UNIVERSITÉ CATHOLIQUE DE LOUVAIN

MASTER THESIS MAP

Diffusion on Social Networks with Application in Disease Spreading

Promoter: Pr. Jean-Charles.Delvenne

Readers: Pr.Philippe Chevalier and Dr.Martin Gueuning Student:

Sibo Cheng

2016-2017

Acknowledgement

First of all, would like to express my deepest gratitude to my master thesis supervisor Professor Jean-Charle Delevenne for having proposed this interesting topic and most importantly for his continuous support and kindness during the whole year.

I would like to thank Martin Gueuning for kindly sharing with me his research experience in similar topics.

Thanks to my friend Kaitong Hu for fruitful discussions.

At last but definitely not at least, my sincere thanks go to my parents: my father Professor Caiyuan Cheng, my mother Doctor Yi Zheng for their advice in both academic and life during my six years of study in Europe.

Sibo Cheng
2017.6.8

Chapter 0: General introduction

"In examining disease, we gain wisdom about anatomy and physiology and biology. In examining the person with disease, we gain wisdom about life."
— Oliver Sacks

In the last five centuries have seen more new diseases than ever before become potential pandemics. With the growing of economics and unavoidable globalisation, people never travel across countries and continents as frequent as today. In fact, in the last 50 years, the passenger numbers of airfreight has grown almost 9% per year [Upham et al. 2003]. Similarly, shipping traffic has also increased over 27 % since 1993 [Zachcial and Heideloff 2003].The efficiency,speed and reach of modern transport networks puts people at risk from the emergence of new strains of familiar diseases, or from completely new disease[Guimera et al.,2005]

As we know, variety of diseases posses different regularities of spreading for example through air (cold, influenza), sexual contact (AIDS) etc . In fact, different ways of spread and different infect objects may lead to a significant difference on the disease spreading which makes it necessary to understand mathematical modellings of different spreading approaches.

Human has already suffered countless disease crisis in our long history. Plague which created the famous black death killed 30 % to 60% of population in Eurasia. It is certainly an important event that has changed the human history. The HIV virus, since it is discovered in 1970s, has become one of the most sensitive and worrying topics. In fact during the past 50 years the number of infected of HIV has always been increasing. Unfortunately, until today a cure treatment for this lethal virus hasn't been found yet. Another recent example is the Ebola outbreak in West African since 2013. This killer virus claimed more than 11 thousand lives in only 14 months. In 2016, the World Health Organization finally declared the end of the outbreak of Ebola virus disease in West Africa.

The control of disease spreading is indeed and will always be one of the most difficult challenges for human beings.
We present two figures below for the infected number evolution curve of two infectious disease that have been just mentioned: HIV and Ebola.

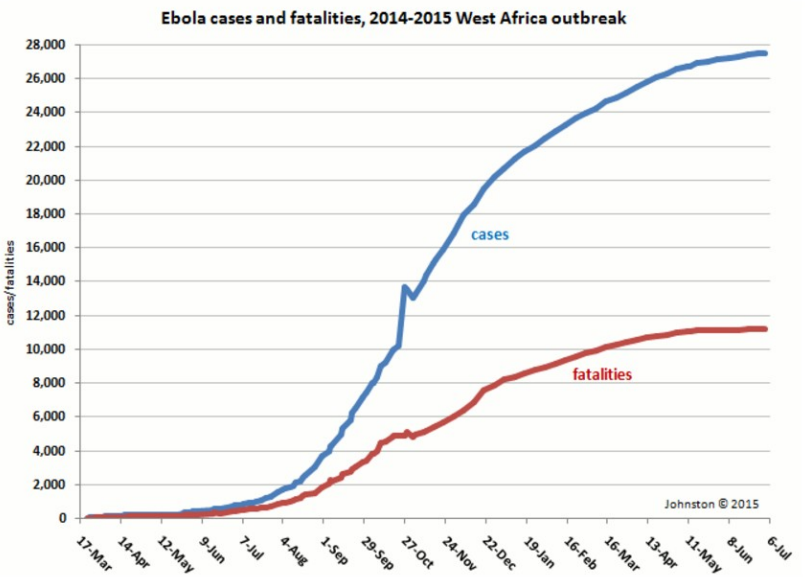


Figure 1: evolution of Ebola disease in 2014-2015, taken from MARYN MCKENNA Science

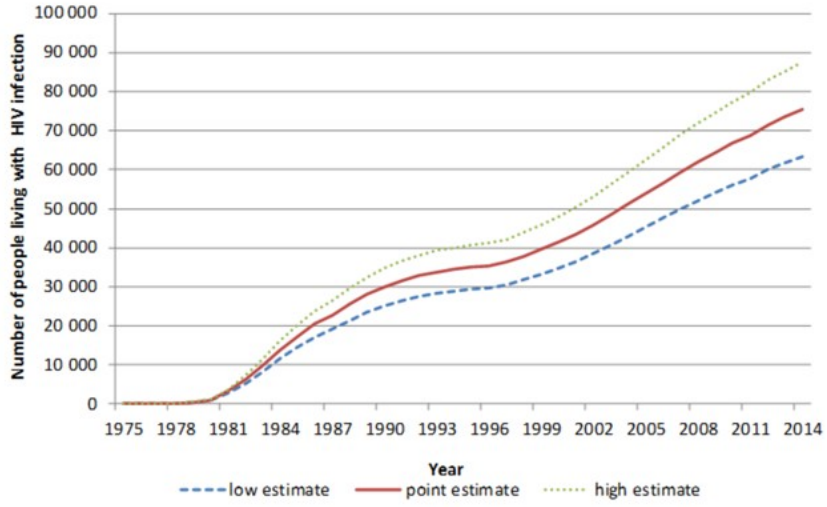


Figure 2: evolution of HIV since 1975,taken from Descriptive Epidemiology

In Figure 1, we observe an exponential increase for Ebola virus in one year. However it disappeared just one year later in 2016. In fact, according to the data given by the World Health Organization the number of Ebola infected doubled every 29 days at the beginning of its outbreak period. Meanwhile, the number of HIV infected has been increasing slowly but stably in the past 40 years. **Why there exists such a significant difference in terms of spreading behavior for different diseases?**

Obviously, there are medical reasons behind but mathematical theories play also a very important role! Diseases spread with different personal contact which construct various mathematical modellings. Therefore, the potential infection network and the probability of contamination could be totally different. In this thesis, we aim

to explain these spreading behaviors using mathematic tools.
The thesis is split into three main chapters.

- Chapter 1 Micro-modelling: In this chapter, we put ourselves into the shoes of one individual in the whole network. The goal is to check how his or her social contact behavior could influence the infection of disease. For example, some friends may meet each other very regularly, say once a week while some others can meet much variously like sometimes several times a week and than loose contact of a couple of weeks. Even in the case where the average numbers of contact are the same for both types of contact, their probability of infect each other could still be quite different. We define a new type of stochastic dominance which is different from the classic ones.
- Chapter 2: Macro-modelling: In chapter 2, we use a more global vision for network diffusion. We are actually interested in the structure of social networks and its consequence on disease spreading. The forms of infection networks literally depend on the types of contacts that may contaminate the disease. For example the virus spread by sexual contact such as HIV will surely not share the same contact network with diseases that could spread through air such as influenza. The key quantity in this chapter is the reproductive ratio. We also define two different models using Erdos-Renyi graphes to simulate the spread of disease based on a given structure of contact matrix.Two real data set have also been used for the simulation.
- Chapter 3: Optimal vaccination strategy. Following the topic of chapter 2, we suppose the social contact network is perfectly known. We also make the assumption that we are only capable to vaccinate a part of the whole population. Therefore, we look for the most efficient measure to decide who should be vaccinated. The goal is to prevent the disease outbreak. We define a new quantity named as the transition capability measure based on the similarity of nodes in the network. We compare it with other measures such as degree and betweenness centrality on different contact networks.

Chapter 1: Micro-modelling in probability distributions between nodes

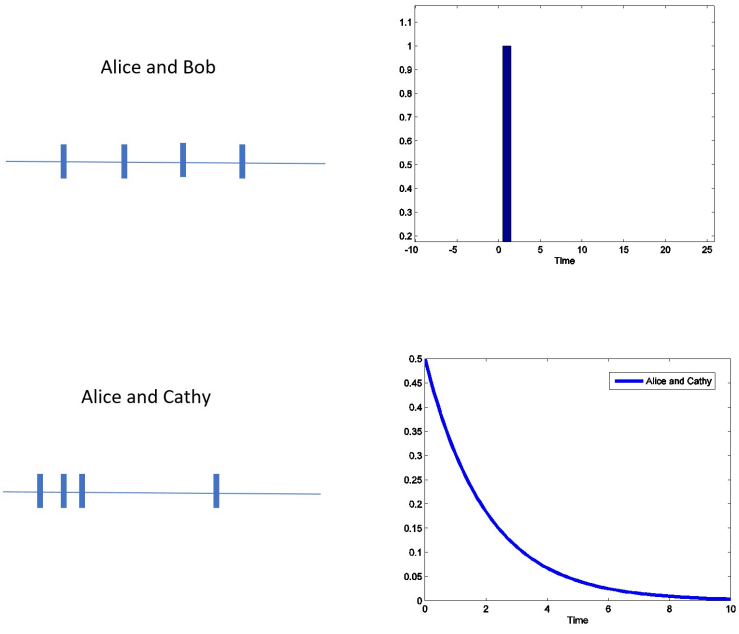
1 Summary

In this chapter, we focus on the individuals' behavior , more precisely the regularity of their social contacts and how this could speed up or slow down the diffusion on a social network. We concentrate a lot on theoretical analysis in chapter 1, two relations of dominance have been defined. We also develop several proprieties around these two new definitions. The chapter is based on a very simple idea : we would like to see which intercontact time distribution can promote the random diffusion. Unlike the chapter 2 and chapter 3, the work in this chapter doesn't depend that much on former articles. To make a clear vision, all the definitions and proprieties developed in this thesis are encircled by red borders.

2 Introduction

In reality, the disease spreading between two persons is a random phenomena. Obviously, the spreading behavior depends a lot not only on the average frequency of contact among people in the social network but also the exact distribution of meeting times. Generally speaking, two persons who meet each other in a regular way will not have the same chance to infect each other as two persons who meet intensively during a small period and stop their contact suddenly.

An example of message transmission has been illustrated as follows where we suppose that the transition time between Alice and Bob is a constant while the transition time between meetings of Alice and Cathy is simulated by an exponential distribution. Although we suppose the average transition time remains always the same, we find that Cathy has more chance in average to receive the message earlier than Bob.



In fact,

$$E(T_{Alice,Bob}) = E(T_{Alice,Cathy}) = E(exp(1)) = 1$$

However, as $T_{Alice,Cathy} \sim exp(1)$

$$P(T_{Alice,Cathy} < T_{Alice,Bob}) = P(T_{Alice,Cathy} < 1) = \int_0^1 e^{-x} dx = 1 - e^{-1} > \frac{1}{2}$$

This result shows that in average Cathy has more chance to receive the message in the first place.

In this chapter, we focus on the influence of different probability distributions of transition time between nodes in the network on the speed and behavior of disease spreading.
The further study will focus on two different models which are denoted as "transition time model" and "meeting time model". As explained in the previous example, in "transition time model" one suppose that once a person received the message/be infected by virus, a transition time will be simulated immediately following certain probability distributions. While in "meeting time model", one suppose the frequency of meeting has been pre-simulated when one individual receives the message/is infected one will have to wait for the next waiting time in the process to transfer the message/spread the disease. To illustrate these two models, some examples are presented as below.

3 Basic examples

In this section, to make readers understand better these two models, some examples have been shown in both cases in a simplified network of three individuals: Alice, Bob and Cathy. It is supposed that one letter will be transferred among them where the transition time between each pair is simulated according to different probability distributions independently.The main objective is to evaluate the accumulated time of each person holding the letter. The more "holding time" a person owns , the more he/she will be considered as the "center" of his/her social network.

3.1 "Transition time model"

As explained previously, in this model, the transition time will be simulated simply when one individual receives the message. Therefore, the transition time is independent on the past of the process. If one node owns two choices of diffusion, it will choose the direction where the time of passage is the shortest.

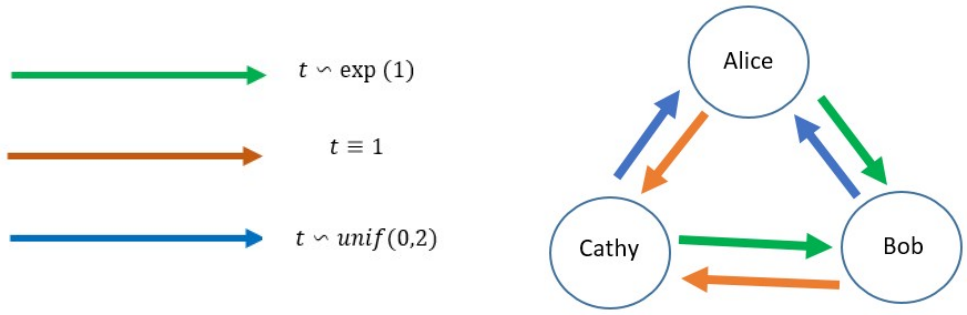


Figure 3: example: triangle of transition time model

For example, when Alice receive the message, two variables will be simulated immediately $X_1 \sim exp(1)$ and $X_2 \equiv 1$ where X_1 present the potential transition time between Alice and Bob while X_2 present the potential tran-

sition time between Alice and Cathy. Suppose $X_1 < X_2$, in this case Alice will transfer the message to Bob. X_1 will be recorded as real transition time and X_2 will simply be dropped.

3.2 "Meeting time model"

Different from the transition time model, here we consider that people(nodes) see each other with certain frequency which is independent on the diffusion in network. The frequency is independent to the message delivery. Using the example in 3, we suppose Alice hold the message at beginning. She will then wait for next time that she meets Bob and then transfer the letter to Bob. Bob will then give it to Cathy and Cathy will give it back to Bob as presented in figure **??**. The transition time in this model will be the rest time between the moment that one receive the letter and his/her next meet with somebody.

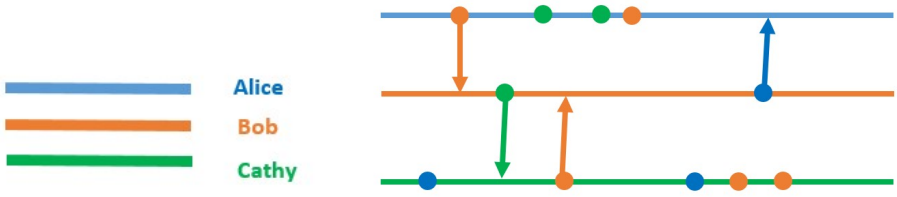


Figure 4: example: triangle of meeting time model

4 Analytical approach

In this section, we define a new stochastic order named "equal dominance" and compare it with the classic definition of stochastic dominance introduced in [J.Hadar, W.Russell,1969]. Several proprieties of this new order is developed and some numerical experiments are made in this section as well. Typically,we generate variables following different probability distributions with same expectation and check the influence they make on the behavior of spread.

Definition 4.1. *We define in this paper that a non-negative random variable X is "equally dominated" by another non-negative variable Y if and only if*

$$\begin{cases} E(X) = E(Y) < \infty \\ P(X > Y) > 0.5 \end{cases}$$

Generally speaking, when variable X is "equally dominated" by Y that means despite the two non-negative variables own the same expectation, the variable Y has more chance to be smaller than X.

This definition looks much like the definition of second-order stochastic dominance.

Definition 4.2.

