously used in other models. The first assumption is to be in the **steady state**:

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

The next step is then to substitute the expression for  $\rho_k$ , obtained thanks to  $\Theta$ :

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

This is the **self consistent equation** for  $\Theta$ .

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  $\Theta$ :

- 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 intersections between the line  $\Theta$  and the function  $f(\Theta)$  and have to be found graphically (or using computational algorithms).

Since  $\Theta$  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:

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

---

<sup>1</sup>One should keep in mind that  $\sum_k kP(k) = \sum_{k'} k'P(k')$ .

that leads to the following condition:

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

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:

$$\frac{\beta \langle k^2 \rangle}{\mu \langle k \rangle} = 1 \quad \rightarrow \quad \beta_c = \frac{\mu \langle k \rangle}{\langle k^2 \rangle} \quad (3.9)$$

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:

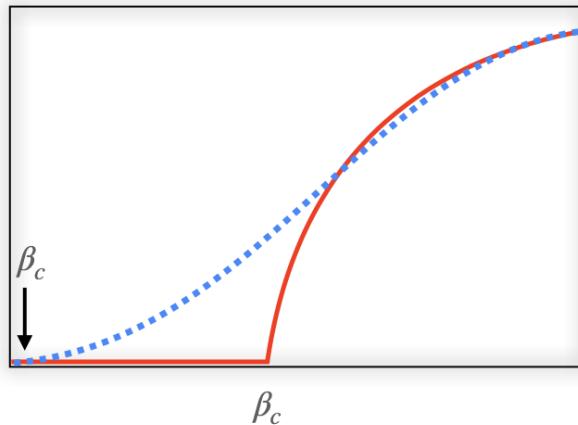
$$\beta_c = \frac{\mu \langle k \rangle}{\langle k^2 \rangle} = \frac{\mu}{\langle k \rangle}$$

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  $\langle k \rangle \rightarrow c$  and  $\langle k^2 \rangle \rightarrow \infty$  as  $N \rightarrow \infty$ . As the network becomes larger also its variance increases, that is:

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

hence the **epidemic threshold vanishes** for  $N \rightarrow \infty$ . 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<sup>2</sup>! 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.

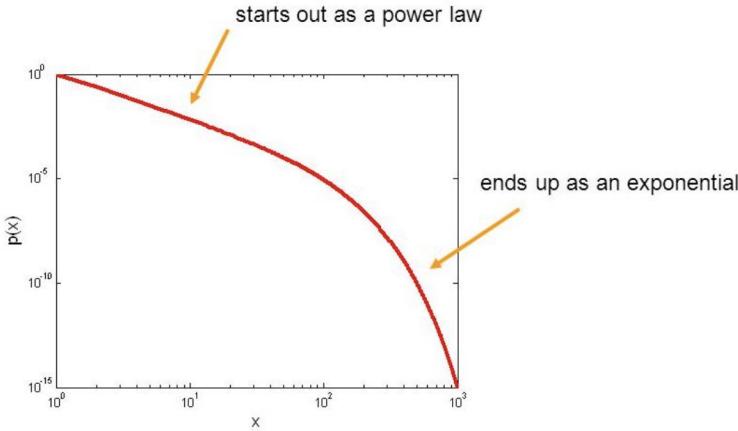


**Figure 3.1:** In scale-free networks (and many heavy-tailed distributions) the epidemic threshold vanishes in the thermodynamic limit.

<sup>2</sup>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:

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

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:

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

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  $\rho_\infty^R = \lim_{t \rightarrow \infty} \sum_k P(k) \rho_k^R(t)$ . Equations take the form:

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

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

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

is the probability of a contact with an infected node, and plays exactly the same role of  $\Theta$  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:

$$\beta_c = \frac{\mu \langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \quad (3.14)$$

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  $P(k)$ <sup>3</sup>.

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:

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

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**

---

<sup>3</sup>the so called "ensemble" of networks!

by, at least, one neighbour. The basic idea for doing this is:

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

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  $j$ .

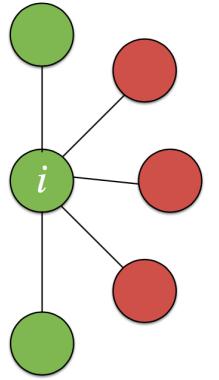
$$[1 - \beta A_{ij} \rho_j(t)]$$

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

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

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:

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



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:

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

it follows that:

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

Now, if we think about what happens when we are in **proximity of the epidemic threshold** (*epidemic onset*), it happens that  $\rho_i^*$  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:

$$q_i^* = 1 - \prod_{j=1}^N [1 - \beta A_{ij} \varepsilon_j^*] \simeq \beta \sum_{j=1}^N A_{ij} \varepsilon_j^* \quad (3.19)$$

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

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

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

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

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

Neglecting second order terms, we have that:

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

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:

$$\beta = \frac{\mu}{\Lambda_i} \quad (3.22)$$

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

$$\beta_c = \frac{\mu}{\Lambda_{max}} \quad (3.23)$$

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 threshold

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:

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

And in particular:

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

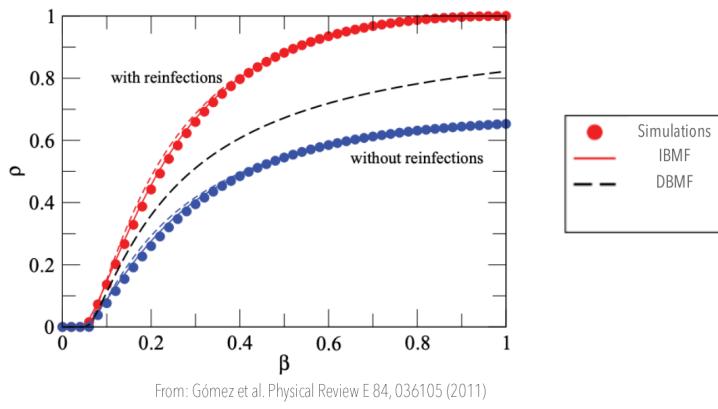
We can conclude that **IBMF** is **more accurate** than DBMF. Due to the approximation, indeed, the **DBMF** is accurate **only** in the **proximity of the epidemic threshold**, while IBMF is accurate for the entire epidemic diagram.

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

$$\beta_c \xrightarrow{N \rightarrow \infty} 0 \quad (3.26)$$

Moreover, the epidemic threshold for IBMF depends on the largest eigenvalue of the adjacency matrix  $\Lambda_{max}$ . The last relations which contain  $\beta_c$  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 29<sup>th</sup>  
October, 2020.  
Compiled: Tuesday  
8<sup>th</sup> December,  
2020.



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

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

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

Let us try to characterize the **annealed network**, in particular we want to see the *adjacency matrix* looks like. We start from:

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

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

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

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 adjacency 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 are susceptible and given that a neighbor of ours was infected:

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

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:

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

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:

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

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.

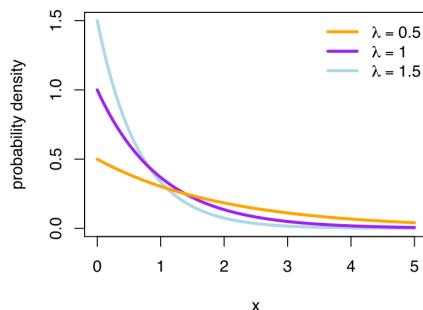
All the models we have seen till now assume that  $\beta$  (infection process) and  $\mu$  (recovery process) are constant rates. This means that movement between compartments takes place at constant rate, or equivalent the probability of jumping from one compartment to another does not depend on the 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 ( $\beta$  and  $\mu$ ) imply that the time spent inside each compartment  $\tau$  follows a **exponential distribution** (Fig. 4.1):

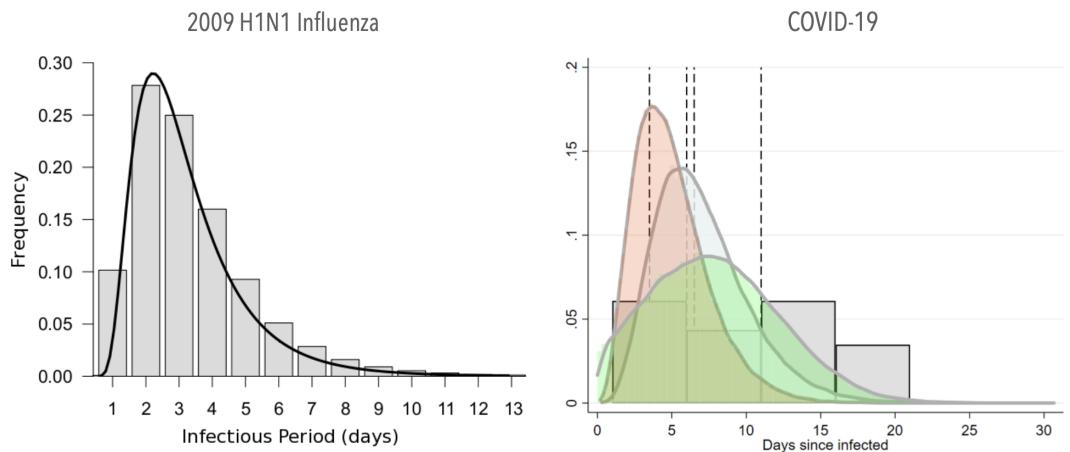
$$P(x) = \tau e^{-\tau x} \quad (4.1)$$

with mean  $\tau = \frac{1}{\mu}$ , i.e.  $\tau$  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 infectious 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 know when it ends. For instance, let us consider the left plot in Fig. 4.2 for 2009 H1N1 Influenza. For this type of disease the plot shows the distribution of the infectious period, which has as probable value 2 days and a half. The most important thing 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 2009 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**:

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

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

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

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:

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

or the **Weibull distribution**:

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

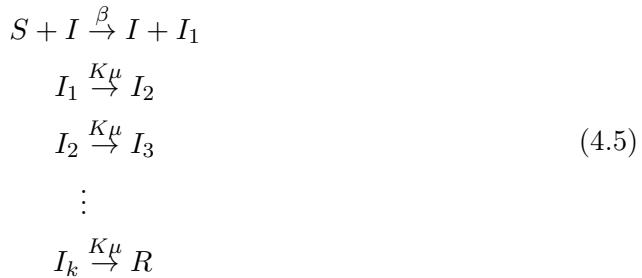
All of these distribution are able to reproduce real histograms.

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 infectious 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 constant rate (exponential 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  $S I_1 I_2 \dots I_k R$ . For instance, we are imposing that these transitions are sequential:



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:

$$\begin{aligned} \frac{ds}{dt} &= -\beta s i \\ \frac{di_1}{dt} &= \beta s i - K\mu i_1 \\ \frac{di_2}{dt} &= K\mu i_1 - K\mu i_2 \\ &\vdots \\ \frac{dr}{dt} &= K\mu i_k \end{aligned} \quad (4.6)$$

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:

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

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  $K \rightarrow \infty$  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  $\lambda$  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 limitaions:

- 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 represents the number of trials that  $i$  try to infect node  $j$  while infected;
4. we generate a sequence of times

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

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 transmissibility of the disease is seen as how many trials I am gonna make to infect. One important thing 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 numbers 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  $i$ :  $v_i S / E[R]$ . Thus we have:

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

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

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

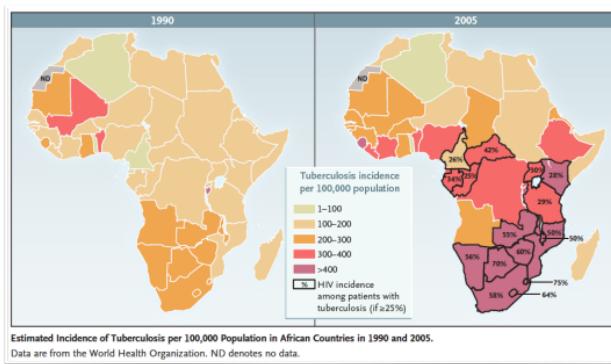
Hence, this is the form in which we can generate a generic model from.

