Exercise 02 - Determining the spreading of epidemics

Federico Agostini, Federico Bottaro, Gianmarco Pompeo

1 Introduction

Computational epidemiology is a scientific field where the spread of diseases and the effectiveness of public health interventions are studied by means of multidisciplinary techniques overlapping Computer Science, Big Data analysis and Network theory.
To simulate such studies, we are given two undirected networks representing two different contact networks among individuals: we will study them by applying three increasingly more sophisticated approximations (Mean Field, Heterogeneous Mean Field and Quenched Mean Field) within the SIS (Susceptible-Infected-Susceptible) model.
Moreover, we are given a dataset describing the early stages of an outbreak of Chicken Pox in 100 different locations, which we will analyze this time applying the SIR model (Susceptible-Infected-Recovered) so as to forecast the outcome of this contagion.

2 Epidemic threshold from the networks

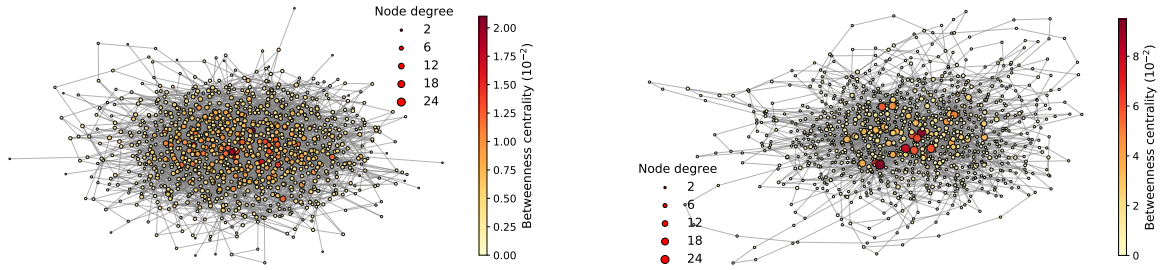
In the SIS framework, two reactions can occur:

$$I \xrightarrow{\mu} S \tag{1}$$

mediates the recovery from the plague of an infected individual I, which becomes susceptible S with a rate μ;

$$I + S \xrightarrow{\lambda} I + I \tag{2}$$

is the reaction describing the contagion of a susceptible when in proximity (network-wise, when linked) of an infected, happening with rate λ.
We have at disposal the following two graphs representing a population of $N_{nodes} = 1000$ people.



(A) *Network A* **(B)** *Network B*

Figure 1

We want to compute the theoretical value of the *epidemic threshold* λ_c, smallest value of the infectivity rate needed for the diseases to undergo an outbreak, using three different approximations:

- **Mean Field (MF)**, where all individuals are assumed to be perfectly equivalent, with the same number of average contacts in a closed population with no demographics; here we have

$$\lambda_c^{(mf)} = \frac{\mu}{\langle k \rangle} \tag{3}$$

where $\langle k \rangle$ is the average degree of the network;

- **Heterogeneous Mean Field (HMF)**, where we add more complexity by sub-dividing individuals in classes according to their node degree k, under the same assumptions as before; here we

$$\lambda_c^{(hmf)} = \mu \frac{\langle k \rangle}{\langle k^2 \rangle} \tag{4}$$

where we also have a dependency on the second moment $\langle k^2 \rangle$ of the degree distribution;

- **Quenched Mean Field (QMF)**, an individual-based approach capable of giving information on specific nodes; one finds

$$\lambda_c^{(qmf)} = \frac{\mu}{\Lambda_{max}(\mathcal{A})} \tag{5}$$

where $\Lambda_{max}(\mathcal{A})$ represents the highest eigenvalue of the adjacency matrix $\mathcal{A}$ for each network.

We set the recovery rate $\mu = 0.5$ and using the provided formulae we computed the thresholds for the three approximations.

	MF	**HMF**	**QMF**
Network A	0.097	0.082	0.080
Network B	0.139	0.076	0.066

Table 1: *Comparison of the values of λ_c for the two networks in the three approximations studied.*

3 Simulation of SIS model

We now wish to simulate the phase diagram for λ for both networks within the SIS model: to do so we will implement a Gillespie algorithm with some variations.
First off, we once again set $\mu = 0.5$ and we will fix the initial number of infected individuals $N_I^0 = 50$: we decide not to modify this value, but we will choose the nodes that are infected at $t = 0$ randomly so as to create variety.
Then for each network, according to the algorithm, we fix the propensity rates