Varibale X has first-order stochastic dominance over Variable Y means that

$$\forall r \in \mathbb{R}, \quad F_X(r) \leq F_Y(r)$$

$$\exists r \in \mathbb{R}, \quad F_X(r) < F_Y(r)$$

X is said to be smaller than Y in the 2nd-order stochastic dominance with equal means, denoted as $X \preceq_{2nd-SD}, = Y, when$

$$\begin{cases} E(X) = E(Y) \\ E[(t-X)+] \geq E[(t-Y)+], \forall t \in \mathbb{R} \end{cases}$$

In fact, they are two different definitions. The property of stochastic dominance can not imply the "equal dominance".
To prove that it is sufficient to give two counter-examples.

- For two discrete variables X and Y presented as below:

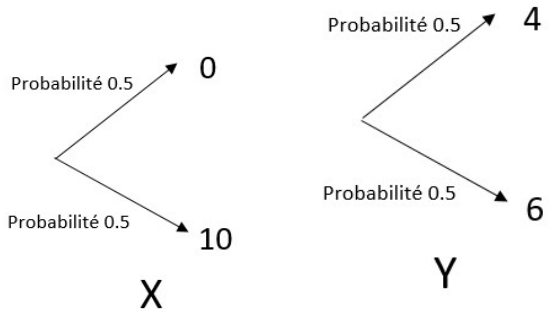


Figure 5: discrete variables

$$P(X = 0) = P(X = 10) = 0.5$$

$$P(Y = 6) = P(Y = 4) = 0.5$$

It is easy to check that the variable X is smaller than Y in the 2nd-order stochastic dominance with equal means.
In fact $E(X) = E(Y)$ and when $t < 6$ one has obviously $E[(t-X)+] \geq E[(t-Y)+]$ as $\forall t, (t-X)+ \geq (t-Y)+$
When $6 < t < 10$,

$$E[(t-X)+] = 0.5t \quad \text{while} \quad E[(t-Y)+] = 0.5(t-4) + 0.5(t-6) = t-5$$

$$\text{thus} \quad E[(t-X)+] > E[(t-Y)+] \quad \text{when} \quad 6 < t < 10$$

It is also quite obvious to prove $P(X > Y) = 0.5$ since $P(X > Y|X = 10) = 1$ and $P(X > Y|X = 0) = 0$. Therefore, X is not "equally dominated" by Y.

- Suppose $X \equiv 1$ and $Y \sim exp(1)$, we have already proved that X is "equally dominated" by Y. However, E[(1-X)+]=0<E[(t-Y)+] ;

Property 4.1.

The "equal dominance" does not own the transitivity, which means there exist a paper -scissors-rock case, ie three different variables X,Y,Z following different probability distributions such that

$$\begin{cases} X \xrightarrow[dominant]{} Y \\ Y \xrightarrow[dominant]{} Z \\ Z \xrightarrow[dominant]{} X \end{cases}$$

Therefore, the "equal dominance" is not transitive.

Proof. It is sufficient to give an example of a paper -scissors-rock case. We have found one as below: Suppose X, Y, Z are three discrete random variables with:

$$\begin{cases} X = & 0.7 & \text{with probability} & 0.4 \\ & 1.2 & \text{with probability} & 0.6 \end{cases}$$

$$\{Y = \quad 1 \quad \text{with probability} \quad 1$$

$$\begin{cases} Z = & 0.8 & \text{with probability} & 0.7 \\ & \frac{22}{15} & \text{with probability} & 0.3 \end{cases}$$

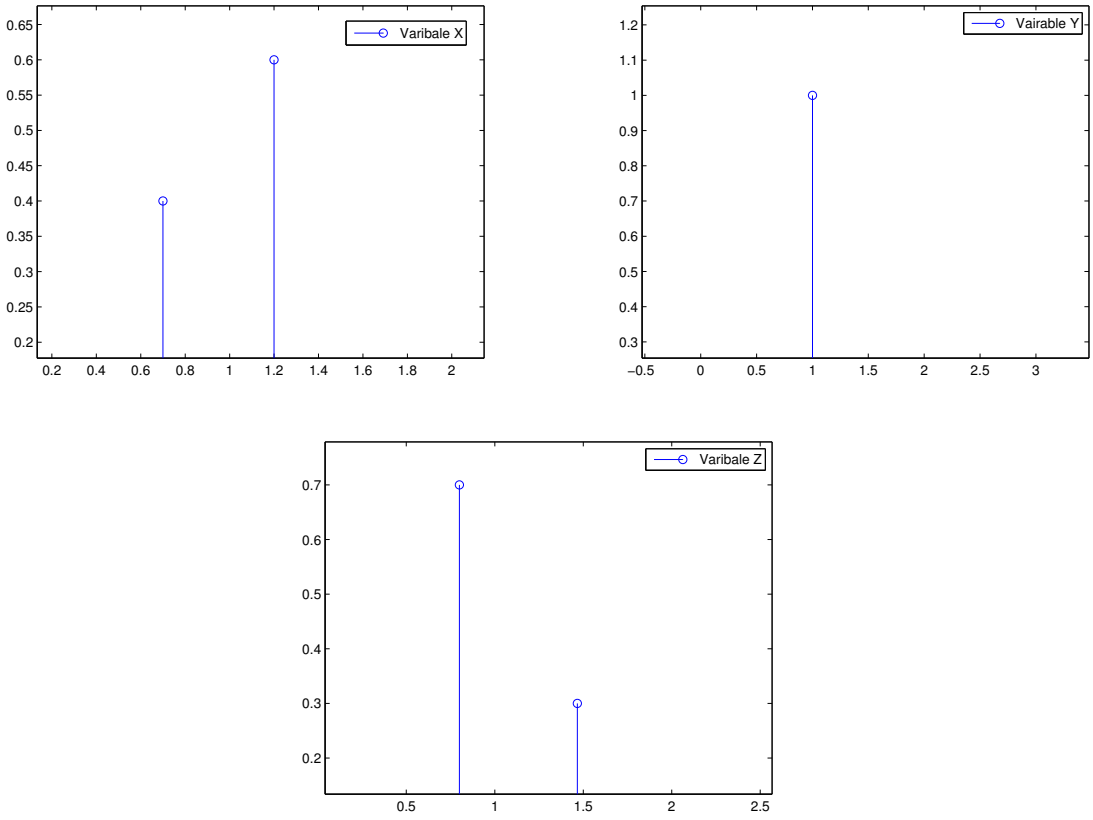


Figure 6: counter-example

It is easy to check that the expectations of three variables are all equal to 1.

$$P(Y > X) = 0.6$$
$$P(Z > Y) = 0.7$$
$$P(X > Z) = 0.4 + 0.6 \times 0.3 = 0.58$$

Thus we have successfully found a counter-example. Actually the existence of this paper -scissors-rock case also proves that the "equal dominance" can not be simply measured by any quantity measures like variance, moments etc otherwise this measure should be transitive.

□

4.1 "transition time model"

In this model, all the potential transition time is simulated independently at the same time following its probability distribution which means the edges only appear when its emission node is active (in our case receive the message or be infected by virus). The transition time is defined as the time period from the moment that the current node receives the message until the first transition takes place.

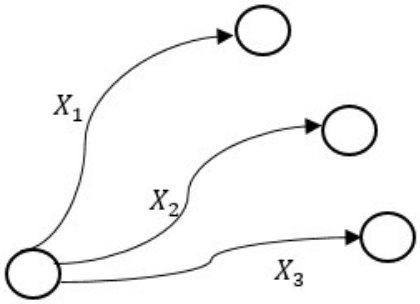


Figure 7: potential transitions

$$T = min(X_1, X_2, ..., X_n) \quad \text{with} \quad E(X_1) = E(X_2) = ...E(X_n)$$

4.1.1 case of two potential transitions

We start with the simple case of two potential transition time X_1 and X_2, denote $T = min(X_1, X_2)$.

Starting by choosing four common probability distribution as presented below, $X_1...X_4$ are four variables that being generated respectively by these four distributions. We are interested in the expectations of minimum values in each pair $E(min(X_i, X_j))$ and the probability $P(X_i > X_j)$. These two values may indicate the importance of each distribution has in the network in terms of accelerating/slowing down the diffusion.
In terms of density function, one can observe from the table below that despite the supports of density function of both exponential and Pareto distributions are infinite, the Pareto density function decreases much slower than exponential in long term.

Variable	probability distribution	expectation	variance	density function
X_1	constant:1	1	0	$P(X_1 = 1) = 1$
X_2	exp(1)	1	1	$f_{X_2}(x) = exp(-x)$
X_3	unif(0,2)	1	$\frac{1}{3}$	$f_{X_3}(x) = 0.5$ for $x \in [0,2]$
X_4	Pareto (1.25,0.2)	1	$+\infty$	$f_{X_4}(x) = \frac{1.25 \times 0.2^{1.25}}{x^{2.25}}$

Table 1: presentation of four random variables

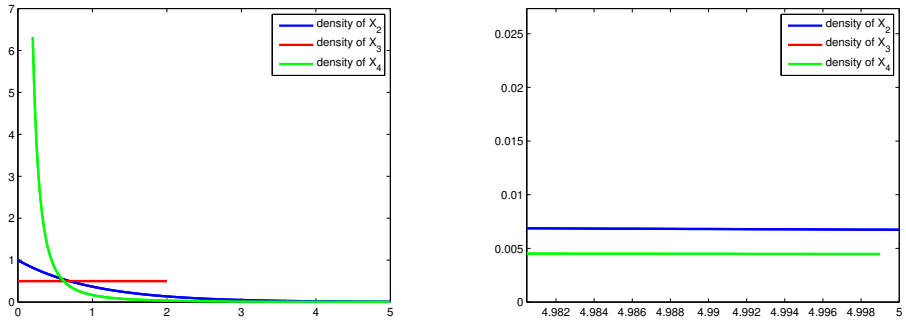


Figure 8: density functions

The estimation of $E(T_{i,j}) = E(min(X_i, X_j))$ is made by a Monte Carlo calculation with $n = 100000$. Ideally, this table should be symmetric.

distribution	constant:1	exp(1)	unif(0,2)	Pareto (1.25,0.2)
constant:1	1.0000	0.6322	0.7500	0.4655
exp(1)	0.6329	0.4991	0.5713	0.3761
unif(0,2)	0.7490	0.5658	0.6665	0.4172
Pareto (1.25,0.2)	0.4654	0.3782	0.4166	0.3326

Table 2: $E(T)$ for each pairs of variables

The probability every variable being chosen as minimum in each pair of variables has also been recorded.

$$P(i,j) = Proba(X_i < X_j) = P(X_i \leq X_j) \quad \text{as} \quad \text{all four distributions are continuous}$$

distribution	constant:1	exp(1)	unif(0,2)	Pareto (1.25,0.2)
constant:1	0.5	0.3681	0.4988	0.1333
exp(1)	0.6317	0.5018	0.5644	0.3796
unif(0,2)	0.4990	0.4340	0.5028	0.2779
Pareto (1.25,0.2)	0.8652	0.6236	0.7244	0.4976

Table 3: $P(X_i \geq X_j))$ for each pairs of variables

We say a probability distribution i is more favourable than the distribution j in terms of accelerating the diffusion if $P(X_i < X_j)) > 0.5$ (ie.X_i "equally dominate" X_j) and for most of other distributions Y, $E(min(X_i, Y)) \leq$

$E(min(X_j, Y))$

From tables 2 and 5, it is easy to observe among these four variables an order of dominance for four distributions in terms of accelerating the diffusion.

$$Pareto(1.25, 0.2) \xrightarrow[\text{dominant}]{} exp(1) \xrightarrow[\text{dominant}]{} unif(0,2) \xrightarrow[\text{dominant}]{} constant1$$

These dominance could also be proved mathematically. In fact, as the column presented with green color in Table 1, this dominance order actually follows the same order of variance. One may expect by intuition that when the variance of a variable is big, it has more chance to be smaller than the other variables which may provide some advantage in this competition. However the analytical study below shows the variance is not the only determinant factor of the "equal dominance" order.

We are now going to find what is the main factor that decides the "equal dominance order". Start with the case of two independent variables X, Y with same expectation $E(X) = E(Y)$ we remind that $T = min(X, Y)$. The goal of the calculation below is to find the distribution funcion of variable T.

$$1 - F_T(t) = P(min(X_1, X_2) > t) = P(x_1 > t, x_2 > t) = 1 - P(x_1 \leq t) - P(x_2 \leq t) + P(x_1 \leq t, x_2 \leq t)$$

X_1 and X_2 are independant, therefore,

$$F_T(t) = F_{x_1}(t) + F_{x_2}(t) - F_{x_1}(t)F_{x_2}(t)$$

We are specially interested in the expectation value of T.

$$E(T) = E(X|X < Y)P(X < Y) + E(Y|Y < X)P(Y < X)$$

In fact,

$$E(X|X < Y) = \frac{E(X \cap X < Y)}{P(X < Y)} = \frac{\int_0^\infty tF_X(t)P(Y \geq t)dt}{P(X < Y)}$$

Thus

$$E(X|X < Y)P(X < Y) = \int_0^\infty tf_X(t)(1 - F_Y(t))dt = E(X) - \int_0^\infty tf_X(t)F_Y(t)dt$$

As the same,

$$E(Y|Y < X)P(Y < X) = E(Y) - \int_0^\infty tf_Y(t)F_X(t)dt$$

Finally,

$$E(T) = E(X) + E(Y) - \int_0^\infty tf_X(t)F_Y(t)dt - \int_0^\infty tf_Y(t)F_X(t)dt \qquad (*)$$

Since $E(X) = E(Y)$ is prefixed, to minimize $E(T)$ is in fact equivalent to maximize $\int_0^\infty tf_X(t)F_Y(t)dt + \int_0^\infty tf_Y(t)F_X(t)dt$. From now on, without losing any generality we suppose that

$$E(X) = E(Y) = \int_0^\infty tf_X(t)dt = \int_0^\infty tf_Y(t)dt = 1$$

If we suppose further more that X and Y are two independent variables which follow the same probability distribution.ie.

$$f_X(t) = f_Y(t) \quad F_X(t) = F_Y(t)$$

The objective function therefore become simple:

$$\max(\int_0^\infty tf_X(t)F_X(t)dt) = \max(\int_0^\infty tf_X(t)\int_0^t f_X(s)dsdt) = \max(\lim_{T\to+\infty}\int_0^T tf_X(t)\int_0^t f_X(s)dsdt)$$

Property 4.2. *In the case that X and Y are identically distributed with $E(X) = E(Y) = 1$ we have*

$$sup(E(T)) = 1 \quad and \quad inf(E(T)) = 0$$

Proof. Let's start with studying the problem when T is a fixed real value.

$$\lim_{T\to+\infty}\int_0^T tf_X(t)\int_0^t f_X(s)dsdt = \lim_{T\to+\infty}\int_0^T t(F_X(t))'F_X(t)dsdt = [tF_X(t)^2]_0^T - \int_0^T f_X(s)F_X(s)dt - \int_0^T F_X(t)^2dt$$

Thus one can easily deduce,

$$2\int_0^T f_X(s)F_X(s)dt = [tF_X(t)^2]_0^T - \int_0^T F_X(t)^2dt$$

As when $T \to +\infty$, $F_X(T) \to 1$, for T big enough, we have

$$2\int_0^T f_X(s)F_X(s)dt \simeq T - \int_0^T F_X(t)^2dt \qquad (**)$$

Once again, we have just transformed the problem from $\max(\int_0^\infty tf_X(t)F_X(t)dt)$ to $\min(\int_0^T F_X(t)^2dt)$ always under the condition of $\int_0^\infty tf_X(t)dt = 1$
We then establish another equality which stands for all the real value T

$$\int_0^\infty t(F_X(t))'dt = [tF_X(t)]_0^T - \int_0^T F_X(t)dt$$

Which is equivalent to

$$\int_0^T F_X(t)dt = [tF_X(t)]_0^T - \int_0^\infty t(F_X(t))'dt$$

when $T \to +\infty$, we have

$$\int_0^T F_X(t)d \simeq T - 1$$

$$Furthermore, \quad F_X(t) < 1 \quad \forall t \quad \Longrightarrow \int_0^T (F_X(t))^2dt \leq \int_0^T F_X(t)d \simeq T - 1$$

With (**) one can easily deduce:

$$2\int_0^T f_X(s)F_X(s)dt \geq T - (T-1) = 1 \Longrightarrow \int_0^T f_X(s)F_X(s)dt \geq \frac{1}{2}$$

Thus using the result of (*), one could have directly $E(T) \leq 1$. Actually this upper bound is attained when $X = 1$ is a constant variable. □

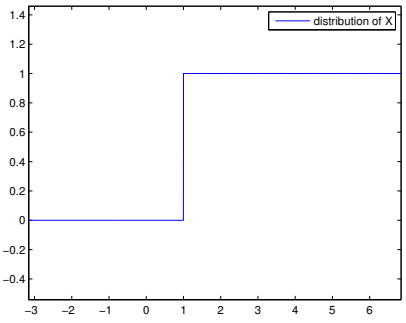


Figure 9: optimal distribution when maximize E(T)

One can easily check for when $X \equiv 1$, the following equality does stand:

$$\int_0^T (F_X(t))^2 dt = T - 1 \quad \forall T > 1$$

Therefore, we have proved $sup(E(T)) = 1$ and to achieve this upper bound, one must have $X = Y \equiv 1$

On the other hand, it is more complicated to find and prove a lower bound of $E(T)$, to do so, we will employ the inequality of Cauchy-Schwarz:

Theorem 1.