## 4.3 Interacting diseases

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

Let us now consider the simplest case, where we have only two different diseases that **cooperate**. That is to say that one disease **boosts** the 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 years or, sometimes, we may have contracted it without it never showing up. Let us compare 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

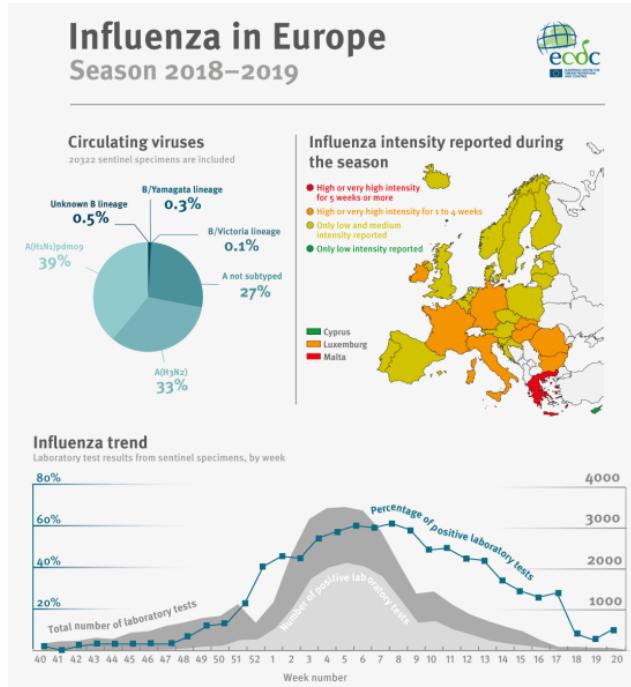
**Figure 4.3:** Map of the prevalence of *TB* for different years (1990 and 2005).

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

**Lecture 11.**  
Thursday 5<sup>th</sup>  
November, 2020.  
Compiled: Tuesday  
8<sup>th</sup> December,  
2020.



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

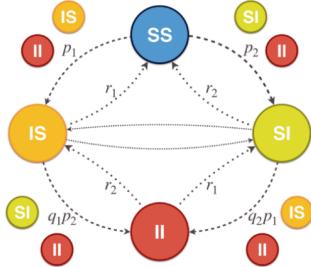
susceptible population is just a reduced fraction of the whole one, and should explain why we do not usually observe influenza pandemics.

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 either one of the following states  $\{SS, SI, IS, II\}$  with its own probability density  $\{[\rho^{SS}]_t^i, [\rho^{SI}]_t^i, [\rho^{IS}]_t^i, [\rho^{II}]_t^i\}$ . Therefore, every disease will be defined thanks to its own **parameters**  $\{\beta_1, \beta_2, \mu_1, \mu_2\}$ . However, some more are to be introduced, namely the ones that **encode** the **interaction** between diseases.

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  $P(k, l)$ .

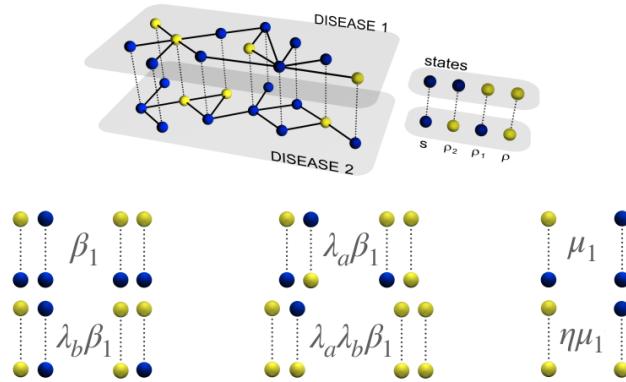


**Figure 4.5:** Coupled SIS model and all possible combinations and different transition probabilities between each compartment.

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  $\lambda_a$ :** being infected of one disease 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  $\lambda_b$ :** once we get infected by one disease, we are less infectious wrt second one
- **modified infectious period  $\eta$ :** 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:

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

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

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

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

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

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

While for network 2:

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

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  $\sigma_1$  and  $\sigma_2$ , 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.

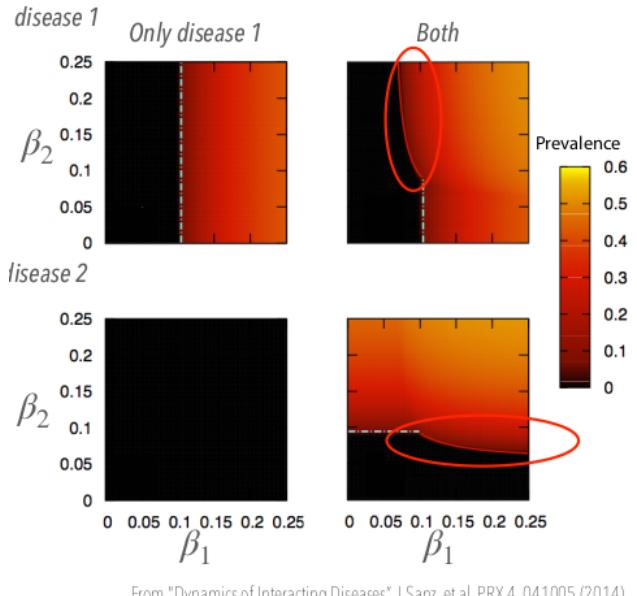
One should see from the formula that the **epidemic threshold** of the **first** disease **depends** on the **prevalence of the other**. An other way to see it is that we are assuming that one disease is already there, and then insert a second one and check the effects on the epidemic threshold.

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  $\eta < 1$ .

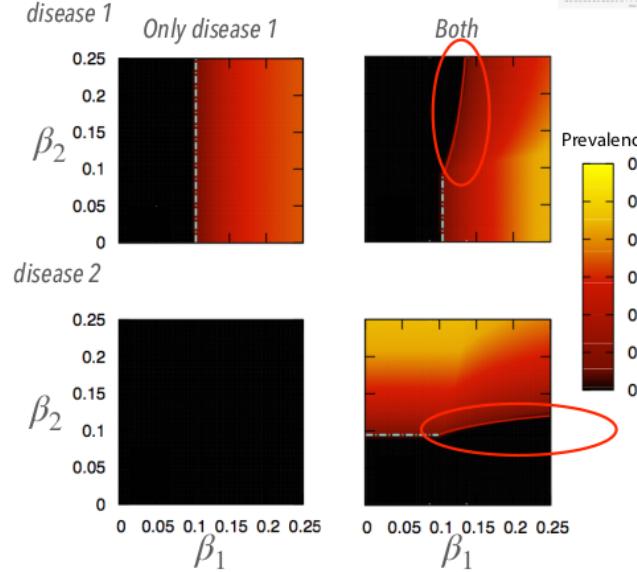
If we consider a single disease, one see that we get back the old results (in 2D): since we have not introduced yet a second disease we are not able to see it. On the other hand, if the parameter  $\beta_2$  is below the critical threshold, we cannot note any difference and the spreading is the same 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



From "Dynamics of Interacting Diseases" J.Sanz, et al. PRX 4, 041005 (2014)

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



From "Dynamics of Interacting Diseases" J.Sanz, et al. PRX 4, 041005 (2014)

**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 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:

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

where:

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

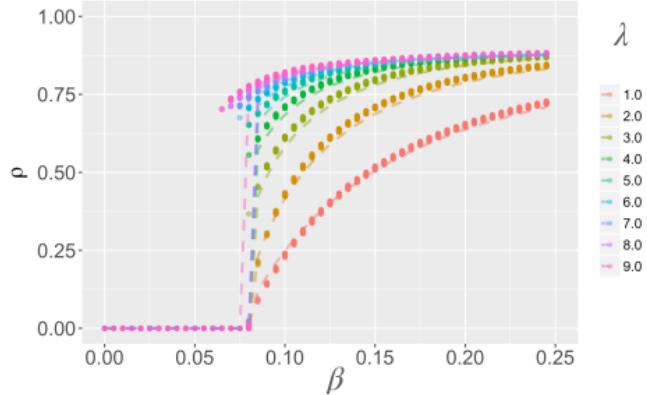
and:

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

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:

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

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

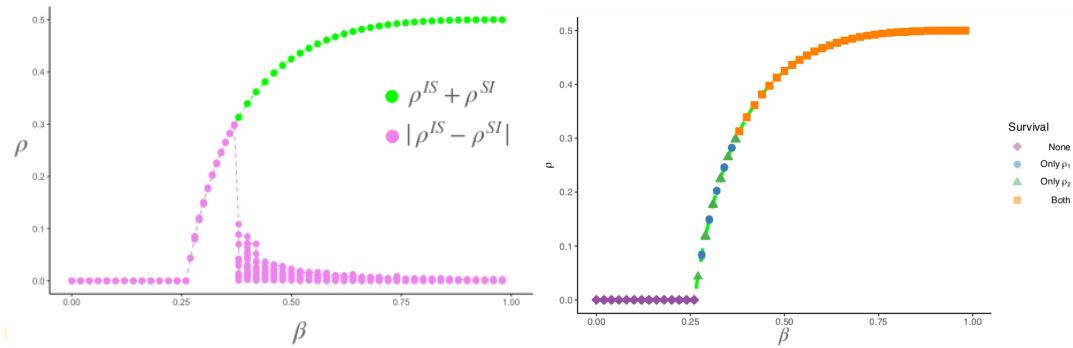


**Figure 4.9:** Cooperating diseases, numerical simulations results for different  $\lambda$ .

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:

$$\rho = \frac{1}{N} \sum_i^N (\rho_i^{SI} + \rho_i^{IS}) \quad (4.17)$$

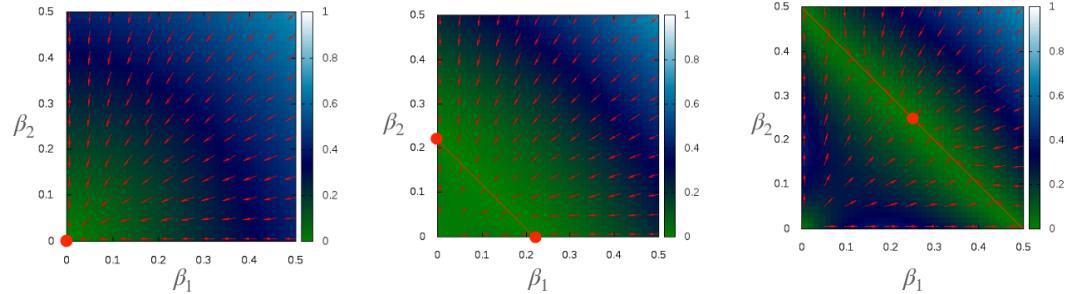
As expected, the prevalence shown in left figure 4.10 does not change at all and is the same as before. While, plotting the difference  $|\rho^{IS} - \rho^{SI}|$  we observe some oscillations in values slightly after the critical threshold. This is explained by saying that only a disease can survive, and which is given by chance, being the two symmetrical (see right fig 4.10). As  $\beta$  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  $\beta$  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  $(\beta_1, \beta_2)$  space: here obviously both of them are absent. We can continuously move these points and increase  $(\beta_1, \beta_2)$  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  $\beta_1$  and  $\beta_2$ .