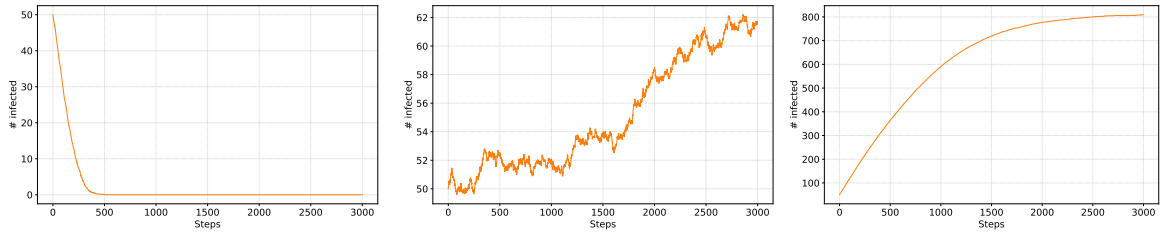
$$a_1 = \mu I \tag{6}$$

for Reaction (1) and we try two possibilities for Reaction (2),

$$a_2 = \frac{\lambda}{2} \sum_{i,j} \varepsilon_{ij} \tag{7}$$

with $\varepsilon_{ij} = 1$ if i and j are linked together and belonging to two different groups, $\varepsilon_{ij} = 0$ otherwise. We also tried using $\hat{a}_2 = \lambda I \frac{S}{N}$, but this propensity rate is an approximated one; in fact we have seen how it is not able to reproduce the expected outbreak of the epidemics for any given λ (discarded plots can be found in the `JN` - Jupyter Notebook).
According to the Gillespie algorithm core principle, a reaction must occur at every time step; the rates we just defined will be used to weight the random selection of which reaction is going to take place. If Reaction (1) is chosen, a random node among the infected ones is selected and it undergoes recovery. If

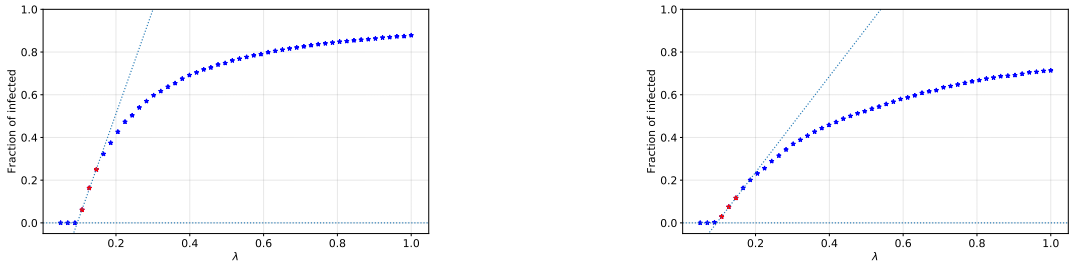
Reaction (2) is instead picked, we again select an infected individual at random but this time we need to make sure that it is connected to at least one susceptible; if that is the case, one randomly-chosen S is turned into an infected, otherwise another I needs to be found again in a random fashion for the reaction to occur.
For every chosen λ, one simulation consists of 3,000 steps after which the systems are updated; we run a total of 100 simulations per λ value and for each step the number of infected individuals gets averaged between the corresponding others. We therefore end up with 3,000 mean values and we once again take the average of the last 500 of them, which we assume is a way to smooth out possible fluctuations happening when the system has reached a stationary regime.
As a first interesting result, we show the dynamics of the epidemics in both networks for three values of λ: a sub-critical one, which causes the disease to die out, a more or less critical value showing great fluctuations and a value for which the outbreak takes place.



(A) $\lambda < \lambda_c$ (B) $\lambda \approx \lambda_c$ (C) $\lambda > \lambda_c$

Figure 2: *Behaviours at different regimes with respect to criticality.*

To build the phase diagram, we just run the whole machinery described above for a list of 50 λ values evenly chosen between 0 and 1 and we plot the density $\rho = \frac{I}{N}$ as a function of the parameter.



(A) *Network A* (B) *Network B*

Figure 3: *Phase diagrams for the two networks, also showing the linear interpolations performed to estimate* λ_c.

The result clearly shows the trend we were expecting for the SIS model: for big enough values, a steep increase right after criticality is followed by stationarity when the disease is at peak.
To have a numerical estimation of $\lambda_c^{(net)}$, we take the first 3 values for which we have a non-trivial fraction of infected and we fit them linearly; then we set the value of $\lambda_c^{(net)}$ as the intersection with the x-axis, since all the preceding values are null. This is an ansatz we make considering that for values of λ that are right above the threshold the approximation appears reasonable.
We find

$$\lambda_c^{(A)} = 0.0951 \pm 0.0074 \qquad \lambda_c^{(B)} = 0.0949 \pm 0.0045$$

Surprisingly, if we compare these numerical values with the theoretical ones obtained in Section 2, we see that the closest estimations are those obtained by the Mean Field model in its basic form.