Cauchy Schwarz inequality for real functions
For two real measurable functions $f, g \in L^2([a,b])$ and two real numbers a,b with $a < b$

$$|\int_a^b f(t)g(t)dt| \leq (|\int_a^b f(t)^2 dt)|^{\frac{1}{2}} |\int_a^b f(t)^2 dt)|^{\frac{1}{2}}$$

The equality can be achieved when $\exists \alpha \in R$ such that $f(t) = \alpha g(t)$ almost every where in $[a,b]$.

According to this theorem, we have:

$$\int_0^T F_X(t)dt \leq (\int_0^T (F_X(t))^2 dt)^{\frac{1}{2}} (\int_0^T 1 dt)^{\frac{1}{2}}$$

In the case when T big enough, we have $\int_0^T (F_X(t))^2 dt \geq \frac{(T-1)^2}{T}$

This leads to

$$2\int_0^\infty t f_X(t) F_X(t) dt \leq T - \frac{(T-1)^2}{T} = 2 - \frac{1}{T}$$

Using (*)

$$E(T) \geq 2E(X) - (2 - \frac{1}{T}) = \frac{1}{T} \to 0$$

The lower bound thus obtained by Cauchy Schwarz inequality is 0, which is quite obvious. Now let us check the condition of equality ie.$\exists \alpha \in \mathbb{R}, F_X(t) = \alpha \times 1$ almost every where when t is non negative.
As $F_X(t)$ is a probability distribution function, it is obvious that $F_X(t) \xrightarrow{+\infty} 1$ so $\alpha = 1$, therefore, to obtain this lower bound, one must have: $F_X(t) = 1$ almost everywhere for t non-negative. We then further notice that $F_X(t)$ is in fact an increasing function with upper-bound 1. In this case

$$F_X(t) = 1 \quad \forall t > 0$$

Otherwise suppose $\exists t \in \mathbb{R}, F_X(t) < 1$ thus $F_X(x) < 1 \quad \forall x \in [0,t]$ as $measure([0,t]) \neq 0$. It is a contradiction of $F_X(t) = 1$ almost everywhere.

By definition, $F_X(t) = P(X < t) = 1, \forall t > 0$, in particular $F_X(0.5) = P(X < 0.5) = 1$ and this is in fact in contradiction with our fundamental assumption $E(X) = 1$

Thus we conclude that the zero lower-bound can not be achieved.

However, we have found a function series F_{X_n} as distribution functions for a series of variables X_n, such that

$$\begin{cases} \int_0^\infty |F_{X_n}(t) - 1|^2 = 0 \\ E(X_n) = 1 \quad \forall n \\ E(min(X_n, X_n')) \xrightarrow{\infty} 0 \quad \text{where } X_n' \text{ is identically distributed and independent to } X_n \end{cases} \tag{***}$$

This series of functions will be given in the proof of next proposition.

Property 4.3.

There exists no random variable with expectation 1 that "equally dominates" all the other random variables. ie.

$$\nexists X \quad with \quad E(X)=1 \quad such\ that \quad \forall Y \quad Y \neq_d X, E(Y)=1, \quad X \xrightarrow[dominant]{} Y$$

where $Y \neq_d X$ means that Y is not identically distributed as X

Proof. We start by construct a sequence of discrete density functions $P(X_n = t)$

$$P(X_n = t) \longmapsto \begin{cases} \frac{1}{n} & \text{if } t = n \\ 1 - \frac{1}{n} & \text{if } t = 0 \\ 0 & \text{if } t \neq 0, n \end{cases}$$

Thus, the distribution functions F_{X_n} for variables X_n

$$F_{X_n}(t) = P(X_n \leq t) \longmapsto \begin{cases} 1 - \frac{1}{n} & \text{if } 0 \leq t < n \\ 1 & \text{if } n \leq t \end{cases}$$

Thus, the distribution functions F_{X_n} for variables X_n

□

We plot the density of function of $F_{X_4}, F_{X_8}, F_{X_{12}}$ which allows us to have a direct vision on the evolution of function series.

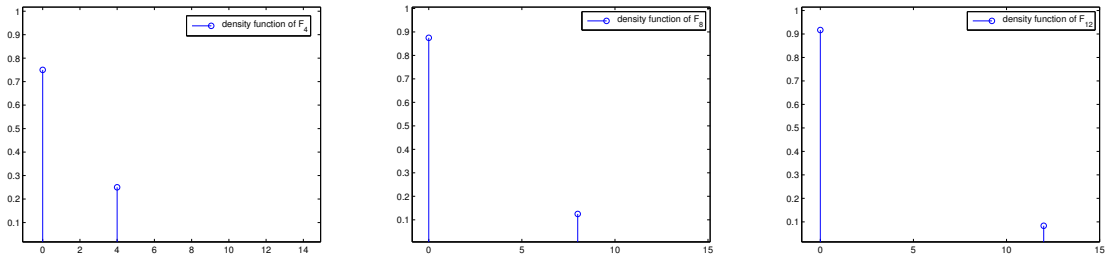


Figure 10: density of f_{X_n}

It is easy to check that

$$E(X_n) = 1 \quad \forall n$$

$$\int_0^{\infty} |F_{X_n}(t) - 1|^2 = n \times \frac{1}{n^2} = \frac{1}{n} \xrightarrow{\infty} 0$$

$$E(min(X_n, X_n')) = n \times P(X_n \neq 0) \times P(X_n' \neq 0)) = \frac{1}{n} \xrightarrow{\infty} 0$$

F_{X_n} verify all three criteria in (***), further more, actually $\forall$ fixed non-negative variable Y with expectation $E(Y) = 1, \exists n \in \mathbb{N}$ such that Y can not "equally dominated" $X_n \quad Y \not\xrightarrow[\text{dominant}]{} X_n$, further more $P(X_n < Y) > P(X_n > y)$

Suppose the probability distribution law of variable Y is fixed and given noted F_Y, suppose $P(Y > 0) = \beta$, obviously $\beta > 0$ as $E(Y) = 1$.
$\exists n \in \mathbb{N}$ such that $n > \frac{1-\beta}{\beta}$ therefore $P(X_n > Y) \leq P(X_n \neq 0) = \frac{1}{n}$ and $P(X_n < Y) \geq P(X_n = 0)P(Y \neq 0) = (1 - \frac{1}{n})\beta$

$$n > \frac{1-\beta}{\beta} \Longrightarrow (1 - \frac{1}{n})\beta > \frac{1}{n} \quad \text{so} \quad P(X_n < Y) > P(X_n > y)$$

4.1.2 case of n potential transitions

Now we come back to the general case when a node in the graph processes n potential transitions targets. The random variables $X_1, X_2, ...X_n$ present the potential transition time. The real time transition time can be expressed as $T_n = min(X_1, X_2, ...X_n)$.

iid variables We start with the most simple case where $X_1, X_2, ...X_n$ are all independent and identically distributed. As always, we suppose $E(X_i) = 1$

$$F_{T_n}(t) = P(T_n \leq t) = 1 - P(T > t) = 1 - P(X_1 > t)P(X_2 > t)...P(X_n > t) = 1 - (1 - F_X(t))^n$$

Suppose $F_X(t)$ is differentiable.

$$\begin{aligned} E(T_n) &= \int_0^\infty t dF_{T_n}(t) \\ &= \int_0^\infty t n f_X(t)(1 - F_X(t))^{n-1} dt \end{aligned}$$

In the particular case, when X_i follows an uniform distribution (in our case $unif(0,2)$ as $E(X) = 1$)

$$F_{T_n}(t) \begin{cases} 1 - (\frac{2-t}{2})^n & t \in (0,2) \\ 0 & t \leq 0 \\ 1 & 2 \leq t \end{cases}$$

So

$$f_{T_n}(t) \begin{cases} \frac{n}{2}(\frac{2-t}{2})^{n-1} & t \in (0,2) \\ 0 & \text{otherwise} \end{cases}$$

$$E(T_n) = \int_0^\infty t f_{T_n}(t) dt = \frac{2}{n+1}$$

One can easily deduce by definition that $E(T_n)$ is a decreasing function of n. We are extremely interested in this section to check the form of curve when X_i follows different probability distribution.
The curves of (T_n, n) have been drawn for four distributions: $exp(1), unif(0.2), Pareto(2, 0.5), Pareto(1.25, 0.2)$

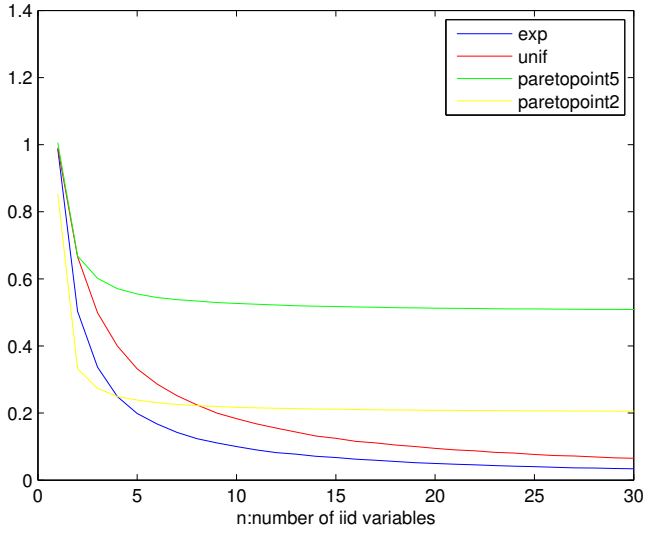


Figure 11: Curve (n, T_n) for iid variables

Before it reaches the lower bound of Pareto distribution, one observes that the Pareto distributions have an advantage in terms of decreasing speed. When the number of targets of transition is limited (n is not large), the Pareto transition time should own an advantage of diffusion speed.

Shifted Pareto distribution In order to avoid the problem that Pareto variables can not go below certain threshold, we have also used the shifted Pareto variables. The main idea is to keep the slow decreasing form of Pareto variables and the expectation value as 1 while to avoid the lower threshold in the density functions of Pareto variables.

Giving an example of $Pareto(0.5,2)$, with its density function:

$$f_X(t) = 2\frac{0.5^2}{t^3}$$

The density function of shifted Pareto variable is:

$$f_X(t) = 2\frac{0.75^2}{(t+0.5)^3}$$

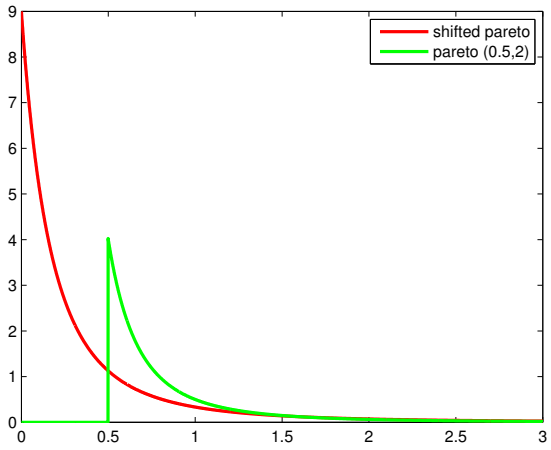


Figure 12: density function of shifted Pareto variable

According to numerical experiences, the advantage of accelerating the diffusion presented in Table 1 becomes more obvious if we use the shifted Pareto variable instead.

$$T_{i,j} = min(X_i, X_j)$$

distribution	constant:1	exp(1)	unif(0,2)	shifted Pareto k=2
constant:1	1.0000	0.6309	0.7487	0.4230
exp(1)	0.6314	0.4979	0.5710	0.3427
unif(0,2)	0.7486	0.5666	0.6632	0.3828
shifted Pareto k=2	0.4239	0.3444	0.3798	0.2502

$$P(i,j) = Proba(X_i < X_j)$$

distribution	constant:1	exp(1)	unif(0,2)	shifted Pareto k=2
constant:1	0.5	0.3674	0.4999	0.1930
exp(1)	0.6317	0.5007	0.5674	0.3446
unif(0,2)	0.5012	0.4314	0.5022	0.2776
shifted Pareto k=2	0.8067	0.6574	0.7264	0.5016

More importantly, when we draw the curve of $(E(T_n), n)$ the lower threshold of Pareto variables disappear. One observes in figure 13 a clear advantage for Pareto variables no matter the value of n.

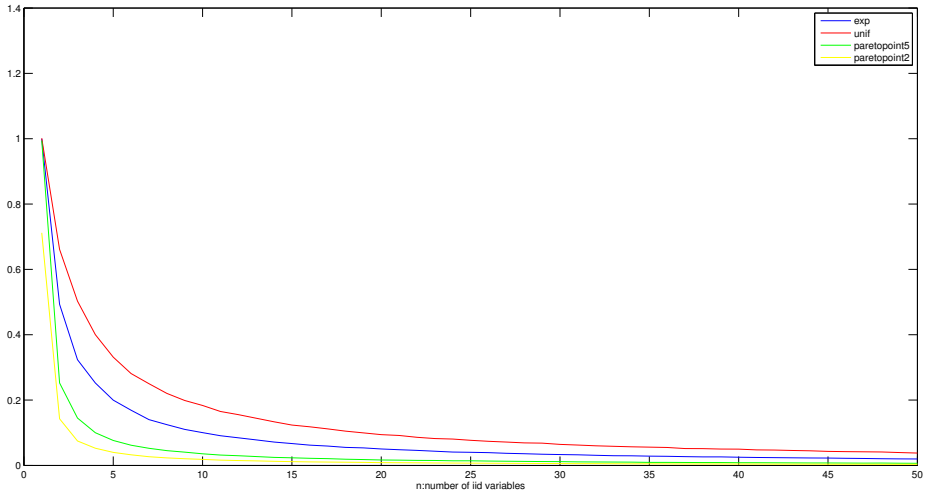


Figure 13: iid variables with shifted Pareto

4.2 "meeting time model"

4.2.1 Bus paradox

In order to explain the idea of the "meeting time model" and make some reminders on stochastic process, we cite here a famous example called Bus paradox.

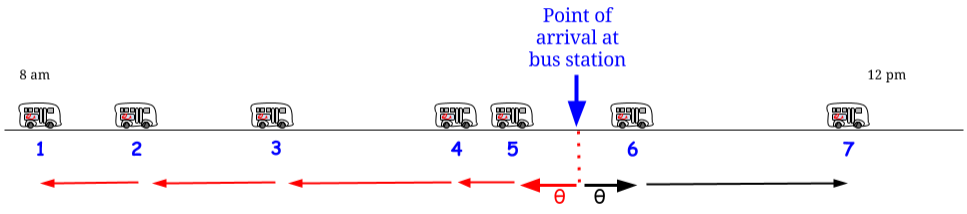


Figure 14: figure taking from https://stats.stackexchange.com/questions/122722/please-explain-the-waiting-paradox

State the problem: At a bus station the time that a bus arrives follows a Poisson process of parameter λ which

means the time between every two buses can be simulated by an exponential distribution. If one person at a random time at the bus station, what will be probability distribution of his waiting time and how long should he wait for the first bus in average?