# 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 half 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  $R \rightarrow S$  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  $\beta$  and  $\mu$  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:

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

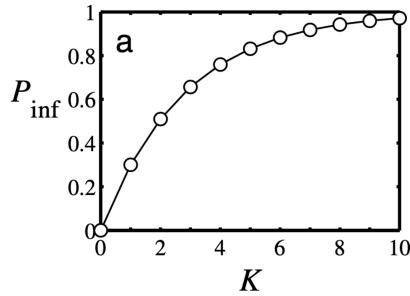
Let us discuss the single terms.  $p^k (1-p)^{K-k}$  is the probability for the contact to be successful. Note as it a *Bernoulli distribution* with  $K$  trials and  $k$  successes. This is multiplied by the Binomial coefficient that takes into account all the possible combinations of  $k$  successes 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:

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

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:

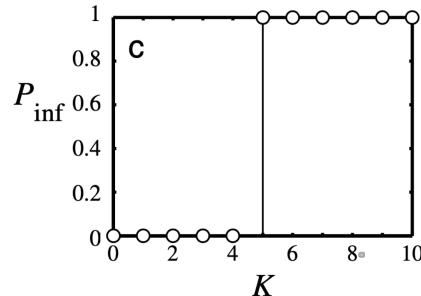
- if the probability of a successful contact  $p < 1$ , the dose has a fixed size  $f(d) = \delta(d-1)$  and fixed threshold  $g(d^*) = \delta(d^*-1)$ , we observe **epidemic spreading** where interactions are independent (see fig. 5.1). In particular, all contacts share the same infection probability 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.

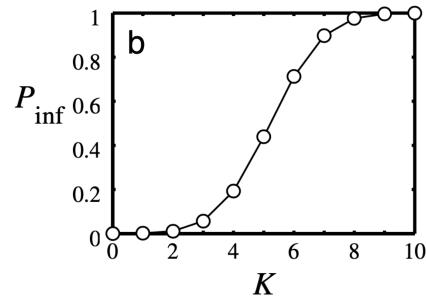
- if the probability of a successful contact  $p = 1$ , the dose has fixed size  $f(d) = \delta(d-1)$  and fixed threshold  $g(d^*) = \delta(d^*-5)$ , we obtain a **deterministic**

**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.<sup>1</sup>



**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.<sup>2</sup>



**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:

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

<sup>1</sup>for instance 5 friends that show to me the same information.

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

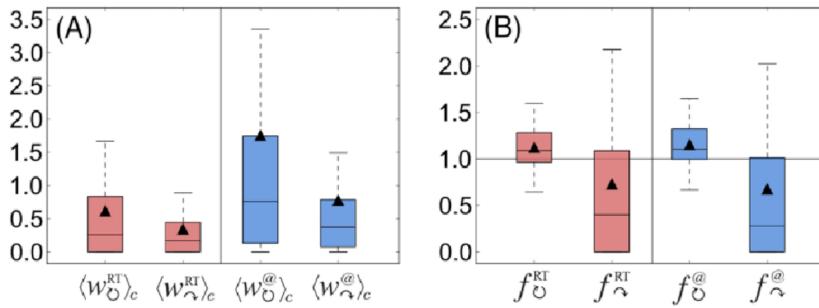
- $\langle w_{\sim} \rangle_c$  is the average weight (number of tweets) per link outside the community.

And the same for users activity:

- $f_{\circ}$  is the fraction of activity inside the community;
- $f_{\sim}$  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			Simulation implementation
Network	Reinforcement	Homophily	
$M_1$	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.

Average behavior

Simple contagion

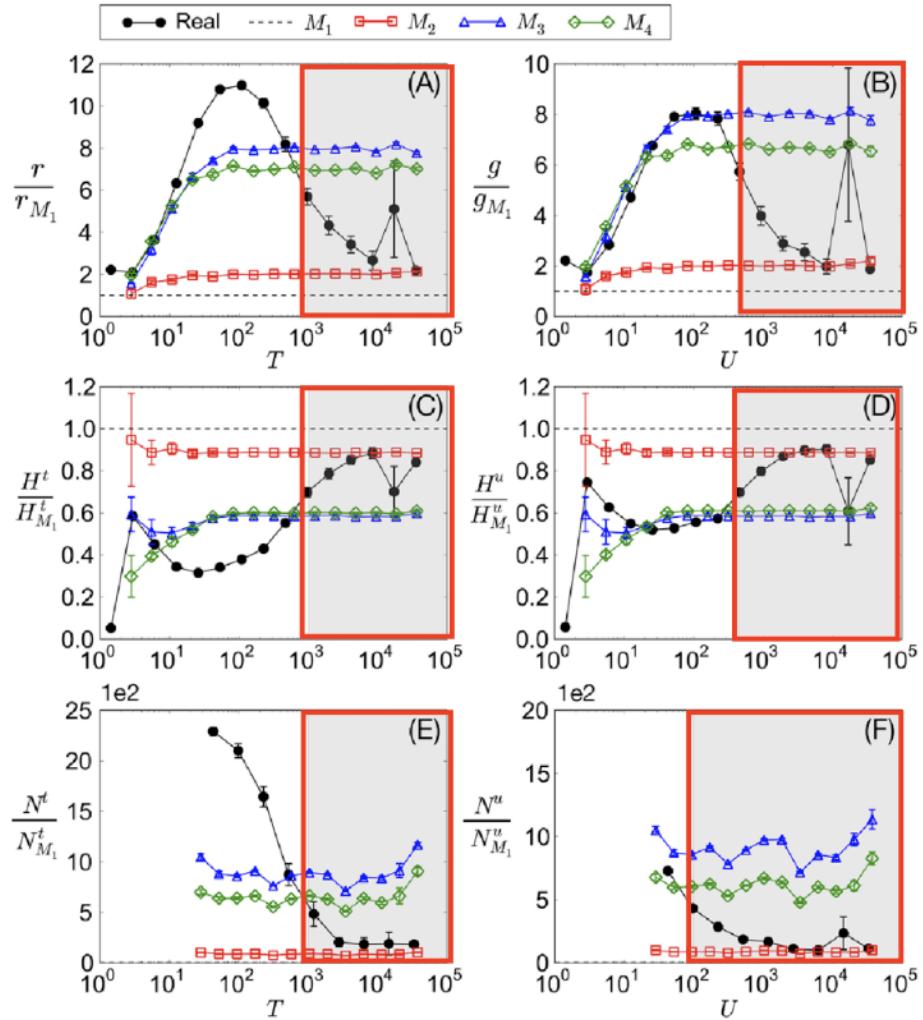
Complex contagion

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

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

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

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



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



## Part II

# Poletto's Lectures



# 6

# Introduction to metapopulation models

## 6.1 Spatial spread of epidemics

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

**Lecture 14.**  
Friday 13<sup>th</sup>  
November, 2020.  
Compiled: Tuesday  
8<sup>th</sup> December,  
2020.

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

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

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

### 6.1.1 Human mobility

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

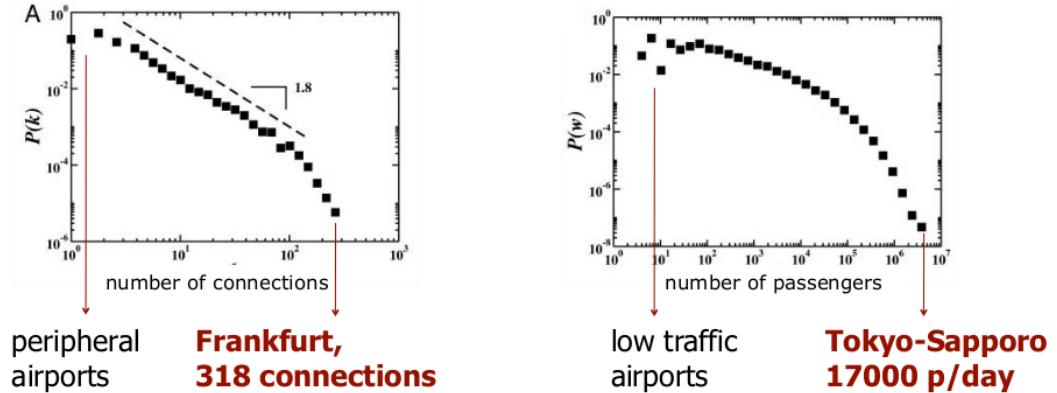
- **segment**: number of seats for each company between two airports;

---

<sup>1</sup>Human mobility: Models and applications, Barbosa et al. Physics Reports 734 (2018)

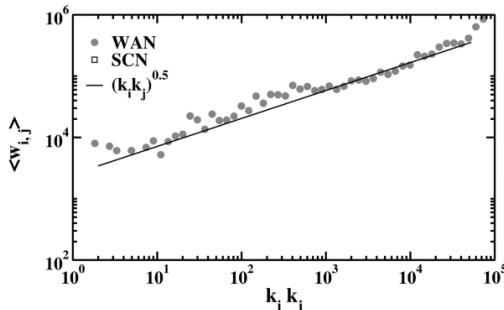
- **origin-destination:** number of passengers travelling between origin-destination, obtained from the tickets purchased.

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



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

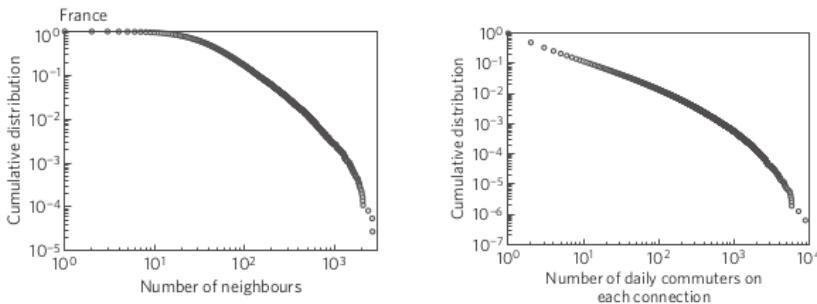
We can indeed find some **scaling relations** between fluxes, number of connections and population, as one can see from Fig. 6.2. The average number of route  $i \rightarrow j$  is determined by a non linear function of the traffic in airports, and it has form  $w_{ij} \sim (k_i k_j)^{\theta}$  with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



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

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

We want now to point out the **differences** between the *air travelling* and the *commuting* networks. In order to do so, first we obviously need to use the same spatial resolution: this is why researchers defined **macro urban areas** centered around airports. Now we are able to look at what differs one network from the other. The first feature one may want to analyze is the average **daily number of travellers** in a certain area: it is about 1000 for the air travel, while 20'000 for the commuting network. Order or magnitudes are indeed different. Another important point is the **daily traveling rate**; it is more probable to commute ( $10^{-2} \text{days}^{-1}$ ) rather than



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

air travel ( $10^{-3} \text{ days}^{-1}$ ). Finally **time scales** are different: one usually travels by plane every days/weeks, whereas commuting is a matter of hour. In conclusion, **commuting** has a **faster dynamics** and leads to a **higher** level of **mixing**.

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

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

As one may imagine **data** is very **heterogeneous**, being heterogeneous the sources where we collect it from. Heterogeneity can be seen in spatial resolution, individuals-level/origin-destination fluxes/seats, broken down per transportation media or per purpose of the trip. Since every dataset provides **partial information**, one may think to try to **combine** some of them. For instance, this cannot be done for air-travel and commuting network, being the spatial ranges very different. In addition,

<sup>2</sup>[www.wheresgeorge.com](http://www.wheresgeorge.com)

combining cell-phones data and commuting we are able to extract commuting proxies from cell-phone data. Finally, we should state that we cannot spot clearly the differences that occur for people of different ages: among air travellers indeed there are few children and old people, and statistically they look like the same and cannot be distinguished.