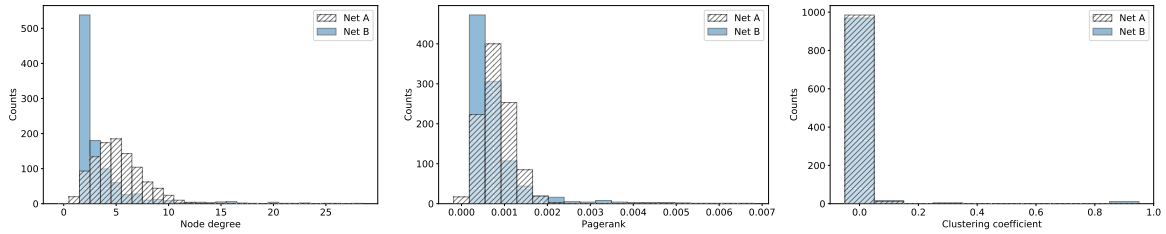
We were not quite expecting such a result, considering that the QMF model is the one that adds the highest degree of complexity and it is therefore expected to be able to replicate real world behaviour more closely.
However, we do not have context as to how these networks were created nor if they represent anything in particular at all, which could be an indication of this unexpected result. From it, we hypothesize that these networks might represent a close group of people whose connections are distributed homogeneously.
It is to be noted, anyways, that also the estimations provided by the MF model are not faithful; in particular, the critical value of λ is highly overestimated for network B.

4 Study of network properties

In this Section we aim at characterizing the networks and understanding how their structural properties may influence the dynamics of the epidemics spreading across them.
First off, we compute the connectivity $\mathcal{C}$ for both network, that is the probability that two given nodes are linked together, $\mathcal{C} = \frac{L}{\binom{N}{2}}$ (where L is the total number of links); we have

$$\mathcal{C}_A = 0.00515 \qquad \mathcal{C}_B = 0.00359 \tag{8}$$

We also show in a graphic way the distributions of other key parameters.



(A) *Degree distribution* (B) *PageRank* (C) *Clustering coefficients*

Figure 4: *Distribution of the studied parameters for both networks.*

From the degree distribution we can see that Network A is more heterogeneous, while Network B seems distributed as a power law: this will have consequences in the effectiveness of different segregation strategies (see Section 5).
As far as the PageRank is concerned, this quantity measures the importance of each node by means of how many connections it has and we can see that the shape is quite similar to the degree distribution. This fact is not surprising because both networks are undirected.
Those differences are reflected in the phase diagrams, not too much on the value of λ_c (which we have seen to be more or less the same) but on the "velocity" of the spreading after the critical point. In fact the fraction of infected in Network A goes up faster than in Network B due to the difference in the degree distribution and the connectivity (higher in the first network).
The clustering coefficient seems instead to play a minor role in this scenario: the two networks show here a pretty much identical distribution, which is strongly peaked in 0. This can be explained by the fact that such graphs do not show tendency to cluster: once the pathogen starts to spread, there is no reason as to why infected or susceptibles should group together. A low value of this coefficient means that the transitivity property does not hold in these systems, which makes sense if we take into account that, given any two infected individuals, they can also recover from the disease with probability μ instead of just passing it around to a third.

5 Impact of segregation

The separation of some critical individuals for special treatment or observation from the group they belong to is known as *segregation* and it represents a possible strategy to slow down the spreading of a disease or prevent it from happening altogether.
We need to perform such segregation with the constraint that we can only remove 1% of the individuals, that is 10 nodes from each network. Clearly, to make the segregation procedure as efficient as possible it seems highly intuitive that the nodes with the highest degree will have to be removed: this will prevent the individuals with the largest number of connections to spread around the disease excessively, if infected.
However, another aspect to keep into consideration is that there might be people who do not have quite as many links but are found in a "strategic" position in the network: for example, those nodes that might create a disconnection between two regions of the network itself if they happen to be removed. This aspect can be accounted for by the betweeness centrality, which quantifies how many shortest paths go through a node, hence it is a way of detecting the amount of influence a node has over the flow (of the virus, in our case) in a graph.
To ponder the influence of these two aspects, for each network we simulated the SIS model after removing: (a) the 10 nodes with the highest degree; (b) the 10 nodes with the highest betweeness centrality (c) the 5 nodes with highest degree and the 5 with highest betweeness.
We compute again the critical values and see how effective each strategy has been.