An important property of Poisson process is its memorylessness. We can simply prove it using the distribution function of exponential variables:

$$\begin{aligned}\forall t_0 \quad P(\text{ Wait at least } t \text{ minutes}|\text{Already waited } t_0 \text{ minutes}) &= P(X > t + t_0 | X > t_0) \\ &= \frac{P(X > t + t_0, X > t)}{P(X > t)} \\ = \frac{\int_{t+t_0}^{\infty} \lambda e^{-\lambda x}}{\int_{t_0}^{\infty} \lambda e^{-\lambda x}} &= \frac{e^{-\lambda(t+t_0)}}{e^{-\lambda t_0}} = exp(-\lambda t) \\ &= P(\text{ Wait at least } t \text{ minutes})\end{aligned}$$

This result is in fact independent on the value of t_0

Denote R_t as the variable that presents the time an individual has to wait for next bus if he/she arrives the station at time t.

$$R_t = \sum_{i=1}^{m} X_i - t \quad \text{for} \sum_{i=1}^{m-1} X_i < t \leq \sum_{i=1}^{m} X_i$$

We note $R_t = WTD(X, t)$, WTD present the waiting time distribution.

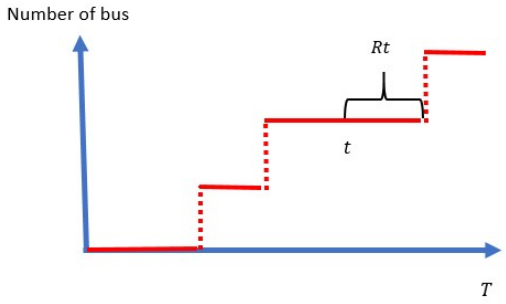


Figure 15: definition of the waiting time R_t

suppose $t_0 = t - \sum_{i=1}^{m-1} X_i$

$$F_R(x) = P(R \leq x) = 1 - P(R > x) = 1 - P(X_m > x + t_0 | X_m > t_0) = 1 - P(X_m > x) = 1 - exp(-\lambda t)$$

Therefore, the distribution of R is the same as X_i, so $R \sim exp(\lambda)$

For any stochastic process when the distribution function of intervals is non-lattice, there is a more general result:

$$\lim_{t \to +\infty} E(R_t) = \frac{E(X_i^2)}{2E(X_i)} \tag{4*}$$

The proof of this result can be found in [K.Van Harn,F.W.Steutel,1994]
We are also interested in the probability distribution of variable $\lim_{t\to+\infty} R_t$ in this general case thus we suppose R is the waiting time variable when an individual arrives at a random time during an infinite long process.

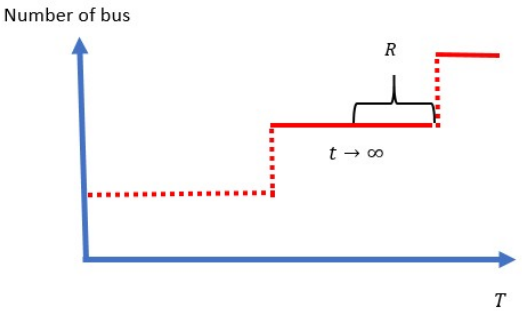


Figure 16: definition of the waiting time R_t

Note the event Au: For a given positive real number u, the individual arrive during the period (between two buses) X_i such that $u \le X_i \le u + du$ where du is an elementary time period

$$\begin{aligned} P(R > t) &= \int_0^\infty P(R > t|Au)P(Au)dt \\ &= \int_0^\infty P(R > t|u \le X_i < u + du)P(u \le X_i < u + du) \\ &= \int_t^\infty \frac{u-t}{u} f_X(u)du \\ &= \int_t^\infty f_X(u)du - \int_t^\infty \frac{t}{u} f_X(u)du \\ &= 1 - F_X(t) - \int_t^\infty \frac{t}{u} f_X(u)du \end{aligned}$$

Therefore,

$$\begin{aligned} F_R(t) &= P(R \le t) = 1 - P(R > t) \\ &= F_X(t) + \int_t^\infty \frac{t}{u} f_X(u)du \end{aligned}$$

As for the density function

$$\begin{aligned} f_R(t) &= \frac{dF_R(t)}{dt} \\ &= \int_t^\infty \frac{1}{u} f_X(u)du \end{aligned}$$

4.2.2 case of two potential transitions

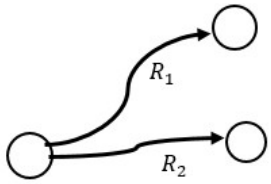


Figure 17: two potential transitions in "meeting time model"

$T = min(R_1, R_2)$, R_1, R_2 are generated by two different process with eaqual average interval time period

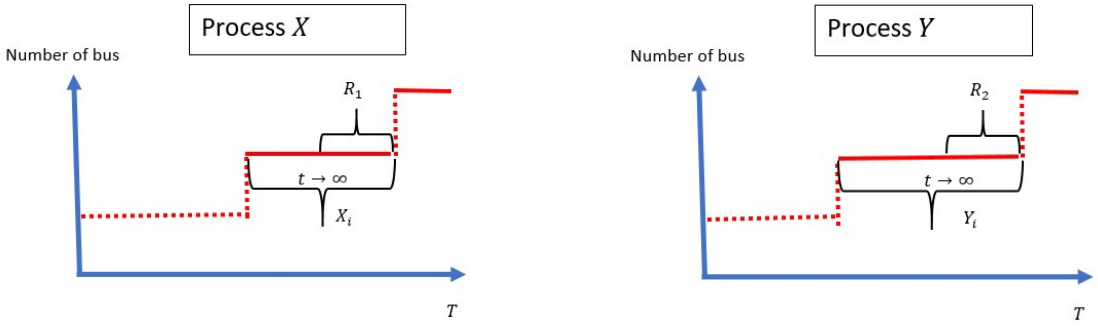


Figure 18: transition time simulated by two independent process

Like in the "transition time model", in order to better compare the influence of different process on the diffusion, we have fixed $E(X_i) = E(Y_i) = 1$. However this doesn't implict $E(R_1) = E(R_2)$
The result in (4*) could be used:

$$E(R_1) = \frac{E(X_i^2)}{2E(X_i)} = \frac{Var(X_i) + E(X_i)^2}{2E(X_i)} = \frac{Var(X_i) + 1}{2}$$

$$E(R_2) = \frac{E(Y_i^2)}{2E(Y_i)} = \frac{Var(Y_i) + E(Y_i)^2}{2E(Y_i)} = \frac{Var(Y_i) + 1}{2}$$

In general case, $E(R) \geq \frac{1}{2}$ and $E(R) = \frac{1}{2}$ if and only if $X_i \equiv 1$ as in this case $Var(X_i) = 0$
In particular, when $X_i \sim exp(\lambda)$, $E(R) = E(X_i) = \frac{1}{\lambda}$
We are now interested in the "equal dominance" of variables R generated by different stochastic process of interval period X_i.

Definition 4.3.

We define in this chapter that a stochastic process with iid interval time variable X is "WTD dominated" by another stochastic process with iid interval time variable Y if and only if

$$\begin{cases} E(X) = E(Y) < \infty \\ P(R_X > R_Y) > 0.5 \end{cases}$$

where

$$R_X = WTD(processX)$$

present the variable of waiting time in $processX$ for current time $t \longrightarrow \infty$

$$R_Y = WTD(processY)$$

present the variable of waiting time in $processY$ for current time $t \longrightarrow \infty$

Property 4.4.

The "equal dominance" does not imply the "WTD dominance" and the converse is also not true. i.e

$$X \xrightarrow[\text{equally dominant}]{} Y \nRightarrow X \xrightarrow[\text{WTD dominant}]{} Y$$

$$Y \xrightarrow[\text{WTD equally dominant}]{} X \nRightarrow Y \xrightarrow[\text{equally dominant}]{} X$$

Proof. It is sufficient to find two variables X and Y with different probability distributions such that

$$X \xrightarrow[\text{equally dominant}]{} Y \quad \text{and} \quad Y \xrightarrow[\text{WTD dominant}]{} X$$

We simply take $X \sim exp(1)$ and $Y \equiv 1$

$$P(Y < X) = F_X(1) = exp(-1) < 0.5 \Rightarrow X \xrightarrow[\text{equally dominant}]{} Y$$

on the other hand, suppose $R_X = WTD(X)$ and $R_Y = WTD(Y)$

As proved before, the exponential stochastic process is memory-less so $R_X \sim exp(1)$. It is also easy to find that $R_Y \sim unif(0, 0.5)$

$$\begin{aligned} P(R_X < R_Y) &= \int_0^{+\infty} P(t < Y) f_X(t) dt \\ &= \int_0^1 exp(-t) \frac{1}{1-0} dt \\ &= 1 - exp(-1) \simeq 0.632 > 0.5 \end{aligned}$$

$$\text{Therefore} \quad Y \xrightarrow[\text{WTD dominant}]{} X$$

□

Similar to the study of "transition time model". We have established numerical results using the four probability distribution presented in Table 1.
We first generate four waiting time variables R_1, R_2, R_3, R_4.

$$T_{i,j} = min(R_i, R_j)$$

distribution	constant:1	exp(1)	unif(0,2)	Pareto (1.25,0.2)
constant:1	0.3342	0.3669	0.3401	0.3335
exp(1)	0.3675	0.4983	0.4416	0.4997
unif(0,2)	0.3431	0.4427	0.4206	0.6295
Pareto (1.25,0.2)	0.3325	0.5641	0.5962	2.3809

Table 4: $E(T)$ for each pairs of variables R_i, R_j

$$P(i,j) = Proba(R_i < R_j) = P(R_i \leq R_j) \quad \text{as} \quad \text{all four distributions are continuous}$$

distribution	constant:1	exp(1)	unif(0,2)	Pareto (1.25,0.2)
constant:1	0.5041	0.6414	0.5494	0.6376
exp(1)	0.3702	0.5018	0.4383	0.5745
unif(0,2)	0.4486	0.5519	0.5056	0.6070
Pareto (1.25,0.2)	0.3649	0.4355	0.4038	0.4922

Table 5: $P(R_i \geq R_j))$ for each pairs of variabales

Comparing table 4 with table 1, it is quite remarkable that the order of dominance in "meeting time model" is almost the inverse of the "transition time model".
As we have already seen the variables with small variance may own an advantage in this competition. Intuitively the reason can be explained by the figures below:

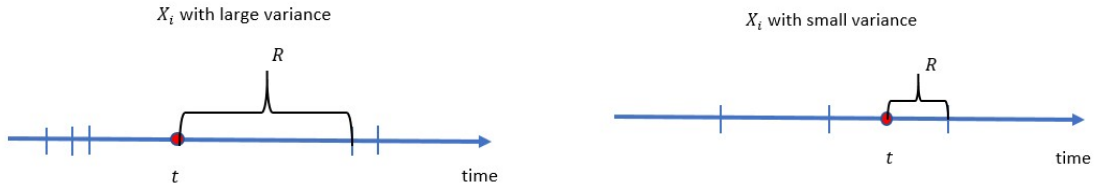


Figure 19: rest time in stochastic process

As presented in figure 19, in this model when time periods X_i are not homogeneous (which leads to a large variance), we actually have more chance to find ourselves in the middle of a large period instead of a small one. Meanwhile when X_i are more homogeneous, the value of R become more predictable and limited.

$$T = min(R_X, R_Y)$$

We could use the same reasoning in the "transition time model"

$$E(T) = E(R_X) + E(R_Y) - \int_0^\infty t f_{R_X}(t) F_{R_Y}(t) dt - \int_0^\infty t f_{R_Y}(t) F_{R_X}(t) dt$$

Suppose that X and Y follow the same probability distribution.

$$E(T) = 2E(R_X) - 2\int_0^\infty t f_{R_X}(t) F_{R_X}(t) dt$$

However, different from the "transition time model", this time $E(R_X)$ is no longer known and the expression of $\int_0^\infty t f_{R_X}(t) F_{R_X}(t) dt$ become more complex as only the distribution of X is known. Therefore, it is very hard to provide a general analysis. As our main interest is how the variance of interval time variable make influence on $E(T)$. We will now limit ourselves on uniform distributions.

As we keep always the condition $E(X) = 1$, the non-negative uniform distributions considered could be written in the form $unif(1-a, 1+a)$ for $0 \leq a \leq 1$

Property 4.5. *There is no "equal dominance" between two uniform variables with same expectation*

Proof.

$$\text{suppose} \quad E(X) = E(Y) = 1, \quad X \sim unif(1-a, 1+a), Y \sim unif(1-b, 1+b) \quad \text{and} \quad a > b$$

$$\begin{aligned} P(X < Y) &= \int_0^\infty P(X < t) f_Y(t) dt \\ &= \int_0^\infty P(X < t) f_Y(t) dt \\ &= \int_{1-a}^{1+a} \frac{t-(1-b)}{2b} \frac{1}{2a} dt \\ &= \frac{1}{2b} - \frac{1-b}{2b} = \frac{1}{2} \end{aligned}$$

This result is quite obvious, as their density functions are symmetric.

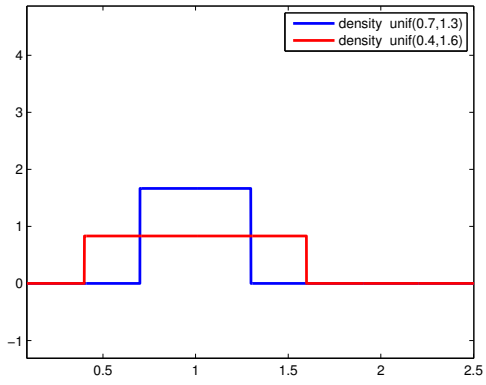


Figure 20: density uniform distribution

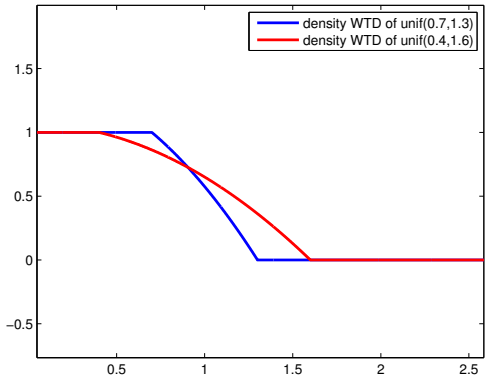


Figure 21: density of uniform WTD

This result shows again the variance is not the only factor that decides the order of "equal dominance".