### 6.1.2 Modelling Human Mobility

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

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

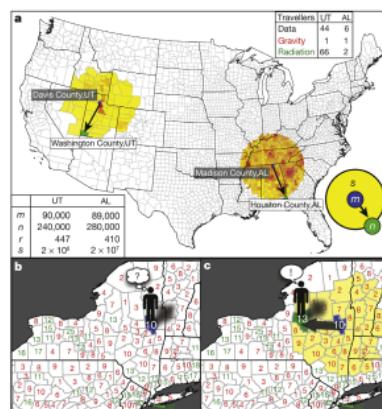
The **Gravity Model** was first introduced by G.K. Zipf. He took inspiration from Newton's law of gravitation in order to describe **mobility flows**:

$$T_{ij} \propto \frac{N_i N_j}{d_{ij}} \quad (6.1)$$

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

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

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



[Simini et al Nature 2012]

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

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

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

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

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

## 6.2 Integrating Human Mobility in Epidemic Models

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



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

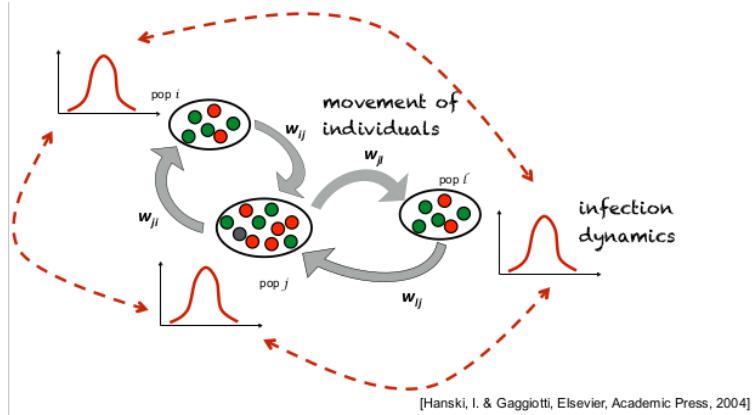
In ecology the dynamics is therefore driven by stochastic effects, which may lead for instance to extinction or recolonisation. This, obviously, will suggest us to find an analogy when dealing with epidemic models. However, one should keep into account that the **discrete nature of individual** is one of the most essential ingredients to describe the dynamics: it is meaningless to state that half an individual travel between two patches. The first models assumed that the mixing between patches occurred homogeneously, while more recently more complexity has been added: **mixing** among patches has started to be **mediated** by the **human mobility network** hence coupling the metapopulation perspective with network theory.

### 6.2.1 SIR metapopulation model

Let us discuss now how we can introduce the metapopulation concept into a model we have studied so many times: the *SIR* model (see fig. 6.6). The only difference

<sup>3</sup>Levins Bull. Entomol. Soc. Am., 15 (3) (1969).

here is that it is **not** an **individual-based** model any more. now we do not keep track of every individual, but we just monitor the occupation number of patches and compartments.



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

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

$$S(t) = S_1(t) + S_2(t) + S_3(t) + \dots + S_V(t) = \sum_i S_i(t) \quad (6.4a)$$

$$I(t) = I_1(t) + I_2(t) + I_3(t) + \dots + I_V(t) = \sum_i I_i(t) \quad (6.4b)$$

$$R(t) = R_1(t) + R_2(t) + R_3(t) + \dots + R_V(t) = \sum_i R_i(t) \quad (6.4c)$$

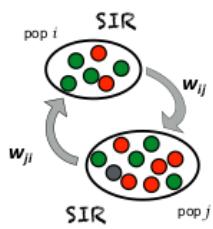
$$N(t) = N_1(t) + N_2(t) + N_3(t) + \dots + N_V(t) = \sum_i N_i(t) \quad (6.4d)$$

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

$$\frac{dS_i}{dt} = -\beta \frac{I_i(t)S_i(t)}{N_i} + \Omega_i^S \quad (6.5a)$$

$$\frac{dI_i}{dt} = \beta \frac{I_i(t)S_i(t)}{N_i} - \mu I_i(t) + \Omega_i^I \quad (6.5b)$$

$$\frac{dR_i}{dt} = \mu I_i(t) + \Omega_i^R \quad (6.5c)$$



**Figure 6.7:**  
Mobility between two different patches occurs with probability  $p_{ij} = w_{ij}/N_i$ .

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

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

$$p_{ij} = \frac{w_{ij}}{N_i} \quad (6.6)$$

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

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

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

Therefore, a formula for  $\Omega_i^X$  can be the following:

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

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

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

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

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

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

## 6.3 Application of metapopulation models

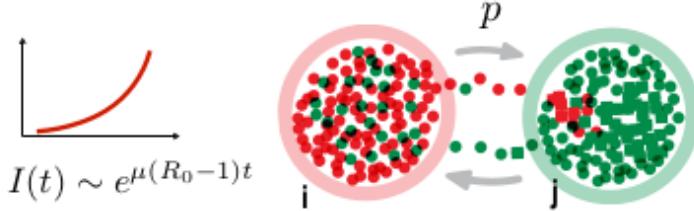
In the last lecture we have introduced the so called **SIR metapopulation model**, deriving it and understanding it *analytically*. Now we want study it further, focusing on the **spatial propagation** and **predictability** under the assumption of Markovian mobility. Note that beside it, there might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

### 6.3.1 Spatial propagation dynamics

Let us discuss the **dynamics** of spatial spreading when we are *above* the **epidemic threshold**. Given that an epidemic has started in a given city  $i$ , we want to

**Lecture 15.**  
*Thursday 19<sup>th</sup>*  
*November, 2020.*  
*Compiled: Tuesday*  
*8<sup>th</sup> December,*  
*2020.*

understand how it will spread to  $j, h\dots$ . Let us define the **seeding time** (or *arrival time*) as it follows: it is the time of arrival of the *first* case in patch  $j$ . Let us focus on a 2 patches model (see fig 6.8). Recalling that  $p$  is the travelling probability from  $i$  to  $j$ ,  $I(t)$  is the number of infectious in patch  $i$ , the **probability** that an infected individual arrives in  $j$  at time  $t$  is  $[1 - (1 - p)^{I(t)dt}]$ .



**Figure 6.8:** Dynamics of spatial spread:  $p$  is the traveling probability  $i \rightarrow j$ , and  $I(T)$  is the prevalence in patch  $j$ .

We can now try to compute the **probability** that the **first** infectious arrives in  $j$  at time  $t$ , and it is given by:

$$P(t_{\text{seeding}} = t) = \prod_{s=1}^{t-1} (1-p)^{I(s)dt} \cdot [1 - (1-p)^{I(t)dt}] \xrightarrow{p \rightarrow 0} P(t_{\text{seeding}} = t) = pI(t)e^{-p \int_0^t I(s)ds} \quad (6.9)$$

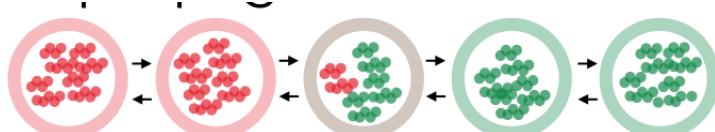
Defining  $a = \mu(R_0 - 1)$ , we can rewrite the probability as:

$$P(t_{\text{seeding}} = t) = pI(t)e^{at}e^{-pa e^{at}} \quad (6.10)$$

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

$$\langle t_{\text{seeding}} \rangle \simeq \frac{1}{a} \ln(pa) \quad (6.11)$$

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



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

Now it occurs that there is **correlation** among patches, since the time at which the patch  $i$  is infected does depend on the previous one, that in turn depends on the one before and so forth. Infected people that travel between patches are actually important only at the starting of each infection process for every patch, but later these can be neglected since their contribution is small compared to the normal spreading dynamics. We can define as  $\Delta_i$  the interval that interoccurs between two consecutive seedings, obviously related to different patches:

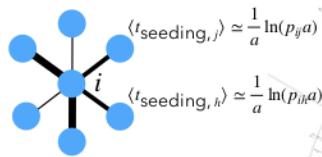
$$\langle t_{\text{seeding},i} \rangle - \langle t_{\text{seeding},i-1} \rangle = \Delta_i, \quad \langle t_{\text{seeding},n} \rangle = \sum_{i=1}^n \Delta_n \quad (6.12)$$

Consequently  $\Delta_i$  are correlated and not identically distributed: incidence dynamics in city  $i$  is a combination of the introductions from  $i - 1$  (travelling rate, infected individual travelling...) and infection transmission with patch  $i$ . However, the simple approximations one can think of is:

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

does not work so bad<sup>4</sup>.

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



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

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

As one can clearly understand from Fig. 6.11, introducing the **degree heterogeneity** decreases the predictability: in this way we are including hubs in our model. On the other hand, if we consider also **pathways** hence bringing in **weight fluctuations** numerical simulations are more likely to return similar outputs (the predictability is increased).

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

$$\langle t_{\text{seeding},T.R.} \rangle \simeq \frac{1}{a} \ln(pa\omega) \quad (6.13)$$

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

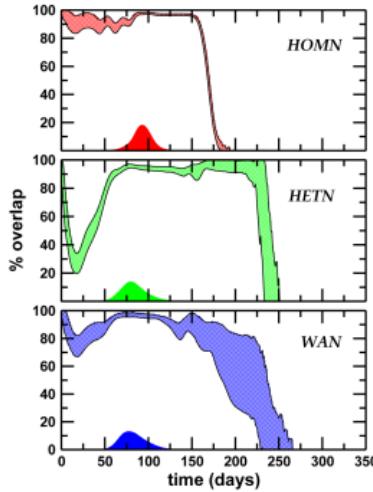
$$\langle t_{\text{seeding},T.R.} \rangle - \langle t_{\text{seeding}} \rangle \simeq \frac{1}{a} \ln(pa\omega) - \frac{1}{a} \ln(pa) = \frac{1}{a} \ln(\omega) \quad (6.14)$$

<sup>4</sup>Gautreau et al JTB 2008.

<sup>5</sup>Brockmann, Helbing, Science 2013.

<sup>6</sup>Colizza, Barrat, Barthelemy et Vespignani, PNAS (2006).

<sup>7</sup>Gautreau et al JTB 2008; Hollingsworth et al Nature Med 2006; Scalia Tomba et al Math Biosci 2008.



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

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

We want now to briefly discuss how these arguments can be applied to **epidemic assessment**, in particular wrt **COVID-19** situation<sup>8</sup>. One of the first early warnings emanated by the Wuhan Municipal Health Committee was the following: “*urgent notice on the treatment of pneumonia of unknown cause*”. This was actually picked up by ProMED-mail, which is an independent program for emerging diseases, that raised an alert to the situation in Wuhan at the beginning of January. Later, some cases of pneumonia were discovered far from the epicentre among travellers, namely 2 in Thailand and 1 in Japan: compared to the 40 *local* cases and to the traffic volume (i.e.  $p$  to travel) this was really out of scale. This argument was actually developed by the Imperial College, stating that<sup>9</sup> this new pneumonia had already started spreading worldwide and provided an estimate for the number of cases. We want now to follow the argument developed by the just mentioned report. Clearly, the number of infected that travel is  $I_{travel} = I_{Wuhan} \cdot p_{travel}$ , where the probability that an infectious travels  $p_{travel}$  is the product between the daily probability to travel  $p_{daily}$  and the time to detect a case  $T_d$ :

$$p_{travel} = p_{daily} T_d \quad (6.15)$$

where  $p_{daily}$  is the probability of travelling out of Wuhan, namely  $p_{daily} = \frac{w}{N}$  and the time to detect a case  $T_d$  is the sum of two terms: *incubation period* and *time to*

---

<sup>8</sup><https://www.nytimes.com/interactive/2020/03/22/world/coronavirus-spread.html>.

<sup>9</sup><https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-1-case-estimates-of-covid-19/>

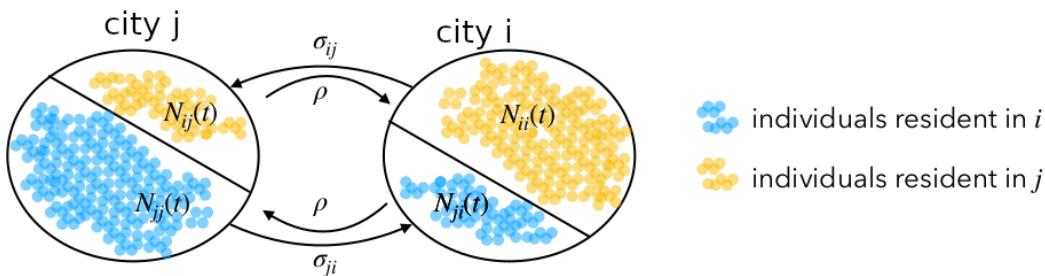
*hospitalization* (first cases were detected only after hospitalization, i.e. when symptoms started to be severe). Using real parameters, namely the passenger per day at Wuhan airport  $w = 3301$ , the catchment population of the just mentioned airport  $N \sim 10^6$  individuals and the incubation period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

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

### 6.3.2 SIR metapopulation model with memory

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

In general, treating mathematically the interplay between mobility and transmission is very difficult. The problem can be solved by using **time scale separation**, that works as a sort of *mean field approximation*. Hence, either the **epidemic unfolds faster than mobility**, that is the case of flu ( $\mu \sim 10^{-1}$  days<sup>-1</sup>) and air traveling (rate  $\sim 10^{-3}$  days<sup>-1</sup>) or alternatively **mobility is faster than the epidemic**, that is the case of commuting (usually we travel twice a day, so the rate is  $\sim 0.5$  days<sup>-1</sup>) and flu ( $\mu \sim 10^{-1}$  days<sup>-1</sup>).