	Original λ_c	Strategy (a)	Strategy (b)	Strategy (c)
Network A	0.0951 ± 0.0074	0.097 ± 0.012	0.096 ± 0.013	0.096 ± 0.014
Network B	0.0949 ± 0.0045	0.109 ± 0.019	0.098 ± 0.010	0.108 ± 0.008

Table 2: *Values of λ_c obtained in the SIS model for the three segregation procedures (the original λ_c is also included for comparison.*

At first glance, we can see that all the values of λ_c increase, meaning the epidemic will outbreak at a higher threshold, which is exactly what segregation is meant to do.
In particular, Network A, which is more homogeneous, only has its values increased by about only 2%; here, removing the nodes with the highest degree has proved to be most effective strategy. In Network B, on the other hand, λ_c increases by $\approx 15\%$ at most: this graph is more subjective to segregation and in particular, since it is distributed as a power law, removing the nodes with the highest betweeness has proven to be the best solution, as the network suffers more from a strategic removal of nodes.
Overall, deciding what the best segregation strategy is highly depends on the network we need to apply it to, just as the degree of its effectiveness.

6 Application of SIR model: chickenpox case study

We are given data of the early stages of an outburst of chickenpox in 100 different locations and we use these 100 sets as independent and identical distributed experiments. This time, we wish to implement the SIR (Susceptible-Infected-Recovered) model, which in comparison with SIS allows for a third stage of the epidemic: when an individual is recovered, it means it is either immune to the disease of it has perished from it.
The goal is to extract a unique value of the parameter R_0, known as *reproduction number* and defined as

$$R_0 = \frac{\beta}{\gamma} \tag{9}$$

where β is the infection rate and γ is the recovery rate: a value $R_0 > 1$ means that the infection will spread among the population.
The differential equations that describe the SIR model are the following

$$\frac{dx}{dt} = \beta s(t)x(t) - \gamma x(t) \qquad \frac{ds}{dt} = -\beta s(t)x(t) \qquad \frac{dr}{dt} = \gamma x(t) \tag{10}$$

where $x(t), s(t), r(t)$ are respectively the fraction of infected, susceptible and recovered of the population as a function of time t. We get the analytical solutions

$$\frac{s(t) - s(0)}{\int_0^t s(\tau)x(\tau)d\tau} = -\beta \qquad \frac{r(t)}{\int_0^t x(\tau)d\tau} = \gamma \tag{11}$$

and here we replace the integrals at the denominator with the sum over all the days since we have a discrete time-series; such summations need to be computed over all the data-taking period, so $s(t)$ and $r(t)$ represent the value corresponding to the last day and $s(0)$ will be the value measured at day 1.
In Fig. 5 we plot the behaviour of these quantities at the early stage of the pathogen spreading.

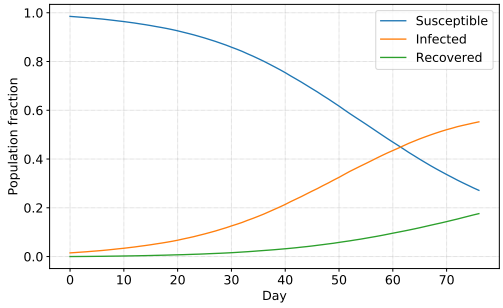
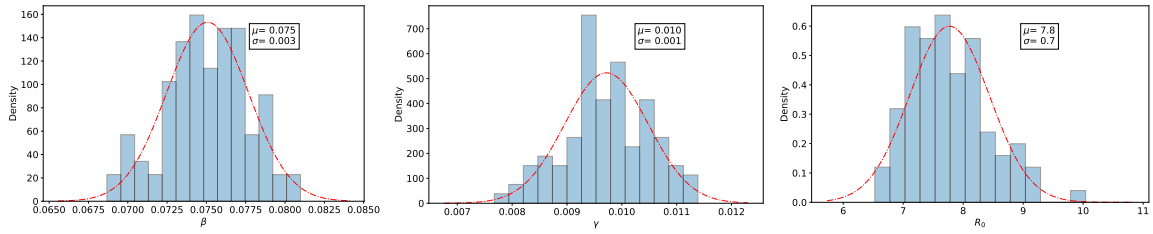


Figure 5: *Mean fraction of infected, susceptible and recovered through the days.*

For each location we estimate a value of β, γ and R_0 and we then build an histogram for each quantity, overlaying it with the fitting Gaussian.



(A) *Histogram of* β **(B)** *Histogram of* γ **(C)** *Histogram of* R_0

Figure 6: *Histograms of the parameters of SIR model with corresponding Gaussian fit.*

We find

$$\beta = 0.075 \pm 0.003 \qquad \gamma = 0.010 \pm 0.001 \qquad R_0 = 7.8 \pm 0.7$$

To conclude this Section, from the value of R_0 obtained we can infer that the chickenpox will become viral, which makes sense considering that the infection rate is bigger than the recovery rate.