□

We then look at the WTD case of uniform distribution.

suppose $X \sim unif(1-a, 1+a), Y \sim unif(1-b, 1+b)$ $R_X = WTD(processX), R_Y = WTD(processY)$

$$\text{We would like to prove} \quad P(R_X < R_Y) > \frac{1}{2} \quad \text{when} \quad a > b$$

In fact, it is not easy to calculate directly the waitting time distribution. We have used a $monte-carlo$ simulation to verify the probabikity $P(R_X < R_Y)$. We take 10 different uniform distributions $unif(0,2), unif(0.1,1.9)....unif(0.9,1.1)$. The pixel i, j in the image present $P(unif(1-0.1\times i, 1+0.1\times i) < unif(1-0.1\times j, 1+0.1\times j))$

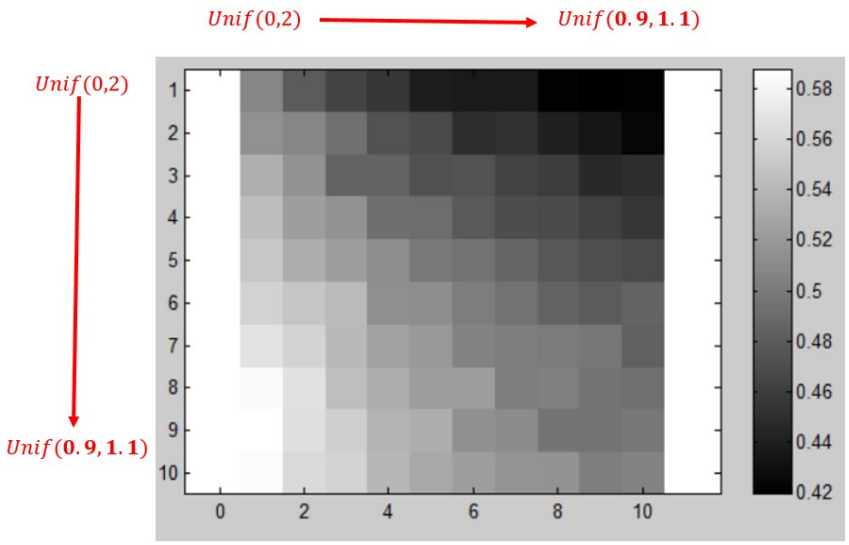


Figure 22: $P(unif(1-0.1\times i, 1+0.1\times i) < unif(1-0.1\times j, 1+0.1\times j))$

From figure 22, we can easy see that $P(R_X < R_Y) > \frac{1}{2}$ for $X \sim unif(1-a, 1+a), Y \sim unif(1-b, 1+b)$ and $a > b$. Therefore, unlike in "transition time model", when meeting time between two person is irregular it seems the diffusion could accelerate.

4.2.3 case of n potential transitions

Similar to the "transition time model", we are interested in how fast $E(T_n = min(R_1, R_2, R_3 ... R_n))$ will decrease when the number of variables n increases. Here the R_i are generated by the same or different types of stochastic process.

Property 4.6.

$$\forall m, \quad T_m = min(R_1, R_2, R_3 ... R_m)$$

$$sup(E(T_m) = +\infty)$$

Proof. We use again the sequence of density functions defined before.

$$P(X_n = t) \longmapsto \begin{cases} \frac{1}{n} & \text{if } t = n \\ 1 - \frac{1}{n} & \text{if } t = 0 \\ 0 & \text{if } t \neq 0, n \end{cases}$$

$$F_{X_n}(t) = P(X_n \leq t) \longmapsto \begin{cases} 1 - \frac{1}{n} & \text{if } 0 \leq t < n \\ 1 & \text{if } n \leq t \end{cases}$$

□

In fact, $X_i \in 0, n$ means several buses arrive at the station at the same time and then we have to wait n hours for the next bus. Thus it is easy to get $R_{X_i} \sim unif(0, n)$

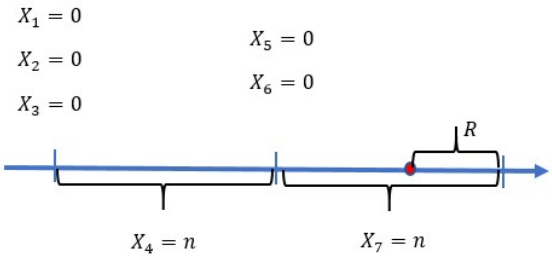


Figure 23: interval time variables of distribution F_{X_n}

suppose $X_1, ..., X_n$ are n iid variables following this distribution law. Then $\forall i, R_{X_i} \sim unif(0, i)$

$$\text{when} \quad n \to +\infty, E(min(R_{X_1} ... R_{X_m})) = E(T_m) \to +\infty$$

The property is thus proved.

iid variables Now we consider the case where R_i are generated by iid process X_i. Similar to the "transition time model", we draw the curve of $(n, E(T_n))$ obtained by Monte Carlo simulations.

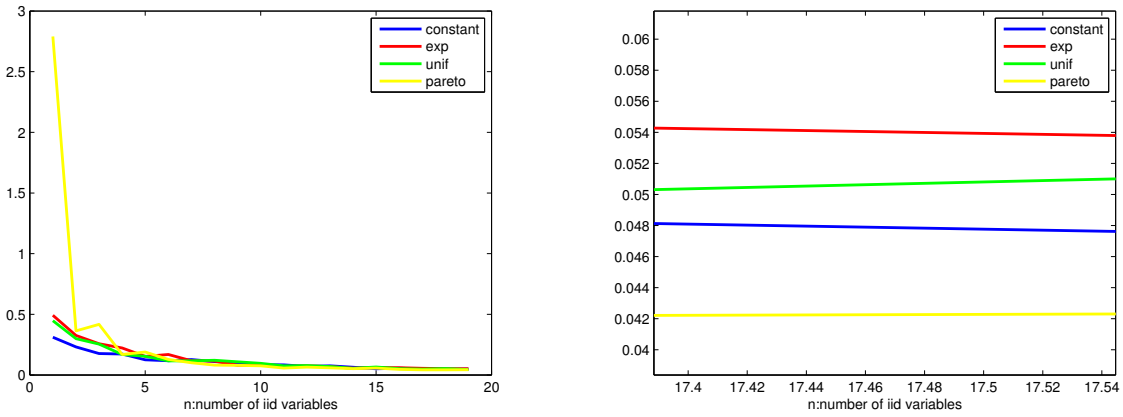


Figure 24: iid variables in meeting time model the figure on the right side is the zoom of the figure on the left

One observes from figure 24 that when n increase, the value of $E(T_n)$ decrease very fast for $Pareto$ process. In fact, when n is sufficiently big, the minimum of iid $Pareto$ variables actually goes below than the other distributions. An explanation to this phenomenon is that when the number of $Pareto$ variables increases, despite the fact we could be inside a large period, we also get a great chance to find ourselves in a extremely tight period generated by $Pareto$ variables. As we take the minimum of generated waiting time, these small values could influence a lot on the final result.**Therefore, when number of potential transition n is large enough, the Pareto distributed meeting time may help to accelerate the diffusion.**

mixed variables Instead of taking iid variables, we are now going to mixed different type of potential transitions. In fact, this model may be more near to the reality as in life we do meet some peope more regularly than the others.

To examine the minimum of mixed variables, we pick 10 waiting time variables generated by two types of distribution of meeting time.For $n \in [0,10]$, n among 10 variables are generated by the first distribution (of meeting time) and the rest are generated by the second one. We then vary n from 0 to 10 and estimate the expectation of the minimum of this ten variables by a Monte Carlo simulation.
The curve of the min value and n for Pareto(1.25,0.2) and constant(1) (ie when n=0, there are just 10 variables generated by constant meeting time) are presented in figure 25:

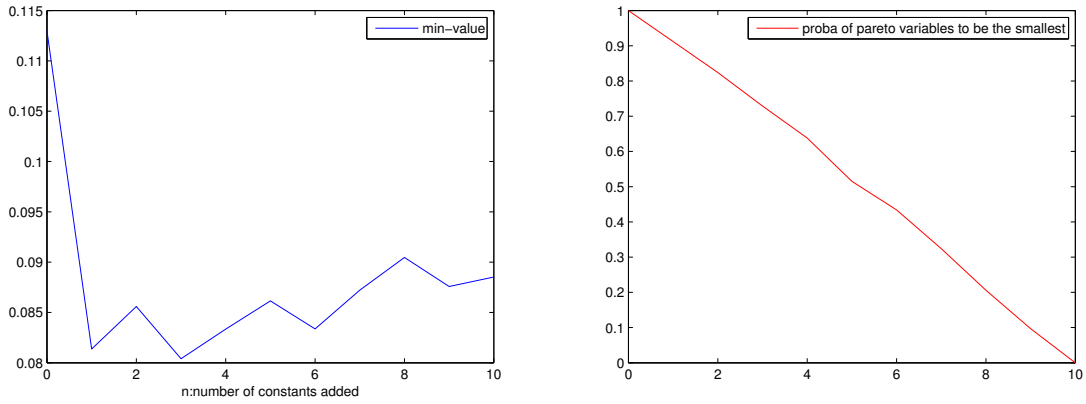


Figure 25: mixed distribution time of Pareto distribution and constant

Here, the Monte Carlo simulation is made by 1000 times experiences, the result is not very precise . However it is enough get the idea: when there is one uniform variable (generated by constant meeting time), it helps to eliminate the extreme case of huge value in transition time variables generated by Pareto meeting time. That explains the huge gap between n=0 and n=1. when n increases (so more uniform variables), the minvalue then re-increase slowly. The right figure present the probability that the minimum among the 10 variables are generated by a Pareto meeting time. It is almost linearly decreasing.
The same experiment using exponential meeting time instead of constant meeting time is made as well and the result is very similar

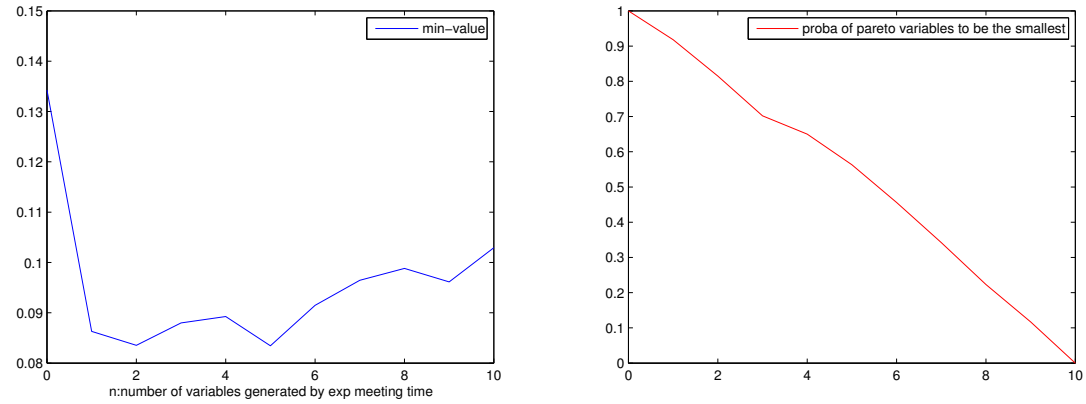


Figure 26: mixed distribution time of exponential distribution and constant

According to the numerical results of these two models, it is easy to see that in a network where different types of potential transition time distribution are mixed, the diffusion could be faster than the network with only one type of potential transition time distribution.

5 Conclusion of chapter 1

In this chapter, based on two different models, we focus on probability distributions of potential transition time among nodes with same average. The main interest is to discover how they may influence on the network diffusion. To do so, we have defined two dominance relations between probability distribution describing the intercontact in a social network. They are

- "equal dominance"
- "WTD dominance"

These two relations could be used to define an "advantage order" respectively in the "transition time model" and "meeting time model". **It seems the shifted Pareto distribution owns an advantage for fast spreading in the "transition time model" among other common continuous probability distributions because of its large variance. On the other hand,it is hard to give a definitive conclusion for which distribution of intercontact time may generate the fastest diffusion in the "meeting time model".** It depends also on the structure of social network. For example, we have seen that the Pareto distribution is often dominated by other distributions when there are only two potential transition targets. However several potential transition generated by Pareto can in fact speed up the diffusion.
Several proprieties and numerical experiments have also been developed around these two relations of dominance.
This work allows us to have a more clear version on this problematic. We understand the regularity of people meeting each could bring an important impact on the diffusion in social network. As have mentioned in the summary of this chapter, there still remains a lot of interesting points that we don't have time to cover in this thesis. We could establish a direct link to the disease spreading. For example, we can suppose 50% of potential transition time follow a Pareto distribution while the other 50% follow the exponential distribution in a social network. We can then change the percentage of two distributions as well as the structure of contact network and we would like to see how it will change the behavior of disease spreading. By doing that, we combine the micro modelling (chapter 1) and macro modelling (chapter 2) together. We believe the result will not be trivial and it should not be easy to understand the mathematical reasons behind.

Chapter 2: Macro-modelling in Graph Structures

"Early diagnosis of disease is the business of the general public even more than of the medical profession."
— J.B.S. Haldane

1 Summary

In this chapter we concentrate ourselves on the influence of disease spreading caused by the structure of social network. The key ingredient is the reproductive ratio of disease introduced by George MacDonald in 1952. It is well adapted that this quantity present not only the average number of secondary infection caused by a primary infected but also the threshold of disease explosion in long period. However, according to different models and systems the formalization has not been clearly precised yet as mentioned in [J.Holland, 2012] (see figure 74 in Appendix).
In this chapter introduce three different methods to simulate the epidemic diffusion: dynamical systems, fixed contact network and random contact network. Using dynamical systems, we consider all individuals are equivalent and symmetric which means they all own the probability to be infected, recovered etc. The reproductive ratio in this case can be easily defined as presented in section **4** while the two result of two other methods strongly depend on the structure of graph connection.
Several analytic tools and numerical simulations have been created in order to find a precise definition of reproductive ratio in both fixed contact and random contact models mainly for **SIS** and **SIRS** system. To do so,different structures of Adjacency matrix have been employed.
The study is split into two cases where in first(resp.second) one each node in the connection graph present one individual (resp. community), therefore the adjacency matrix is binary (resp. weighted). The analysis using two real data set is made in section **7** of this chapter in order to verify the validity of our methods. A brief conclusion at the end of the chapter highlights our contribution in this paper.
The key message of this chapter could be summarized as : **While using a graph based model for disease spreading, the non-homogeneity among nodes brought by adjacency matrix makes it difficult to define the reproductive ratio quantity which shows the threshold of explosion.**

2 Introduction and Motivation

In this chapter we concentrate on the adjacency matrix modelling of temporary interconnection graph. The transition time between each nodes will be considered as a constant, our interest is to check how different structure of adjacency matrix may influence the performance of epidemic spread.
To achieve this objective, we model the contact of individuals in the social network by random adjacency matrices. Two different modellings have been created, first we suppose once the random matrix of contact is simulated, it remains constant during the period of spreading and for the second modelling we suppose conversely that the temporary random adjacency matrix change its forms at each step.

3 R_0 theory and examples of social networks

R_0 is known as basic reproduction rate, which is used to estimate the transmission potential of disease spread. It is generally defined as the average number of secondary infections produced by a typical case of an infection in a population that is totally susceptible. However, there exists a few different ways to calculate concretely the value of R_0(See appendix). The idea is in fact simple, for example if $R_0 = 10$ for certain disease in a population, that means a patient will infect other 10 people before he is cured, we can then expect an exponential increasing of infection. The value of R_0 can be affected by several factors, such as:

- The frequency of contact between infectious agent and susceptible agent;

- The probability of infection being transmitted during contact;
- The duration of infectiousness before patient is cured or eventually dead

For example

- Adults with the flu are typically contagious for up to eight days, while children can be contagious for up to two weeks. The longer the infectious period of a disease, the more likely an infected person is to spread the disease to other people. A long period of infectiousness will contribute to a higher R_0 value.
- The diseases that spread the most quickly and easily are the ones that can travel through the air, such as the flu or measles.In contrast, diseases that are transmitted through bodily fluids, such as Ebola or HIV, are fortunately not so easy to catch or spread.

In fact, R_0 of a disease is not a constant during a long period. Because of medical reasons and immunity, $R_{0,t}$ is normally considered as a decreasing function of t. However, due to certain incontrollable factors (such as variation of virus), the $R_{0,t}$ could also evolve randomly.

It is believed that a threshold of invasion on the reproduction rate is 1: for values above this, an infection can grow in the population and the disease can successfully invade. Generally for a disease to be capable to continue spreading, it should own a value R_0 strictly higher than 1. This means i a current patient should infect at least one person in average for the disease can eventually survive with non-zero probability. Otherwise in the case $R_0 < 1$, the extinction of disease will take place.