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

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

<sup>10</sup>Global Epidemic and Mobility Model.

<sup>11</sup>Sattenspiel, L. & Dietz, K. Math. Biosci. 128, 71–91 (1995)

<sup>12</sup>Keeling, M. J. & Rohani, P. Ecol. Lett. 5, 20–29 (2002)

traveling to  $j$ , while we approximate the **population** for a given patch  $i$  as *constant*:  $N_i = N_{ii}(t) + \sum_j N_{ij}(t)$ . In other words we are assuming no immigration and no births nor deaths. In this way we can keep track of from where and individual left.

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

$$\partial_t N_{ii} = - \sum_j \sigma_{ij} N_{ii}(t) + \rho \sum_j N_{ij}(t) \quad (6.16a)$$

$$\partial_t N_{ij} = \sigma_{ij} N_{ii}(t) - \rho N_{ij}(t) \quad (6.16b)$$

The linear ordinary **differential equation** of order 1 to be solved can be rewritten as:

$$\partial_t N_{ii}(t) + (\rho + \sigma_i) N_{ii}(t) = N_i \rho \quad (6.17)$$

The first solution is:

$$\begin{aligned} N_{ii}(t) &= e^{(\rho+\sigma_i)t} \left( C_i i + N_i \rho \int_0^t e^{(\rho+\sigma_i)s} ds \right) = \\ &= \frac{N_i}{1 + \sigma_i/\rho} + \left( N_{ii}(0) - \frac{N_i}{1 + \sigma_i/\rho} \right) e^{-\rho(1+\sigma_i/\rho)t} \end{aligned} \quad (6.18)$$

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

$$\begin{aligned} N_{ij}(t) &= \frac{\sigma_{ij} N_i / \rho}{1 + \sigma_i / \rho} - \frac{\sigma_{ij}}{\sigma_i} \left( N_{ii}(0) - \frac{N_i}{1 + \sigma_i / \rho} \right) e^{-\rho(1+\sigma_i/\rho)t} + \\ &\quad + \left[ N_{ii}(0) - \frac{\sigma_{ij} N_i / \rho}{1 + \sigma_i / \rho} - \frac{\sigma_{ij}}{\sigma_i} \left( N_{ii}(0) - \frac{N_i}{1 + \sigma_i / \rho} \right) \right] e^{-\rho t} \end{aligned} \quad (6.19)$$

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

$$[\rho(1 + \sigma_i / \rho)]^{-1} \sim \rho^{-1} = \tau \quad \text{since } \rho \gg \sigma_i \quad (6.20)$$

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

$$N_{ii}(t) = \frac{N_i}{1 + \sigma_i / \rho}, \quad N_{ij}(t) = \frac{\sigma_{ij} N_i / \rho}{1 + \sigma_i / \rho} \quad (6.21)$$

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

$$N_i^* = N_{ii} + \sum_j N_{ji} = \frac{N_i}{1 - \sigma_i / \rho} + \sigma_j \frac{N_j \sigma_{ji} / \rho}{1 - \sigma_j / \rho} \quad (6.22)$$

The ratio actually  $\sigma_i / \rho$  quantifies the proportion of time spent outside and in the residence population.

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

- $\sigma_i \rightarrow 0 \implies N_{ii}(t) \rightarrow N_i; N_{ij}(t) \rightarrow 0; N_i^* \rightarrow N_i$  in this case people rarely leave their residence, thus the non travelling individuals approaches the population of residents;
- $\rho \rightarrow \infty \implies N_{ii}(t) \rightarrow N_i; N_{ij}(t) \rightarrow 0; N_i^* \rightarrow N_i$  people return home immediately thus the non travelling individuals approaches the population of residents;

- $\rho \rightarrow 0 \implies N_{ii}(t) \rightarrow 0; N_{ij}(t) \rightarrow \frac{\sigma_{ij}}{\sigma_i} N_i; N_i^* \rightarrow \sum_j \frac{\sigma_{ji}}{\sigma_j} N_j$  here the case collapses to **migration** process: people never get back and the population of resident in  $i$  is distributed among the neighbouring destinations  $j$ .

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

$$X_{ii}^{[m]} = \frac{X_i^{[m]}}{1 + \sigma_i/\rho}, \quad X_{ii}^{[m]} = \frac{\sigma_{ij} X_i^{[m]}/\rho}{1 + \sigma_i/\rho}, \quad X^{[m]} = S, I, R \quad (6.23)$$

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

$$\partial_t I = \lambda S(t) - \mu I(t), \quad \lambda = \beta \frac{I(t)}{N(t)} \quad (6.24)$$

Instead of explicitly modelling mobility, we now directly compute the effects of the other patches on the risk of infection, i.e. we break down the force of infection in its contributions.  $S_i$  is distributed among patch  $i$  and all possible destination  $j$  according to the proportions:

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

and the risk of infection has the form:

$$\lambda_i = \frac{\lambda_{ii}}{1 + \sigma_i/\rho} + \sum_j \frac{\lambda_{ij}\sigma_{ij}/\rho}{1 + \sigma_i/\rho} \quad (6.25)$$

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

$$\lambda_{ii} = \frac{\beta_i}{N_i^*} \left[ I_{ii} + \sum_j I_{ji} \right] = \frac{\beta_i}{N_i^*} \left[ \frac{I_i}{1 - \sigma_i/\rho} + \sum_j \frac{I_j \sigma_{ji}/\rho}{1 - \sigma_j/\rho} \right] \quad (6.26)$$

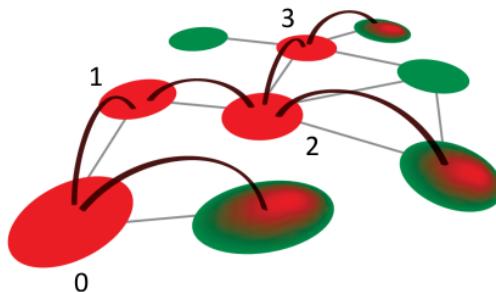
Whereas  $\lambda_{ij}$  susceptible resident in  $j$  that are travelling and so are exposed in other patch  $i$  to infected people:

$$\lambda_{ij} = \frac{\beta_j}{N_j^*} \left[ I_{jj} + \sum_l I_{lj} \right] = \frac{\beta_j}{N_j^*} \left[ \frac{I_j}{1 - \sigma_j/\rho} + \sum_l \frac{I_l \sigma_{li}/\rho}{1 - \sigma_l/\rho} \right] \quad (6.27)$$

Note as some of the terms are similar to the steady state ones in equation 6.22. These last expressions are actually useful in **numerical simulations**, and they allow us to understand the relative role of mobility and infection parameters on the epidemic dynamics in order to speed up the simulations, too. One should actually remember that both *infections* and *travels* are **stochastic processes**, and last results hold only when mobility is faster than the epidemic time scale.

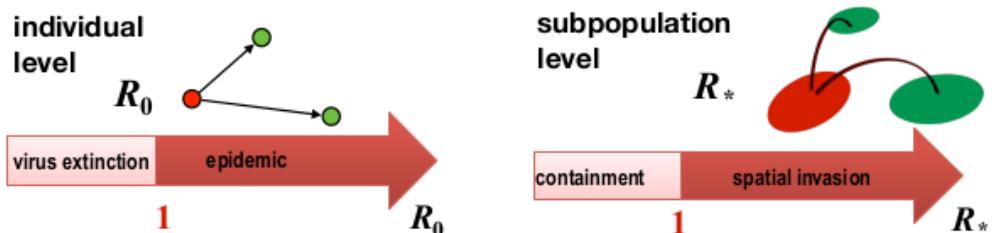
## 6.4 Global Invasion Threshold

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



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

In order to pursue our goal, we make the so called **coarse graining**: starting from the formalism typical of a single subpopulation we decrease the number of degrees of freedom by making some averages or finding some more general rules. Quantities must therefore be scaled accordingly:  $R_*$  is the correspondent to  $R_0$  for metapopulations (if  $R_* > 1$  we observe an outbreak), and the typical **timescale** is now the *duration of an outbreak* in a patch. Hence, we follow the spread from one subpopulation to another, by the mean of **mapping**<sup>13</sup> <sup>14</sup> the **spreading dynamics among subpopulation** into the spreading on a **network**.



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

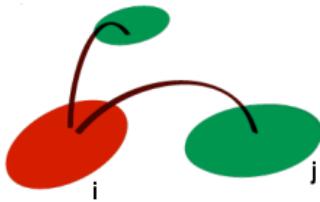
We are indeed approximating as a **branching process** and try to follow the invasion dynamics at the subpopulation level, denoting by  $D^n$  the diseased subpopulations at the  $n$ -th generation. Hence, we do not follow the dynamics in term of times, but in terms of *generation*.

### 6.4.1 Homogeneous networks

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

<sup>13</sup>Colizza & Vespignani, PRL 2007, JTB 2008

<sup>14</sup>Cross, et al. JRSoc Interface 2007



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

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

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

$$P_{ext} = \left( \frac{1}{R_0} \right)^{\lambda_{ij}} \quad (6.28)$$

where  $R_0$  is the usual term and  $\lambda_{ij}$  is the total number of infectious individuals sent from  $i$  to  $j$  during the local outbreak:

$$\lambda_{ij} = \frac{w_0}{\mathcal{N}} \frac{\alpha \mathcal{N}}{\mu} \quad (6.29)$$

Note that the *recovery rate*  $\mu$  is included as well and the first ratio is the *probability* to travel.

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

$$D^n = (\langle k \rangle - 1)(1 - P_{ext}) \left( 1 - \sum_{m=0}^{n-1} \frac{D^m}{V} \right) D^{n-1} \quad (6.30)$$

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

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

$$R_* = (\langle k \rangle - 1)(1 - P_{ext}) \quad (6.31)$$

Moreover, if  $R_0$  is really low<sup>15</sup>, the probability of having an outbreak is:

$$1 - P_{ext} = 1 - \left( \frac{1}{R_0} \right)^{\lambda_{ij}} \simeq \lambda_{ij}(R_0 - 1) = \frac{\alpha w_0}{\mu} (R_0 - 1) \quad (6.32)$$

---

<sup>15</sup>Actually this last approximation does not hold for COVID19, since it was able to spread given that its  $R_0$  was high.

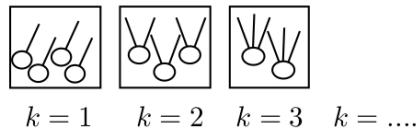
which actually simplifies our expression for  $R_*$ . Indeed, recalling that according to our assumption every patch has same number of connections and travellers, we obtain:

$$R_* = (\langle k \rangle - 1) \frac{\alpha w_0}{\mu} (R_0 - 1) \quad (6.33)$$

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

#### 6.4.2 Heterogeneous networks

Let us introduce a more realistic model and consider that networks are actually different from the homogeneous we have seen up to now. Indeed, **real systems** are **highly heterogeneous**. As we have already pointed out (see Sec. 6.1.1 and Fig. 6.1 and 6.2): the number of connections and travellers along the connections is heterogeneous<sup>16</sup>. This implies that average quantities, such as  $\langle k \rangle$ , are not representative of the properties of the patches: **homogeneous approximation is bad**. However, we were able to find some **scaling relations**: that are some approximate laws that make computations feasible and describe the **degree-block** description (coarse-graining).