There exist several different definitions of R_0 (Please see figure 74 in Appendix). Most of them are based on the concept of outbreak threshold. The two most common methods to define the reproductive ratio is:

- The dominant eigenvalue value of the transition probability matrix which is identical to the network adjacency matrix if we suppose each contact with infected can contaminate the disease.
- The expected number of secondary infected caused by a random patient in the network.

We have also find an interesting definition of R_0 in [Leon Danon et $al.$,2011]. They use the next-generation matrix defined as:

$$K_{k,m} = \frac{[km](m-1)}{m[m]}p$$

Here $K_{k,m}$ presents the number of cases for individuals with k contacts from an individual with m contacts. $[m]$ presents the number of individuals with m edges in the network and $[km]$ is the number of edges between individuals with k and m contacts, respectively. In addition, p is the probability of infection eventually passing across the edge between a susceptible-infectious pair. The basic reproductive ratio is given by the dominant eigenvalue of the next-generation matrix.

$$R_0 = \text{spectral radius}(K_{k,m})$$

The condition of this definition is that the contact network should remain constant during the period of spreading. Which is the case of our fixed contact model simulation see section **6.2**.
We have implemented this algorithm in Matlab, the code can be found in Appendix. Its result will be compared with other methods later in this chapter.
The estimate R_0 values of some well-known disease such as HIV and Measles are presented in Figure 27

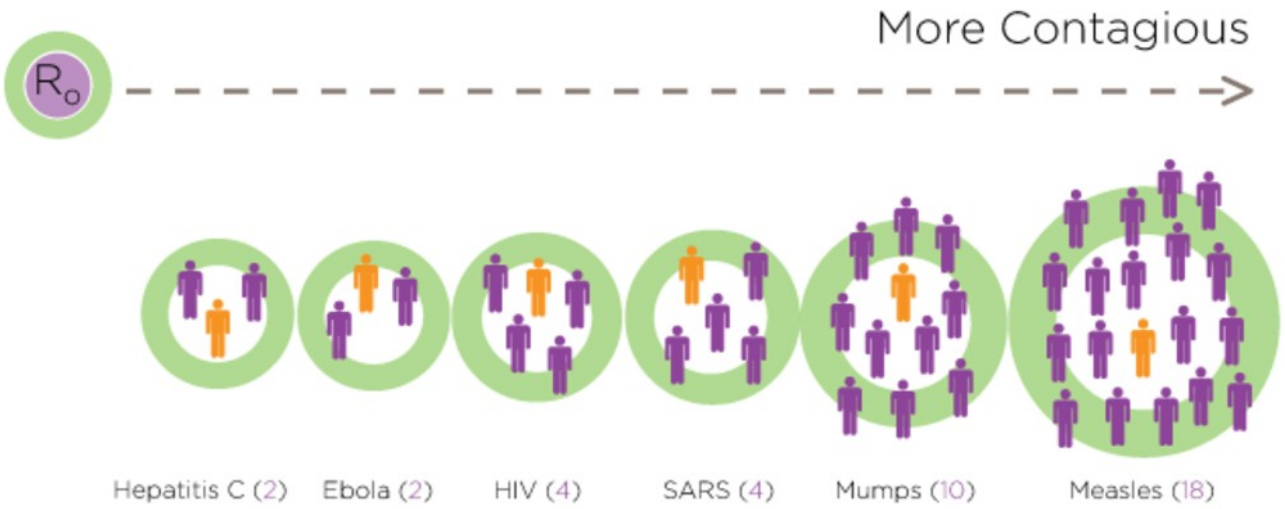


Figure 27: example of R_0 for different diseases from Healthline `http://www.healthline.com/health/r-nought-reproduction-number#calculation3`

It is known Measles virus could spread by coughing and sneezing while HIV virus spread with a contact much more intimate like sexual behaviors or sharing needle or syringe use. This explains the huge difference of R_0 and also in epidemic curve.

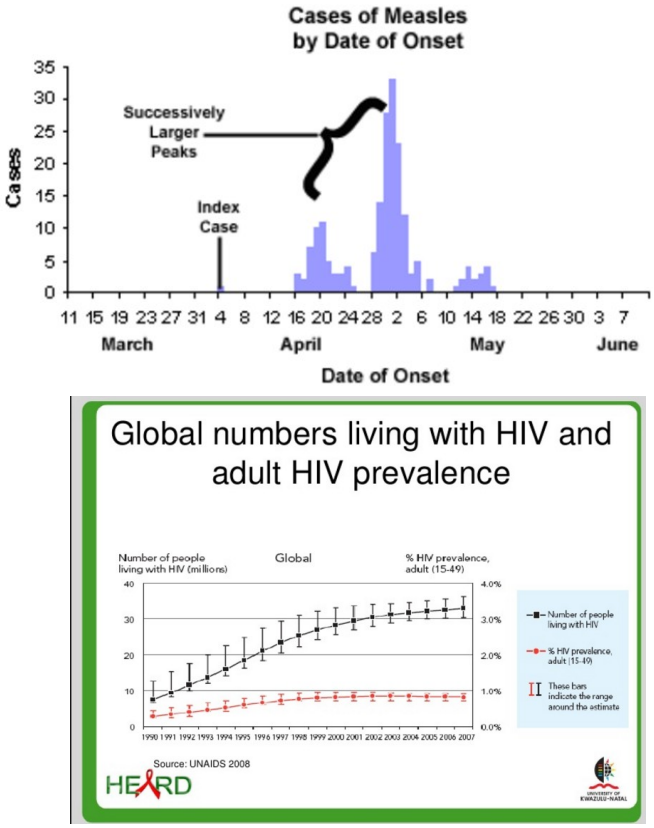


Figure 28: epidemic curve of Measles and Hiv from Alain Whiteside, Descriptive Epidemiology

epidemic curves An "epidemic curve" shows the frequency of new cases over time based on the date of onset of disease.

From figure 28, one observes that there exists a huge difference between Measles and HIV in appearance of new cases. In the past 50 years since its first appearance, the number of infected of HIV stays stable and increases gently. Meanwhile one observes from the epidemic curve of Measles that there often exists some larger peaks which correspond to a sudden break out of disease. The successive waves tend to involve more and more people, until the pool of susceptible people is exhausted or control measures are implemented.

Different types of contact which could be considered as potential transmission for different diseases form naturally a social network. Obviously, the network formed by different personnel contact can be very different in terms of densities and shapes.In this paper we consider all the networks which present the individual contacts are binary.

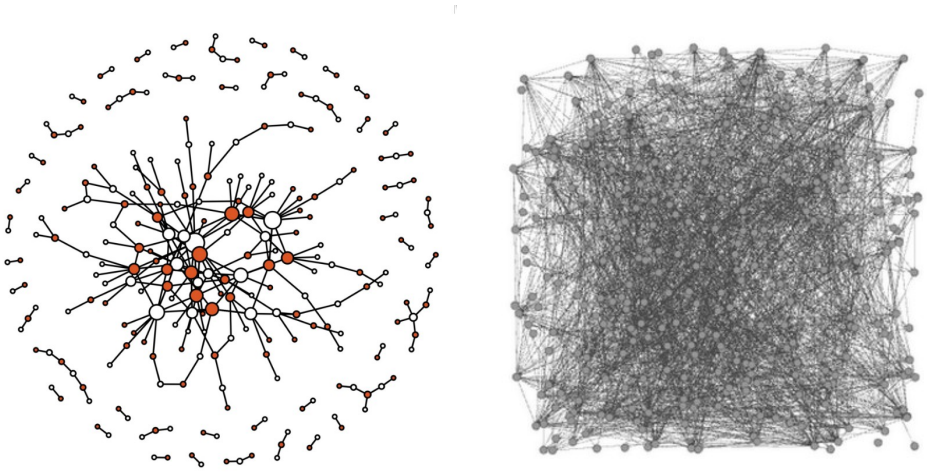


Figure 29: left: network of sexual contacts among teenagers in a high school in USA from immuno -epidemiology right: network of facebook connections, David Ellis from Lancaster University

The two networks in Figure 29 from the American Journal of Sociology and David Ellis research blog represent respectively the sexual contact in an American high school and interconnections among a group of people at facebook. The first can be considered as potential infections for diseases like AIDS while the last could be considered as potential infections for more propagated disease like Measles. One can easily observe that in the network of sexual behavior, the contacts are more isolated by small groups except a large group mainly created by individuals who are generally more active in sexual activities. Actually even in this central group, the density of edges is still sparse enough.Meanwhile, in the network of facebook connection, the graph is much more dense and random. furthermore, for diseases that can be spread by air the infection between two total strangers could also be possible. Thus the real potential infection network for Measles should even have more edges.

Despite the lack of realized numeric givens, we have tried to reconstruct the adjacency matrix of two networks presented in Figure 29 using a graph structure created by the same structure of networks.

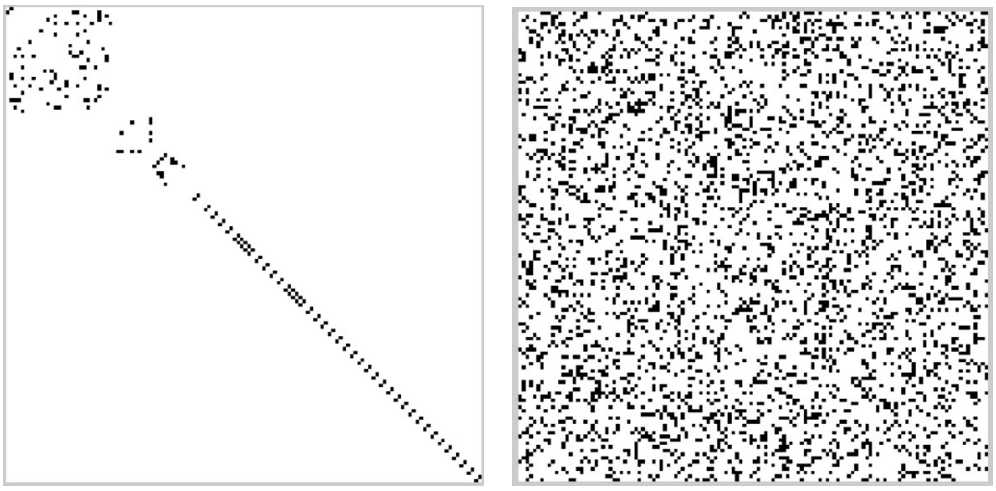


Figure 30: form of Adjacency matrices for networks given in Figure 29

Figure 33 shows that the difference between two networks is not only on density but also on the structure of connections. For the adjacency matrix present sexual contacts, it obviously owns a diagonal block structure including a lot of small closed loops connect two three people which can be considered as isolated part and could hardly provide any contribution to disease spread. We reminder that for a relatively larger community (such as a city or a country) there could be several large disjoint or fewly connected groups. However in the network of facebook, it is hard to separate the connections into small groups. It looks more like a whole block of random connections.

However, as explained before, the value of R_0 not only depends on the way of infection but also on other factors, such as the probability of infection on each contact. Obviously, a disease spread by air such as Measles has much dense potential contamination pair than the one spread by sexual contact such as HIV but the probability of infection of each pair is also much lower.Therefore, a disease spread with weak contact doesn't necessarily own a higher R_0.

A node in the social network could also present a community (ex. a city or a country). As people move frequently among countries, this type of network could also be used for disease spread. Different from the network of individuals, these networks are no longer binary, instead they are weighted by the quantity of exchanged population between two nodes.Here the connected graph of global air-flight transport of 227 nations has been created using the numeric givens found on the internet of a data sharing forum:

`https://github.com/gsmanu007/Complex-network-analysis-of-Airport-network-data/blob/master/airport_CnToCn.csv`

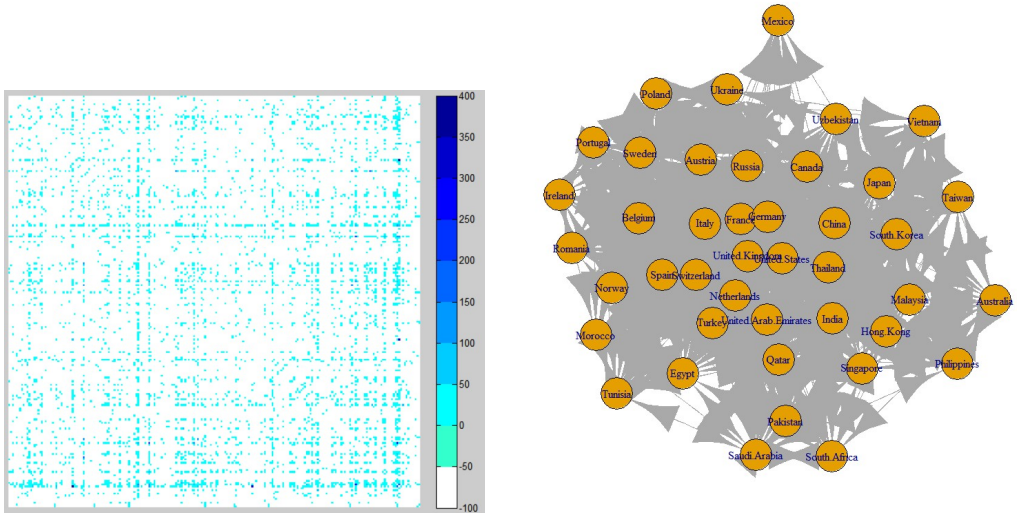


Figure 31: network of global air-flights

The left hand side of figure 31 shows the weighted adjacency matrix where the node (i, j) present the number of flies from country i to j. The white part shows no airlines between two countries.In order to make more contrast when drawing the adjacency matrix, we have replaced the value 0 by -100 in the matrix. Obviously, some nodes are more active than others. To have more clear vision, we have also showed the connection graph of 50 most active countries in the network on the right hand side.Most of them are well developed countries. This data set will be used later for simulations in section **6** of this chapter.

4 Epidemic model

Besides R_0, several other notations and quantities are widely used in network of disease spread.

- **S** : Susceptibles
- **I** : Infectives
- **R** : Recovered with immunity
- β : Contact rate at each period
- $\frac{1}{\gamma}$: Average infectious period
- **N** : Total population

SIR model

In this model created by [W. O. Kermack, A. G. McKendrick, 1927]. we neglected the probability of death for the infected. Once a person is infected, he/she will be cured after infectious period and then get immunity of the disease for ever.Therefore, the number of susceptible goes down when the disease spread.

As the total population $N(t) = S(t) + I(t) + R(t)$ stays as a constant, the dynamical system can be described as

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1}$$

An obvious case of equilibrium of system is when $I(t) \equiv 0$ which signs there are no initial patients.

Using the package $ode45$ in Matlab, we present in Figure 32 the evolution of $I(t)$

Concretely, we fixe $\gamma = 1, N = 100$, variate β from 0.1 to 10. (Therefore, R_0=β from 0.1 to 10) with initial conditions $N(t) = 100 = S(0), I(0) = 1$ which present the assumption that one percent of total population is illed at the beginning of the process.

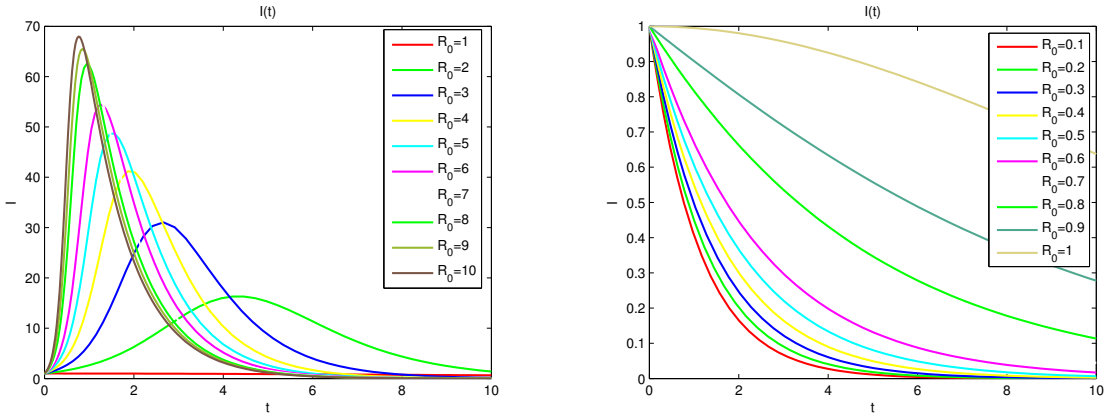


Figure 32: **SIR** model with different R_0

As explained before, when $R_0 >> 1$ we observe an explosion of disease at first stage. However, as the infected become recovered after the infectious period, the number of susceptible shrinks quickly. $I(t)$ thus converge to 0 in long term.

In fact, in the case that the disease has not exploded (which means the number of infected and recovered people is negligible before the whole population) $N(t) \approx S(t)$. This implies $R_0 \approx \frac{\beta}{\gamma}$. We denote $T = \frac{1}{\gamma}$ as the average infections period, thus $R_0 \approx \beta T$. Furthermore in the critical case of $\beta = \gamma$ (ie.$R_0 = 1$) the second equation in (6.1) becomes $\frac{dI}{dt} = (\beta \frac{S(t)}{N(t)} - \gamma) I \approx 0$

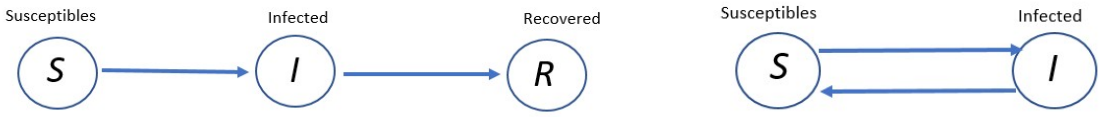


Figure 33: **SIR** and **SIS** model

SIS model

There are only two states in this model, the difference from SIR is that once a patient is cured, he/she will then become susceptible again which means this individual could be infected by same disease more than one time.
For the differential equations in the dynamical system, we only need to remove the part of recovered population and add them into susceptibles.

$$\frac{dS}{dt} = -\frac{\beta SI}{N} + \gamma I$$
$$\frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I$$

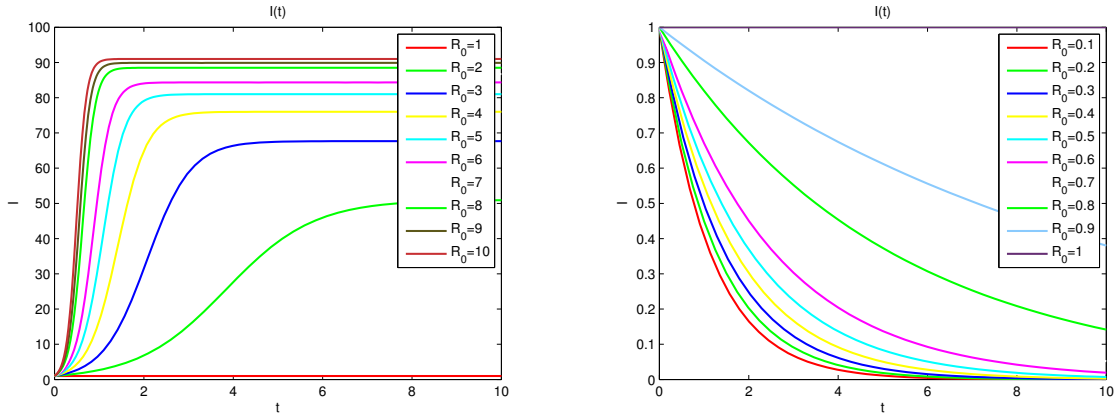


Figure 34: **SIS** model with different R_0

Different to the **SIR** model, since all infected person come back to the susceptible group, there will be no lack of "infect resources". The system equilibrium achieves at a strict positive value of $I(t)$ when $R_0 > 1$ is above the critical point.

There exists a lot of other epidemic models such as SIRS, SEIR whcih we will not discuss in this paper. In a real case of disease spreading, there should be other factors such as the rate of death. In this paper, we are specially interested in the first stage of propagation. The goal is to analyze which initial condition could bring an explosion of disease.

5 Galton–Watson process

The Galton–Watson process is a stochastic process firstly arose by [Francis Galton,1842] to explain the extinction of family names(Wikipedia). Suppose that family names are passed on to all male children by their father and the number of a man's sons to be a random variable distributed in $\mathbb{N}$. More importantly we suppose in addition that the numbers of different men's sons to be independent random variables.
Respecting all the assumptions above, the number of males sharing the same surname has been defined. The mathematical model can be described as:

$$X_0 = 1$$
$$X_{n+1} = \sum_{j=1}^{X_n} \zeta_j$$

Where $\zeta_j \in \mathbb{N}$ are all independent identically distributed variables.They represent the number of sons in the j^{th} family with this surname .

The extinction probability is defined as

$$Pr_{extinction} = \lim_{n \to +\infty} Pr(X_n = 0)$$

It is well known that $Pr_{extinction} = 1$ if $E(\zeta_j^{(n)}) < 1$ and $Pr_{extinction} < 10$ otherwise. In the case when $E(\zeta_j^{(n)}) = 1$ we are also almost sure to have the extinction of family name unless we are in the situation that each man has exactly one son.

This model sounds very similar to the disease spread problem where we can make an analogy between R_0 and $E(\zeta_j^{(n)})$ intuitively.

R_0=average number of secondary patients infected by one primary patient
$E(\zeta_j)$=average number of sons for each male human being

We are interested in the critical case when $E(\zeta_j) = 1$. Generally speaking, this condition shows that each male has one son in average. Imagine a society of n families where n is a large integer. As we know in average each male has one son, therefore the total number of males in the whole population should stay constant. However according to the result presented before, one has $Pr_{extinction} = 1$ as $E(\zeta_j) < 1$. This means family name will extinct in the society. In order to keep the population (male) constant, some families should have a very large number of descendants.

We have made a Monte Carlo numerical test to verify our intuition.
Suppose

- $\zeta_j \sim Poisson(1)$;
- n=10^5;

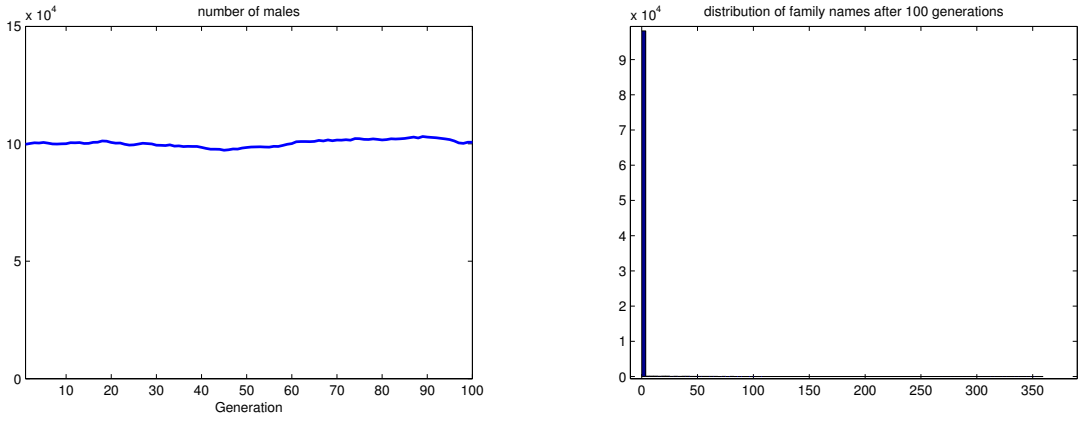


Figure 35: family names after 100 generations

From figure 35, one observe that during the whole period of simulation the number of males in the society remains almost constant. After 100 generations, over 99% of family names have extincted meanwhile the most prosperous family have 359 male members.

We would like to find a familiar phenomenon in disease spread. However the model of disease does not fulfill an important assumption in Galton–Watson process. The infections actually depend on the connections of individuals, therefore they can not be independent. One could understand this model as a "graph based Galton–Watson process".

6 Mathematical models and anlysis

In this section, we focus on the essential problematic: how can the structure of adjacency matrix influence the behavior of network diffusion (disease spread). These different structures present different spreading characteristics of diseases.
To do so, we keep the graph density constant (ie. same number of edges) and split our study into two different models which are named "fixed contact model" and "random contact model". In first model, we assume that the adjacency matrix of connection remains constant during the whole process of diffusion. Meanwhile in second, the adjacency matrix will be re-simulated at each step following the constraints of block structure. Which means people in same community see each other randomly with a even probability.
We split the problem into two different models which depend on the randomness of adjacency matrix.

- **fixed contact model**: We suppose the adjacency matrix stays constant during the whole period of spreading which means people always own the same contact.
- **random contact model**: In this model, the adjacency matrix changes its form at each generation of disease respecting the structure of communities.

6.1 description and assumptions

The contact of n persons in a social network is presented as a directed graph with adjacency matrix A. The transition time is supposed to be constant as 1 unit. The object is to check if the disease will spread explosively or eventually disappear.Therefore we start with a weak number of initial patient and suppose all the adjacency matrix of infection are sparse. The validity of the last assumption will be judged in section **6**.

An integer vector V_0 has been created to present the initial index(patient) in the social network (person who carry virus).

In the case when each node in the graph presents one individual

$$V_{0,i} = \begin{cases} 1 & \text{if person i is infected at time 0} \\ 0 & \text{otherwise} \end{cases}$$

In the case when each node in the graph presents a community (ex. a city or a country)

$$V_{0,i} \in \mathbb{N}^+ \quad \text{number of initial patients in community } i$$

According to the assumption of weak initial index, the number of initial patient should be negligible before the whole population n.

$$\sum_{i=1}^{n} V_{0,i} = o(n)$$

A function of four entries $next-generation(A, V_i, model, p, r)$ is defined in order to simulate the spread in one period.

- A: adjacency matrix of contact (We suppose that every contact can lead to an infection. If it is not the case, we can generate a matrix of infection using binomial laws and the contact matrix like we do in section **7** of this chapter)
- V_k: binary vectors present patient infected at k^{th} step for i from 0 until the present time
- $model$: which epidemic model being applied for the problem, we mainly consider three models: **SIR**, **SIS**,**SIRS**
- p: Infectious time (suppose to be a constant in this chapter,equal to ∞ in **SI** model)
- r: recovered time after being cured (only useful in **SIRS** model, for **SIR** model $r = +\infty$)

The output of function $next-generation$ is a binary vector of length n that present people being infected at this step of disease spread.

Therefore,

$$V_{k+1} = next-generation(A_k, V_i(\ i = 1...k), model, p, r)$$

All current infected at time k are presented by a vector I_k. We give an example of **SIS** model for readers to have a more clear vision.We remind that in this model, once the infected being cured, he/she will become again susceptible immediately. To simplify the notation, we introduce the symbol of positive part $()_+$ as

$$Q' = (Q)_+ \Longrightarrow \forall (i,j) \quad Q'_{i,j} = max(0, Q_{i,j}) \quad \text{for Q to be a matrix, vector or scalar}$$

Thus this model could be described as

$$I_k = \sum_{i=(k-p)_+ +1}^{k} V_i \quad \text{and} \quad V_{k+1} = A_k I_k \quad \Longrightarrow V_{k+1} = A_k(\sum_{i=(k-p)_+ +1}^{k} V_i) \tag{2}$$

Here I_k is a vector present current infectious person. **In the case when each node in the graph present one individual** I_k should be a binary vector as

$$\forall i \quad I_{i,k} = \max(\sum_{i=(k-p)_+ +1}^{k} V_{i,k}, 1)$$

The **SIR** and **SIRS** models are in fact more complicated as a part of population could be recovered from the disease. The recovered appears in the network when $k \geq r+1$. A vector R_k of length n has been defined to present the recovered at time k

Therefore, in **SIR** model we have

$$R_k = \sum_{i=1}^{(k-p-1)_+} V_i \quad \text{and} \quad V_{k+1} = A_k(\sum_{i=(k-p)_+ +1}^{k} V_i - R_k)_+$$

while in **SIRS** model only the definition of R_k is different

$$R_k = \sum_{i=(k-r-p)_+}^{(k-p-1)_+} V_i$$

Erdos-Renyi graphs:from contact to infection

The traditonal Erdos-Renyi graph is a one block stochastic matrix. Each pair of nodes in the matrix share a same probability P to be linked by one edge. Here we use the same idea and extend the Erdos-Renyi graph to a random matrix formed by a prescribed block structure on a reduced graph where the edges appear with probability $p_{in} \in [0,1]$ if they belong to the same prescribed block and $p_{out} \in [0,1]$ if not.We assume by default $P_{in} > P_{out}$

Here we present an example below for constructing an Erdos-Renyi random graph from a blocked structure reduced graph of three self connected clustering.Therefore, the adjacency matrix is also diagonal-block structured which is often the case of a social contact network.
If one fix $P_{out} = 1$ then the adjacency matrix generated by Erdos-Renyi method will be purely block diagonal structured which means there are three disjoint communities with no connections among them.

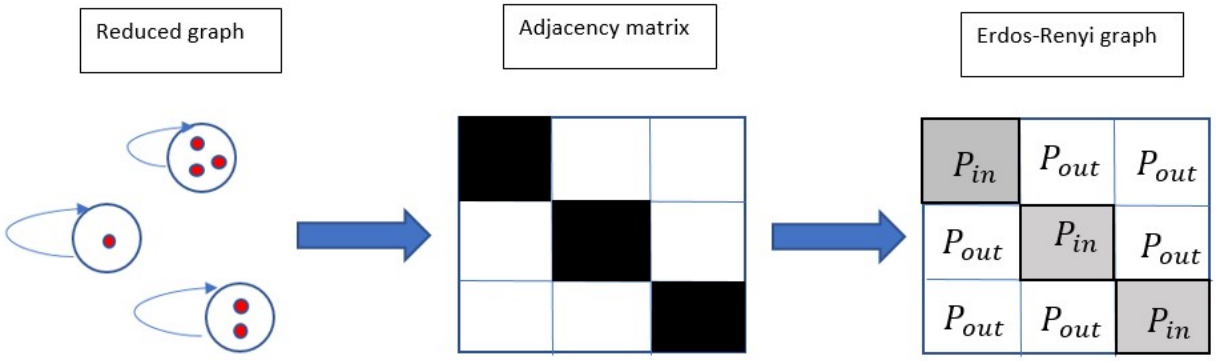


Figure 36: construct Erdos-Renyi graph from reduced graph

6.2 Fixed contact model

6.2.1 problematic

As explained in the previous part, in this model the adjacency matrix is considered as a constant, $A_k = A$. Using the Erdos-Renyi graphs presented in the end of the last section, we have constructed several numerical experiments

to reveal the importance of graph structure on the diffusion. We first limit ourselves to the 2×2 block structured matrices (of size 200×200 with density $P = 0.01$) with four equal size blocks presented in figure 37 below.