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

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

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

$$N_k = N_0 k^\phi$$

Whereas the flux between two nodes  $k$  and  $k'$ :

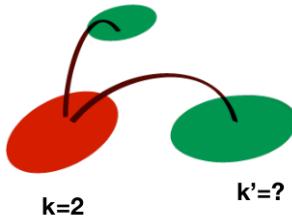
$$w_{kk'} = w_0 (kk')^\theta$$

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

$$p_{kk'} = \frac{w_0}{N_0} \frac{(kk')^\theta}{k^\phi} \quad (6.34)$$

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

<sup>16</sup>Colizza & Vespignani, PRL 2007, JTB 2008



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

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

$$D_k^n = \sum_{k'} D_{k'}^{n-1} (k' - 1) P(k'|k) \left( 1 - \sum_{m=0}^{n-1} \frac{D_{k'}^m}{V_{k'}} \right) (1 - P_{ext}(\lambda_{k'k})) \quad (6.35)$$

which looks like 6.30, except for the sum over all possible degrees  $k'$  that was introduced since we are dividing our patches in same-degree classes. Let us analyze every term: the  $k - 1$  is the number of mobility connections through which the seeding may potentially occur. This is multiplied by the probability that contact has degree  $k'$ , given we are starting from a node with degree  $k$ : explicitly *in random networks*  $P(k'|k) = k' \frac{P(k')}{\langle k \rangle}$ . In this step, we have assumed that the network is uncorrelated. This however implies that when we are making connections at random we are more likely to connect to hubs, leading to the so called *friendship paradox*, a.k.a. "rich gets richer", where nodes that have already many connections tend to accumulate more. These last two terms are given by the **topology** of the network. The second last term (red) is equal to the one previously obtained: it is indeed the probability that the contact patch belonging to  $k'$  class is disease-free. Last term (orange) is, as before, the probability for the epidemic to not get extinct before the global outbreak. One should note that, as previously done, this last term can be approximated to  $\lambda_{k'k}(R_0 - 1)$  assuming this is a *branching process*, and moreover for heterogeneous networks:

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

Finally, 6.35 can be rewritten as it follows:

$$D_k^n = (R_0 - 1) \frac{\alpha w_0}{\mu} \frac{k^{1+\theta} P(k)}{\langle k \rangle} \sum_{k'} D_{k'}^{n-1} (k' - 1) k'^\theta \quad (6.36)$$

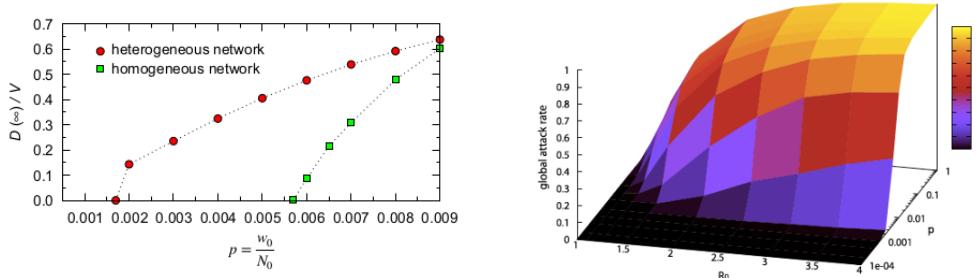
If we define the highlighted factor in red as  $\Theta^{n-1}$  for a more compact writing, we can multiply both rhs and lhs of Eq. 6.36 by  $\sum_k (k - 1) k^\theta$  thus obtaining:

$$\Theta^n = (R_0 - 1) \frac{\alpha w_0}{\mu} \frac{\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle}{\langle k \rangle} \Theta^{n-1} \quad (6.37)$$

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

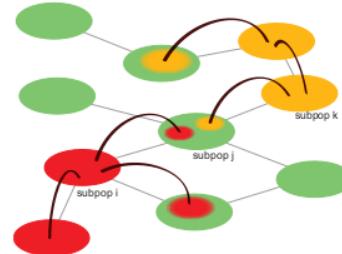
$$R_* = (R_0 - 1) \frac{\alpha w_0}{\mu} \frac{\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle}{\langle k \rangle} > 1 \quad (6.38)$$

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



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

## 6.5 Spatial spread of competing diseases

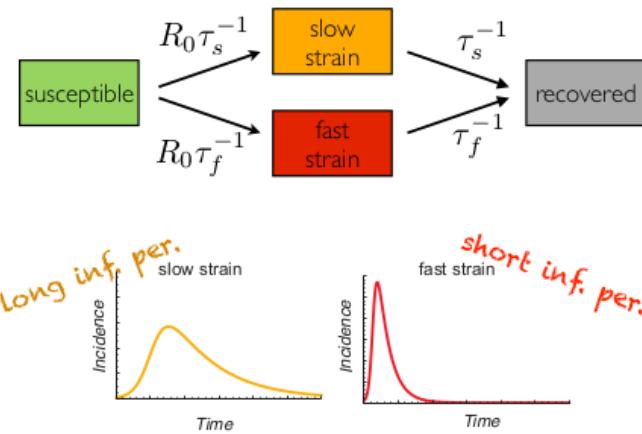


**Figure 6.19:** Two strains present in the same metapopulation network.

We will present now an application to understand better what we have explained so far: let us assume that there are two **competing pathogens** that are in the same metapopulation system. The **compartimental model**<sup>17</sup> that we obtain is the one in figure 6.20: there are two strains of a certain disease present, and one individual might be infected by either one of them. Once he recovers, he acquires the **full-cross immunity**, that is to say that he cannot be infected by both of them any more. The main difference between the two is that the infectious period of one is actually **faster** than the others ( $\tau_{slow} > \tau_{fast}$ ), whereas they share the **same  $R_0$** .

Since the two strains have actually different infectious periods, they lead to **different infectious dynamics**. Indeed,  $\tau$  affects both the slope of the *exponential growth*, as well as the *peak* and the presence of the disease at longer times. For instance, keeping  $R_0 \propto \beta/\mu$  **fixed**, if  $\tau$  is small we stay infected for less time but the

<sup>17</sup>Poletto, PLOS Comp Biol PRL 2007, JTB 2008.

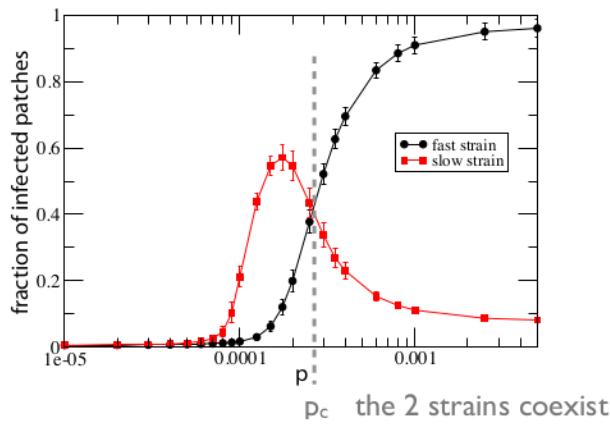
**Figure 6.20:**

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

transmissibility  $\beta$  must be actually higher, hence the spreading explodes. On the other hand, if  $\tau$  is large, we stay infected more and therefore the disease has more time to spread despite the lower  $\beta$ . At the end of the epidemic, we will have reached the same amount of population but ended up with very different dynamics.

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

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

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

At the end of simulation there might be a little chance for the two strains to coexist if they reach the same patch, however, more likely the fast strain will make the other die out when the *probability of travelling* is *higher*. On the other hand, if the *travelling probability* is *less* they even might not encounter each other, and therefore the slower strain is favored to last longer and still be present at the end of simulation time. We can find a **probability**  $p_c$  for which the two diseases coexist. One should note that, for higher traveling probability, we approach the *homogeneous*

mixing regime.

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

$$R_* = (\langle k \rangle - 1) \frac{\alpha p N}{\mu} (R_0 - 1) \quad (6.39)$$

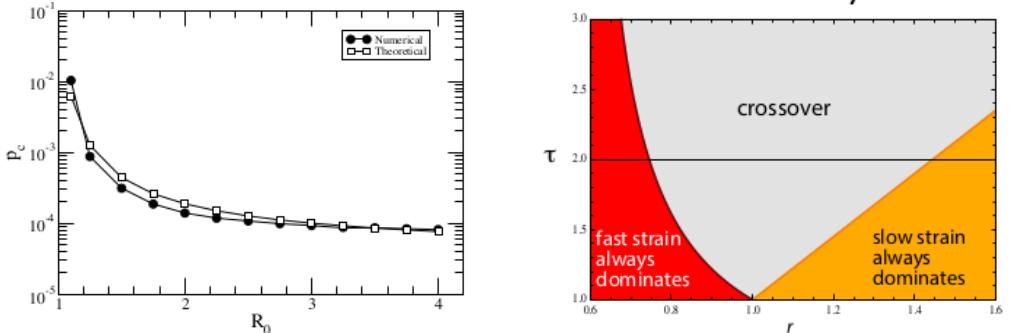
while the **number of diseased patches** is  $D(t) \sim e^{\frac{1}{T}(R_*-1)t}$ . One should note that, since we are assuming that all patches are the same, the outbreak duration  $T$  on average is the same for every patch.

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

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

$$\frac{D_s(t)}{D_f(t)} \sim e^{\left(\frac{R_*^s-1}{T_s} - \frac{R_*^f-1}{T_f}\right)t} = 1 \quad (6.40)$$

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



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

Now let us **drop** the assumption that  $R_0$  is **fixed**, and allow for different  $R_0$ <sup>18</sup>:

$$R_0^s = r R_0^f \quad (6.41)$$

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

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

$$G^s > G^f \implies R_*^s > R_*^f, \quad R_*^f > R_*^s \implies G^f > G^s \quad (6.42)$$

<sup>18</sup>Poletto et al. Sci Rep 2015.

that defines our colored areas where either one strain dominates (the one that comes with larger  $R_*$ ). However, if  $G^f > G^s$  and  $R_*^s > R_*^f$  we may observe **crossover**. More particularly:

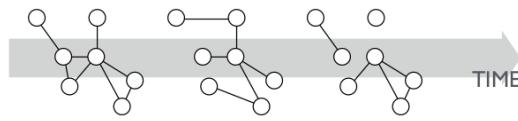
$$\begin{aligned} r > 1 \rightarrow \quad G^f > G^s \quad \Rightarrow \quad rR_0^f - 1 < \tau(R_0^f - 1) \\ r < 1 \rightarrow \quad R_*^s > R_*^f \quad \Rightarrow \quad \tau\alpha_s \log(R_0^s) > \alpha_f \log(R_0^f) \end{aligned}$$



# Temporal Networks

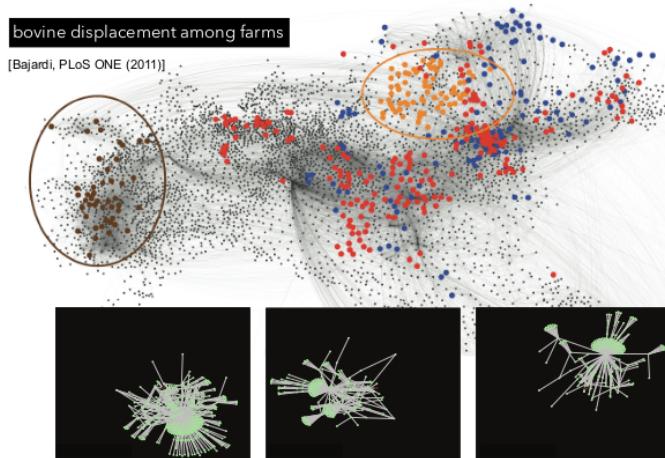
Let us discuss about networks that change structure in time (see Fig. 7.1). This kind of networks are called **temporal networks**. There are actually many of them: for instance let us consider the *face-to-face interactions* network. It is obviously a temporal network: nodes are indeed individuals and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

**Lecture 17.**  
 Thursday 26<sup>th</sup>  
 November, 2020.  
 Compiled: Tuesday  
 8<sup>th</sup> December,  
 2020.



[Holme, Saramaki Phys. Rep. (2012)]

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

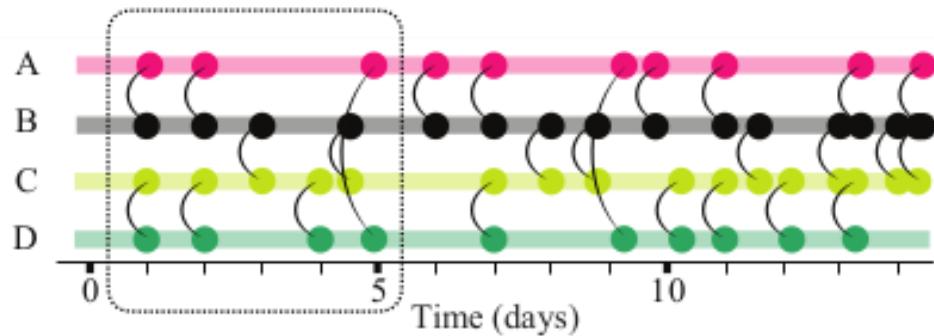


**Figure 7.2:** Bovine displacements among farms network.

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

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

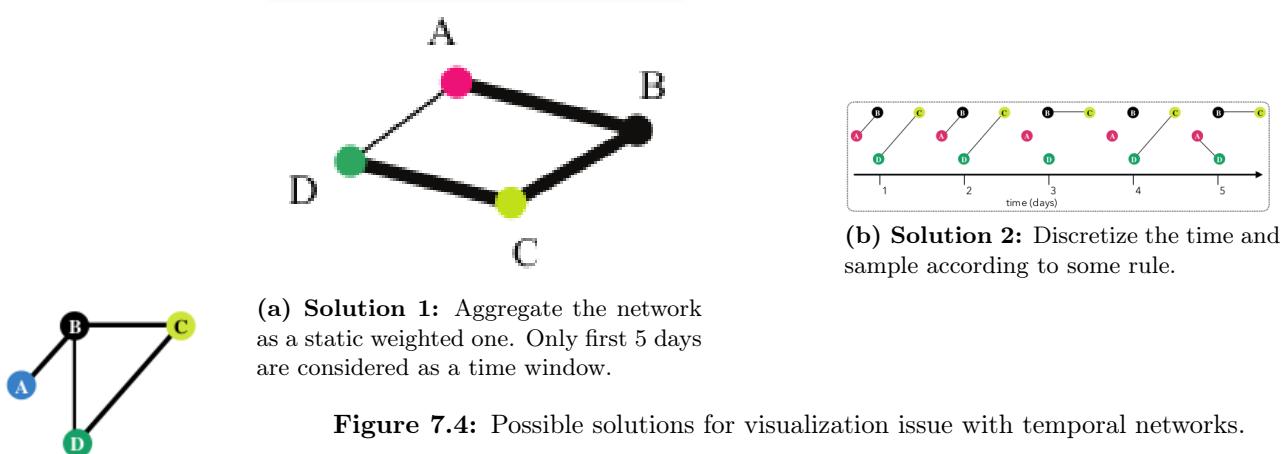
see how they change. However, the **sequence of links** is not a good representation to spot communities/topology: indeed we see what happens and at what moment, but in this way it is difficult to have a general overview. It allows us to preserve the whole information but, actually, it cannot be easily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



**Figure 7.3:** Example of temporal network.

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

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

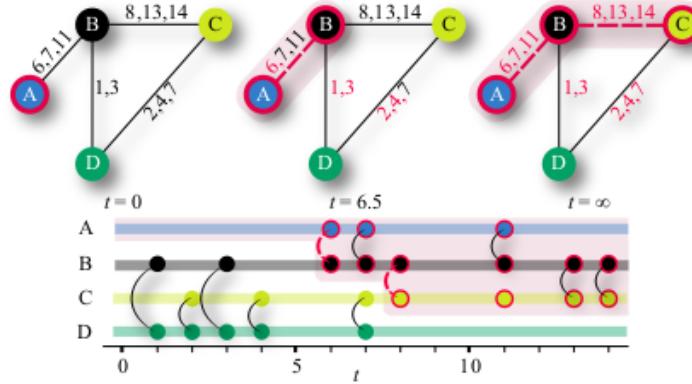


**Figure 7.4:** Possible solutions for visualization issue with temporal networks.

**Figure 7.5:** Reachability issue is not important in static networks.

The second **issue** one may face when dealing with temporal network is **reachability**. We say that  $i$  is **reachable** from  $j$  if it exists a path  $i \rightarrow j$ . This is indeed a very easy problem to tackle when dealing with *undirected static networks*: every node belonging to the same **connected component** is reachable from the other members. We recall now that we are losing much information: if, for instance, two nodes share a link only at the very beginning, we see them as connected even though they are not present any more after a certain time instant. It actually shows lots of path that might be **not** available any more. Hence, the **existence** of a *time respecting path* does depend on the window  $[t, T]$  of observation!

However, in an **undirected temporal network**,  $j$  is reachable from  $i$  only if there exist a **time respecting patch**  $i \rightarrow j$ . That is to say that there is a **sequence of contacts** that connects  $i \rightarrow j$  with each contact in the path coming sequentially one after the other from node  $i \rightarrow j$  (see Fig. 7.6).



**Figure 7.6:** The existence of temporal paths between two nodes depends on time.

- |  |  |
|--|--|
| <p>WO For <math>t = 6.5</math> there is a path from A to C<br/>WO For <math>t = 11.5</math> there is no path from A to C</p> <p>(a) Reachability depends on the choice of time window.</p> | <p>WO For <math>\mu^{-1} = 3</math> days YES<br/>For <math>\mu^{-1} = 1</math> days NO</p> <p>(b) Whether infection spreads depends both on the disease, in particular <math>\mu</math>, and on the time window.</p> |
|--|--|

**Figure 7.7**

One should recall that we are dealing with **epidemic spreading**: if we aggregate network into a static one, possibilities to infect other nodes are much larger rather than what happens for temporal network. Moreover, infection has a certain duration that is given by the **recovery rate**  $\mu$ : the longer it is, the more likely we stay infected and are able to infect other nodes spreading the disease. On the other hand, when infection duration is small we may recover without having the time to spread it to other nodes that at that moment are not connected to us.

The third **issue** that arises is the *activation frequency* or, in other words, the **contact heterogeneities**. In static network, we know that there are actually some nodes that have more contacts whereas some have few. The analogous for temporal networks is that, besides the *number of contacts* a node may have, we must consider how **often activation occurs**. That is to say: if a node activates often and has many contacts, for sure it will lead to a different spreading rather than a node that activates less often and has few contacts. It is also different in terms of epidemic spreading if, once fixed the number of contacts, we have them as "one-shot" (we go to a party and meet lot of people) or diluted wrt time (daily we meet few people): at the end the number of total contacts will be the same. Indeed, the **cumulative number of contacts** results from *activation frequency* and *number of contacts per activation*.

Another **issue** we have to tackle is the **non homogeneous activation**, that is to say that **inter-contact** times (i.e. *inter-occurrence* time) between activations is not exponential and the number of contacts cannot be modeled as Poisson. Therefore, individuals do not have some sort of general *activation rate*: it is a more complex process, since according to an exponential distribution longer periods are not allowed (see fig.7.8). Indeed, empirical data suggests us that human behavior is **bursty**,

and this *burstiness* is reflected in broader-than-expected distribution of inter-contact times. Some datasets, for example face-to-face interactions, emails, or phone calls suggested that the best model that reflects human social activity is a power law with a cutoff:

$$P_E(\tau) = A\tau^{-\alpha}e^{-\tau/\tau_E}$$

Indeed, we usually stay inactive for a while, and then have a burst: we start reply to just received messages and eventually start a conversation (*causality effect*).



**Figure 7.8:** Inter-contact times can be modeled according different distributions, which lead to very different behaviors.

The last **issue** is the most complex one to be captured by models and involves **temporal correlations**. Every time a node activates, there is the chance that the **contacts** it makes are **correlated** to the contacts it had in the **past**. A possible workaround is to introduce the so called **social strategy**<sup>1</sup>:

$$\gamma_{i,t} = \frac{k_{i,t}}{s_{i,t}} \quad (7.1)$$

where  $k_{i,t}$  and  $s_{i,t}$  are respectively the *degree* and the *weighted degree* of  $i$  in the network aggregated over the interval  $[t - \delta, t]$ . For  $\gamma \rightarrow 0$  we have a *memory-driven behavior*, so a node tends to make contacts always with the same nodes (social keeper), while for  $\gamma \rightarrow 1$  the behavior is memoryless, therefore a node shows a more socially exploratory behavior.

## 7.1 Temporal networks dynamics

Let us consider now, once we have a dataset with temporal description of the network, how its main **properties affect the dynamics**. Up to some years ago, temporal networks were not considered of much interest: for instance indeed under some assumption, namely the **time scales separation**, the time dimension seemed not to be relevant in the spreading process and therefore was dropped.

Let us refer to the **average infectious duration** as  $\mu^{-1}$ , while to the **average contact time**  $\tau$ . For  $\tau \ll \mu^{-1}$  we do not need to take into account contacts with a high resolution in time, being the network faster than the disease: considering the average network would be actually sufficient. Conversely, if the network is slower than the disease (e.g. migration network), namely  $\tau \gg \mu^{-1}$ , one may want to exploit the static network and drop the time information. We want to understand when  $\tau \sim \mu^{-1}$ , so when **timescales are comparable**. This can be done obviously if and only if **timescales are definite**: there might be some distributions whose mean is not informative, since its variance is high and a timescale cannot be defined out of them.

<sup>1</sup>Miritello, et al, Sci Rep 2013.

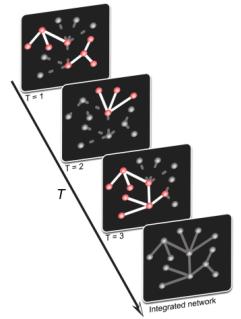
We can follow two main approaches when dealing with temporal networks in epidemiology: **bottom-up** perspective and **top-down** approaches. In the first one, we start *building mathematical models* for human interactions and the spreading dynamics. However, this is not at all a simple approach and requires too much work. The second approach is to take some *empirical networks* and *run numerical simulations* over them, by randomising (i.e. neglecting/throwing away) some properties whose we want to study the effects. In this way, we are able to understand the *impact of a specific property* in the spreading process.

### 7.1.1 Activity driven model

It is the first and the simplest model<sup>2</sup> that was introduced to study human interactions, where nodes activate according to a certain frequency.

The main **ingredients** for this model are:

- $\Delta t$  timestep, indeed time is discretized;
- $N$ : number of nodes;
- $x_i$ : activity potential, that is the number of activations of  $i$  during  $\Delta t$  and is normalized over the total number of activations ( $\varepsilon \leq x_i \leq 1$ );
- $F(x)$ : distribution of the activity potential;
- $a_i = \eta x_i$  activation rate, where  $\eta$  is a rescaling factor chosen to tune the average number of active nodes per unit time in the system  $\bar{N} = \eta \langle x \rangle N$ ;
- $m$ : number of connections made at each activation, it is the same for all the nodes.



**Figure 7.9:** Activity driven model.

The **algorithm** for the model is indeed the following:

- at each time step  $t$  the network  $G_t$  starts with  $N$  nodes that are *all* disconnected;
- node  $i$  activates with probability  $a_i \Delta t$  and makes  $m$  links with other randomly selected nodes. Non-active nodes are still able to receive connections from active nodes;
- at the next time step  $t + \Delta t$  all the edges are deleted, so all links last  $\tau_i = \Delta t$ .

This model actually captures the issue introduced when talking about **heterogeneity in activations**: now we are indeed modelling Poisson activations, despite the rate of activation is heterogeneous. In addition, it does not take into account possible correlations between activations. Moreover, *network at each timestep has on average  $E_t = m\eta \langle x \rangle N$  edges* and  $\langle k \rangle_t = \frac{2E_t}{N} = 2m\eta \langle x \rangle$  **average degree**. Network is indeed **homogeneous**<sup>3</sup> wrt its topology!

Now we want to understand how properties change when we **integrate the network over a time window  $T$** . The **degree** of a node  $i$ , in the aggregated network is:

$$k_T(i) = k_T^{OUT}(i) + k_T^{IN}(i) \quad (7.2)$$

Let us focus on the first term  $k_T^{OUT}(i)$ :  $i$  makes  $T a_i m$  links. We want to understand how many different nodes  $i$  connects to, without counting twice repeated links with the same node. This should resemble the *Urn problem*, where we want to count the number of *different* balls extracted from an urn with  $N$  balls after  $T a_i m$  extractions.

<sup>2</sup>Perra et al, Sci Rep 2012.

<sup>3</sup> $P(k)$  is Poisson

We know that the probability for each ball (node) to be extracted at least once is  $p = 1 - [1 - 1/N]^{Ta_{im}}$  and the probability of extracting  $d$  balls is binomial with parameters  $\text{Bin}(p, N, d)$ . Hence, the average number of balls  $k_T^{OUT}$  is:

$$\langle k_T^{OUT} \rangle = Np = N[1 - e^{-\frac{Ta_{im}}{N}}] \quad (7.3)$$

as  $N \rightarrow \infty$  and the time window is such that  $T/N \rightarrow 0$ .

Let us focus on the other term  $k_T^{IN}(i)$ , that is the number of nodes that make connections with  $i$  among those that were not targeted by  $i$ . In this way we avoid double counting. The probability that a node was not a target of  $i$  is the complementary as before, that is:  $[1 - 1/N]^{Ta_{im}} \sim e^{-\frac{Ta_{im}}{N}}$ . Given that the average number of edges formed at each timestep is  $mN \langle a \rangle$  and they can connect to  $i$  with probability  $1/N$ , we have that:

$$\langle k_T^{IN} \rangle = m \langle a \rangle e^{-\frac{Ta_{im}}{N}} \quad (7.4)$$

Therefore, the degree formula can be written as function of each activity potential  $a_i$ :

$$k_T(i) = N[1 - e^{-\frac{Ta_{im}}{N}}] + m \langle a \rangle e^{-\frac{Ta_{im}}{N}} \simeq N[1 - e^{-\frac{Ta_{im}}{N}}] = N[1 - e^{-\frac{T\eta x_i m}{N}}] \quad (7.5)$$

where we assumed that  $N \rightarrow \infty$  and  $T/N \rightarrow 0$ .

Rewriting now the **activity potential** as a function of the *degree*  $x(k)$ :

$$x(k) = -\frac{N}{\eta m T} \ln \left( 1 - \frac{k}{N} \right)$$

We can revert both the latter formula and  $P_T(k)dk \sim F(x)dx$ , thus obtaining **degree distribution** of the aggregated network for observation in window of length  $T$ :

$$P_T(k) \sim F[x(k)] \frac{dx(k)}{dk} = \frac{1}{Tm\eta} \frac{1}{1 - k/N} F \left[ -\frac{N}{\eta m T} \ln \left( 1 - \frac{k}{N} \right) \right]$$

If the time window dimension is small, namely  $T \rightarrow 0$ , then also  $k/N \rightarrow 0$ , and the latter expression can be further approximated as:

$$P_T(k) \sim \frac{1}{Tm\eta} F \left[ \frac{k}{Tm\eta} \right] \quad (7.6)$$

This implies that **nodes activate and form a heterogeneous network** despite they activate heterogeneously. In other words, heterogenous topology in the aggregated network (i.e. there are hubs), over a window  $T$ , results from a heterogeneous activity potential (i.e. they activate more often).

Let us try to understand what are the **effects** on the network dynamics on epidemic spreading. For instance, let us consider how the **epidemic threshold** changes according to the activation dynamics. In order to pursue our goal, let us use the **activity block approximation** that works as the same way as the *degree block approximation*, and consider a *SIR* model. Let us define the *probability of transmission per contact* as  $\beta$ , moreover let us assume  $m = 1$ : every time a node activates, it makes a single connection. The *SIR* equation, classifying nodes according their activity, for infected nodes at time  $t + \Delta t$  within class  $a$  is:

$$I_a^{t+\Delta t} = -\mu \Delta t I_a^t + I_a^t + \beta(N_a^t - I_a^t) a \Delta t \int da' \frac{I_{a'}^t}{N} + \beta(N_a^t - I_a^t) \int da' \frac{I_{a'}^t a' \Delta t}{N} \quad (7.7)$$

The *green* term describes the probability for a node in class  $a$  to activate and get in contact with infected nodes of any other classes  $a'$ , from which it contracts the disease.

The blue term, on the other hand, returns the probability we have to be infected by other infectious nodes that instead activate while we are not. Defining the last integral  $\theta^t = \int da' \frac{I_a^t a' \Delta t}{N}$ , we are able to write the **total number of infectious** as:

$$\int da I_a^{t+\Delta t} = I^{t+\Delta t} = I^t - \mu \Delta t I^t + \beta \langle a \rangle I^t \Delta t + \beta \theta^t \Delta t \quad (7.8)$$

Multiplying both rhs and lhs of the last equation by  $a$  and integrating over the latter, we obtain:

$$\theta^{t+\Delta t} = \theta^t - \mu \theta^t \Delta t + \beta \langle a^2 \rangle I^t \Delta t + \beta \langle a \rangle \theta^t \Delta t \quad (7.9)$$

Up so far, we obtained two equations in two variables  $I$  and  $\theta$ . Rewriting them as differential equations:

$$\partial_t I = -\mu I + \beta \langle a \rangle I + \beta \theta \quad (7.10a)$$

$$\partial_t \theta = -\mu \theta + \beta \langle a^2 \rangle I + \beta \langle a \rangle \theta \quad (7.10b)$$

Let us understand now when these expressions return as a growth in the number of infectious. The tool we will use is **linear stability analysis**. The Jacobian is:

$$J = \begin{vmatrix} -\mu + \beta \langle a \rangle & \beta \\ \beta \langle a^2 \rangle & -\mu + \beta \langle a \rangle \end{vmatrix} \quad (7.11)$$

whose set of eigenvalues is  $\Lambda_{1,2} = \beta \langle a \rangle - \mu \pm \beta \sqrt{\langle a^2 \rangle}$ . We want the *largest eigenvalue* to be **positive**, in order to have the number of infectious growing.

The condition for the *largest eigenvalue* to be *positive* is thus:

$$\frac{\beta}{\mu} > \frac{1}{\langle a \rangle + \sqrt{\langle a^2 \rangle}} + \mathcal{O}\left(\frac{1}{N}\right) \quad (7.12)$$

and it set a threshold for  $\beta$  that is function of the moments of the activity distribution. Note as if activity distribution is heterogeneous, then its variance becomes large and the threshold decreases, hence favoring the spread.

Summarizing, the **activity driven model** captures the realistic property of human behavior (face-to-face, sexual contacts, phone call, email, tweets) and takes into account heterogeneous activity rate. Moreover, the contact network at a **given instant** is *sparse* and has homogeneous degree. Once we want to stress that the **aggregated network** over a certain window is well connected with *heterogeneous degree* networks!

So if we do not make any assumption on the pattern of activation that may unfold at the same time scale of the spreading process, computations are made possible within the *activity-block* approximation, which follows the same scheme as the degree-block approximation. In conclusion, **contact heterogeneity lowers the epidemic threshold**.

### 7.1.2 Randomised Reference Models

Let us now discuss an example of the other type of approaches one may want to use when dealing with *temporal networks*: **top-down** approaches<sup>4</sup>. Let us recall we want to understand how the **temporal structure** of the network **impacts** the **spreading**. We therefore compare the epidemics on real data with the outcome in suitable null models that randomize (i.e. destroy) some properties over some others. However, we keep the topology of the aggregated network fixed.

We are going to study 3 different types of randomizations. Let us define the following quantities:

---

<sup>4</sup>Gauvin et al Sci Rep 2013.

- $P(\tau)$ : inter-contact time distribution;
- $\omega_{AB}$ : cumulated contact durations of an arbitrary link;
- $P(\omega)$ : distribution of the cumulated contacts duration;
- $n_{ab}$ : number of contacts per link of an arbitrary link. These are the total times A, B are in contact, regardless of how much;
- $P(n)$  distribution of the number of contacts per link.

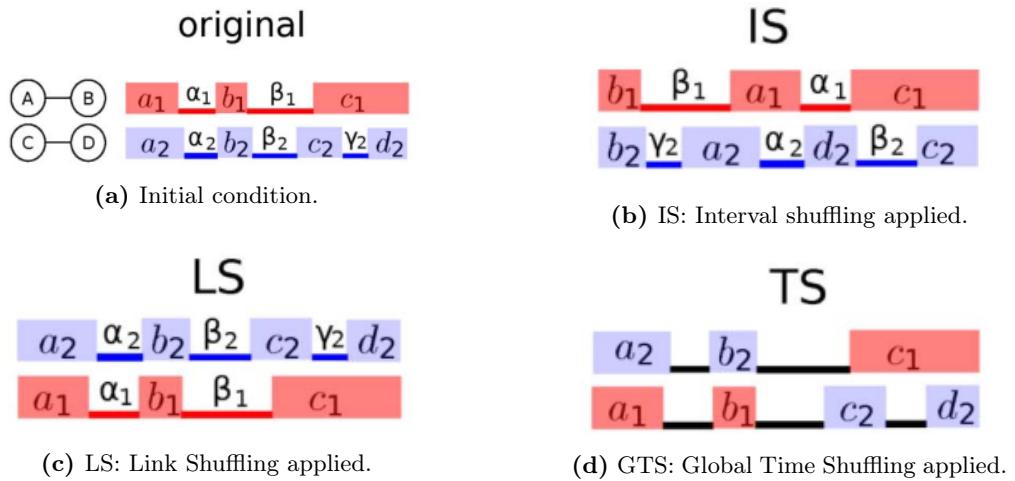


Figure 7.10

Let us discuss the three different of shuffling (see Fig. 7.10) we may have:

- **Interval Shuffling**: the sequences of contact and inter-contact durations are reshuffled for each link separately. The only property we destroy is the *causality*: correlations between historical information of link are destroyed, since their chronological order (sequence) is randomized.
- **Link Shuffling**: the unaltered sequence of events (i.e. contacts) are swapped between link pairs. In this case we are destroying *causality*: by reshuffling links we destroy causal correlations between pair nodes. Moreover, we are also dropping  $\omega_{AB}$  and  $n_{AB}$ , since we are assigning to each link a history that has been taken randomly from other nodes.
- **Global Time Shuffling**: build a global list of the contact durations. For each link, we generate a synthetic activity timeline by sampling with replacement the global list according to the original number of contacts for that link. We are preserving, by construction, the topology, the number of connections between each pair and their distribution. On the other hand, we are destroying all other ones *causality*,  $P(\omega)$ ,  $\omega_{AB}$ ,  $P(\tau)$  since now we are drawing randomly activation times.

See Fig. 7.11 in which the main shuffling characteristics are summarized.

Comparing the empirical data and the other to the ones we obtain by using the aforementioned shuffles, we are able to understand what are the **main features** of our network. In addition, we can see how these either enhance or contrast the spreading of a disease. For instance, if we destroy a property and we note that spread does not strongly change, we are even allowed to not collect data related to that property any more, thus saving time, memory and resources. On the other hand, if after shuffling it strongly changes, then we can conclude that that property plays an important role in the spread process.

RRM	Topology	Causality	$P(\tau)$	$\omega_{AB}$	$P(\omega)$	$n_{AB}$	$P(n)$
<b>IS</b>	✓	✗	✓	✓	✓	✓	✓
<b>LS</b>	✓	✗	✓	✗	✓	✗	✓
<b>TS</b>	✓	✗	✗	✗	✗	✓	✓

**Figure 7.11:** Table depicting all different types of randomizations and what property we drop.