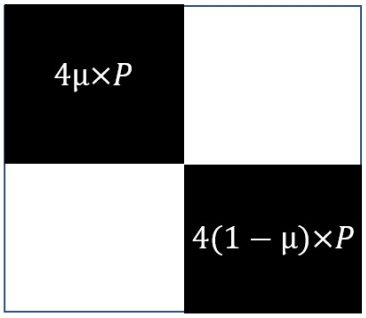


Figure 37: bock structure matrix where the densities are expressed by parameter μ

One can easily find that no matter how the value of μ changes the global density of the adjacency matrix stay invariant ($= P$). Which means the average degree of the associated graph is an invariable. As we have mentioned in the beginning of the chapter, R_0 can be seen as the average number of secondary infected caused by one primary infected. By this definition, the value of R_0 should be exactly the average degree of all nodes in the network. Therefore, the R_0 of all diseases (ie. whatever μ is) sharing the contact network in the form as Figure 37 should have the same R_0. Using the formulas in section (6.1), we have simulated 20 generations of infection for fixed matrices in a SIS model using different values of μ. When starting the simulation we suppose there are 5% (ie.10/200) initial infected distributed randomly for every simulation($sum(V_0) = 10$). We then draw the curve of infected after 20 iterations (ie.$sum(V_{20})$) in function of μ

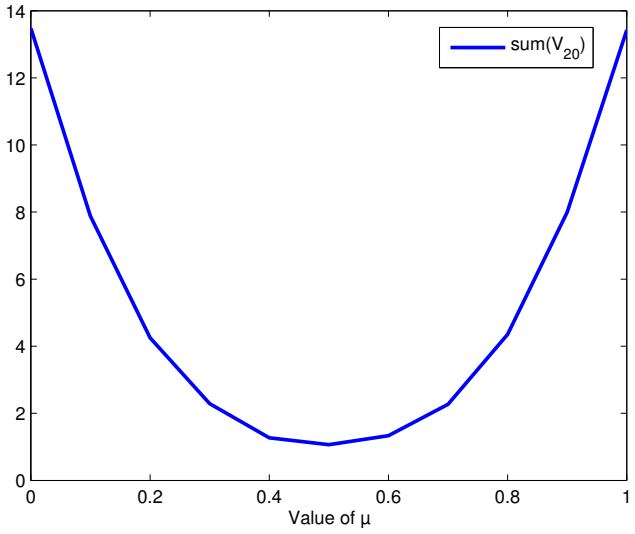


Figure 38: infected population size after 20 generations in function of μ obtained by monte carlo simulation for $n = 1000$

The figure 37 shows that although the R_0 value should probability stay constant when μ varying, there exists a significant difference on the result of spreading. The number of infected after 20 generations is much smaller

when the two blocks are homogeneous (ie. μ approach 0.5).
Here we present the same numerical experiment with a series of double symmetric (ie. the two diagonal (resp.anti-diagonal) blocks have always the same density).

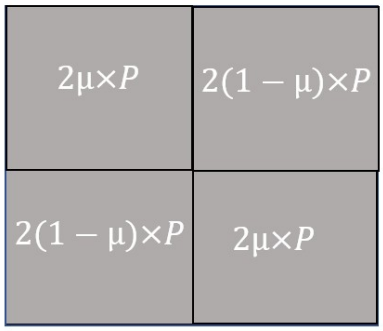


Figure 39: double symmetric structure

Similarly we draw the curve of infected number after 20 iterations in the network in function of μ. One observe from figure 40 that in this case the number of infected finally stay almost constant for all values of μ.

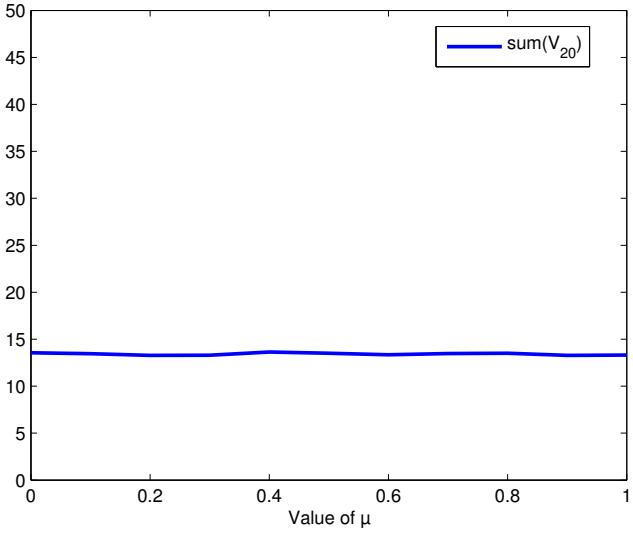


Figure 40: infected after 20 generations in function of μ

Why does that happen? Does the fact that the contact among people is more homogeneous slow down the diffusion?

To reveal the truth, we start by developing the expression (2) for A constant.

$$V_{k+1} = A^2\Big(\sum_{i=(k-p)_+ +1}^{k} \sum_{j=(i-1-p)_+ +1}^{i} V_j\Big) = A^3\big(\sum_i \quad \sum_j \sum_q \ldots\big)$$

It is obvious to see that the power of matrix A play a key role in the spread model. Especially the convergence of matrices power A^n depend on the dominant eigenvalue of A which is also known as spectral radius of matrix A.

6.2.2 Study of eigenvalues of binary matrices

Property 6.1. *For a directed graph of n nodes G, denote A as its adjacency matrix and $d_1, d_2....d_n$ as degree of each node. For $\lambda_1 = \rho(A)$ (ie.dominant eigenvalue of A), $d_{\max} = \max(d_1, d_2....d_n)$ and $d_{mean} = \frac{\sum_{i=1}^{n} d_i}{n}$, we have*

$$d_{average} \leq \lambda_1 \leq d_{max}$$

Proof. Using the property of Rayleigh quotient, the lower bound can be proved easily :

$$\lambda_1 = \max_x \frac{x^T A x}{x^T x} \geq \frac{1^T A 1}{1^T 1} = \frac{\sum_{i,j} A_{i,j}}{n} = \frac{\sum_i d_i}{n} = d_{mean}$$

For proving the upper bound, let's suppose v is the eigenvector associated with the dominant eigenvalue λ_1, so $Av = \lambda_1$. Let v_i be the largest component in vector v ie. $\forall j \in [1, n], v_j \leq v_i$.

$$\lambda_1 = \frac{(Av)_i}{v_i} = \frac{\sum_j A_{i,j} v_j}{v_i} \leq \frac{\sum_j A_{i,j} v_i}{v_i} = \sum_j A_{i,j} = d_i \leq d_{max}$$

□

As showing in the proof above, the eigenvalue of a adjacency matrix could also be understood as a sort of weighted average of nodes' degree.

In fact, as described in chapter 0, $\rho(A) > 1$ implies $A^n \longrightarrow \infty$ when $n \longrightarrow \infty$. Generally speaking when the spectral radius of adjacency matrix A is higher than 1, we are almost sure to have an explosion of disease. Therefore, the respective ratio R_0 is often defined by the $\rho(A)$.

remark [Richard.P.Stanley 1987] has proved a more precise upper bound for a symmetric 0-1 matrix (which is the case of most disease transmission)

[Richard.P.Stanley 1987]
The spectral radius $\rho(A)$ of the adjacency matrix A of a graph G with e edges satisfies $\rho(A) \leq \frac{1}{2}(-1 + \sqrt{1+8e})$. Equality occurs if and only if $\exists k \in \mathbb{N}$ such that $e = \binom{k}{2}$ and G is a disjoint union of the complete graph K_k and isolated vertices.

In fact the number of edges e could be considered as $n \times d_{average}$. Therefore, the theorem above shows that in a non directed graph, $\rho(A) = O(n \times d_{average})$. As we suppose the graph of contact is sparse, we are interested in the influence of density (number of ones in the adjacency matrix) on the spectral radius of a size-fixed graph. On the left part of figure 41, we present the curve of eigenvalues as function of the density for a sparse matrix of size 1000 × 1000 simulated using Monte Carlos method for 1000 experiments. On the right part we find the probability of the spectral radius is higher than 1.

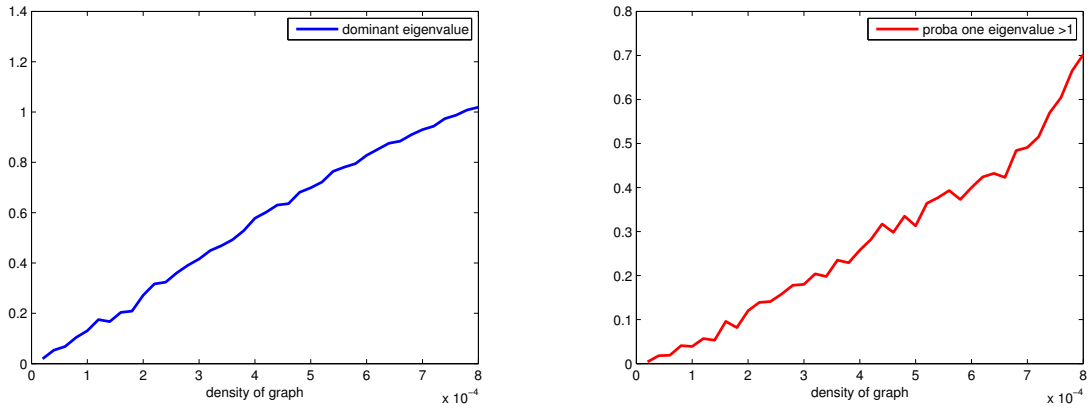


Figure 41: dominate eigenvalue of random binary matrix

One observes that the increasing of dominant eigenvalue is almost linear to the density of graph when the graph is sparse enough (ie. relatively low density). However the probability of having a bigger eigenvalue than one increases rapidly(convex function) with the augmentation of density.

6.2.3 block structured matrices

As mentioned in section 2 of the chapter, an important part of diseases owns a diagonal block structured matrix as potential infectious contact especially for the diseases spread by sexual contacts or drug injections like HIV. Each diagonal block in the adjacency matrix presents a community in social network. By intuition, we think the disease could spread very fast if infected appears inside a community. However it will be much more difficult for the disease to spread across the communities.

To illustrate this phenomenon, a 2×2 block matrix A has been constructed. We suppose the two communities in the network are with the same size.

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}$$

Here $A_{i,j}$ are sub-matrices of the same size that represent the link between two communities as presented in figure 42

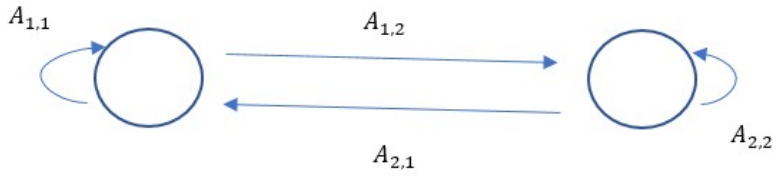


Figure 42: communities in graph G

Now we are interested in the spectral radius of these matrices by blocks and how the densities in different blocks could influence these eigenvalues.

Property 6.2. *The dominant eigenvalues(ie.spectral radius) of blocks $A_{1,1}$ (respectively $A_{1,2}, A_{2,1}, A_{2,2}$) are denoted as $R_{1,1}$ (respectively $R_{1,2}, R_{2,1}, R_{2,2}$). We have*

$$min(|R_{1,1}|+|R_{1,2}|, |R_{2,1}|+|R_{2,2}|) \leq \lambda \leq max(|R_{1,1}|+|R_{1,2}|, |R_{2,1}|+|R_{2,2}|)$$

Proof. Suppose the dominant eigenvalue of matrix A is λ with associated eigenvector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

Where $dim(x_i) = dim(A_{i,i})$ for $i = 1,2$
Therefore:

$$Ax = \lambda x \Rightarrow \begin{cases} A_{1,1}x_1 + A_{1,2}x_2 = \lambda x_1 \\ A_{2,1}x_1 + A_{2,2}x_2 = \lambda x_2 \end{cases}$$

$$\Rightarrow |A_{1,1}x_1| + |A_{1,2}x_2| \geq |\lambda x_1|$$

$$\Rightarrow \begin{cases} |R_{1,1}|||x_1|| + |R_{1,2}|||x_2|| \geq \lambda||x_1|| \\ |R_{2,1}|||x_1|| + |R_{2,2}|||x_2|| \geq \lambda||x_2|| \end{cases}$$

Without losing generality, one may suppose $||x_1|| \neq 0$, thus

$$\Rightarrow \begin{cases} |R_{1,1}| + |R_{1,2}|\frac{||x_2||}{||x_1||} \geq \lambda \\ |R_{2,1}| + |R_{2,2}|\frac{||x_2||}{||x_1||} \geq \lambda\frac{||x_2||}{||x_1||} \end{cases}$$

$$min(|R_{1,1}|+|R_{1,2}|, |R_{2,1}|+|R_{2,2}|) \leq \lambda \leq max(|R_{1,1}|+|R_{1,2}|, |R_{2,1}|+|R_{2,2}|)$$

□

Remark: λ reaches the maximum when the sub-vectors x_1 and x_2 are co-linear. This is obviously the case of structure 2-2, as the null vector is co-linear to all vectors which makes in practice the disease spread more efficient for the transition matrix of type 2-2 than type 1. This phenomenon will be illustrated later by numerical experience.

When the matrix is in purely diagonal block structure as the case of a lot of infectious contact, the reduced adjacency matrices of blocks can be written as:

$$A = \begin{pmatrix} A_{1,1} & 0 \\ 0 & A_{2,2} \end{pmatrix}$$

This matrix could be seen as the next-generation matrix for two independent communities (ex.two cities far away from each other). In this case it is easy to see $\rho(A) = max(\rho(A_{1,1}), \rho(A_{2,2}))$. However the behavior of disease spread will heavily depend on the initial patients. For example, if there is no initial patient in one community, its population will never be infected no matter the value of R_0 is.

In some specific situation, the block structure of adjacency matrix could also be anti-diagonal.[J.Holland,2007] give an example for a sexually transmitted disease in a completely heterosexual population. Define f as the expected number of infected women and m as the expected number of infected men given contact with a single infected member of the opposite sex in a completely susceptible population.
In this case,

$$A = \begin{pmatrix} 0 & f \\ m & 0 \end{pmatrix}$$

with one community of all males and another of all females. Obviously the disease could only spread across communities. We assimilate m and f as the eigenvalues of two anti-diagonal blocks. Therefore, $R_0 = \rho(A) = \sqrt{mf}$.

Keeping the global density of adjacency constant, we would like to further check how the distribution of edges in each block could influence the value of R_0

We still limit ourselves in the case of 2×2 block structured matrices with two communities of same size. We further assume that all four blocks are for the same size. The constant density (number of edges over number of all pairs) is denoted as P which means the average density of four sub-matrices $A_{1,1}, A_{1,2}, A_{2,1}, A_{2,2}$ should be $P_i n$.Using the result in figure 41, we mix the two definitions of R_0 (ie.eigenvalue of adjacency matrix and average number in secondary infection). Therefore,

$$R_0 = \rho(A) = \rho\left(\begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}\right) = \rho\left(\begin{pmatrix} R_{1,1} & R_{1,2} \\ R_{2,1} & R_{2,2} \end{pmatrix}\right) \quad \text{with} \quad \frac{R_{1,1} + R_{1,2} + R_{2,1} + R_{2,2}}{4} = P$$

the problem become simply evaluating the dominant eigenvalues of a 2×2 matrix.One can easily get:

$$\rho(A) = \frac{R_{1,1} + R_{2,2}}{2} + \sqrt{\left(\frac{R_{1,1} + R_{2,2}}{2}\right)^2 - (R_{1,1}R_{2,2} - R_{2,1}R_{1,2})}$$

We further assume that $R_{1,1} = R_{2,2} = a$ and $R_{1,2} = R_{2,1} = b$, the expression above becomes simply:

$$\rho(A) = a + b = 2P$$

This result shows that theoretically no matter the distribution of density of each block, when the adjacency matrix is symmetric in both diagonal and anti-diagonal directions (ie $A_{1,1} = A_{2,2}$ and $A_{1,2} = A_{2,1}$) the value of R_0 will always stay the same. Therefore, the speed of disease spread should remain the same for any adjacency matrices with the same density and the constraints of symmetry.

Now we consider another scenario where we have two isolated communities which means $A_{1,2}$ and $A_{2,1}$ are null matrices. However we remove the condition that the matrix is anti-diagonal symmetric. Thus

$$R_\mu = \begin{pmatrix} 4\mu P & 0 \\ 0 & 4(1-\mu)P \end{pmatrix}$$

$$\rho(A_\mu) = \rho(R_\mu) = max(4\mu P, 4(1-\mu)P)$$

To observe more clearly the different evolution of infected, we have chosen several block structures and drawn their curves of evolution.

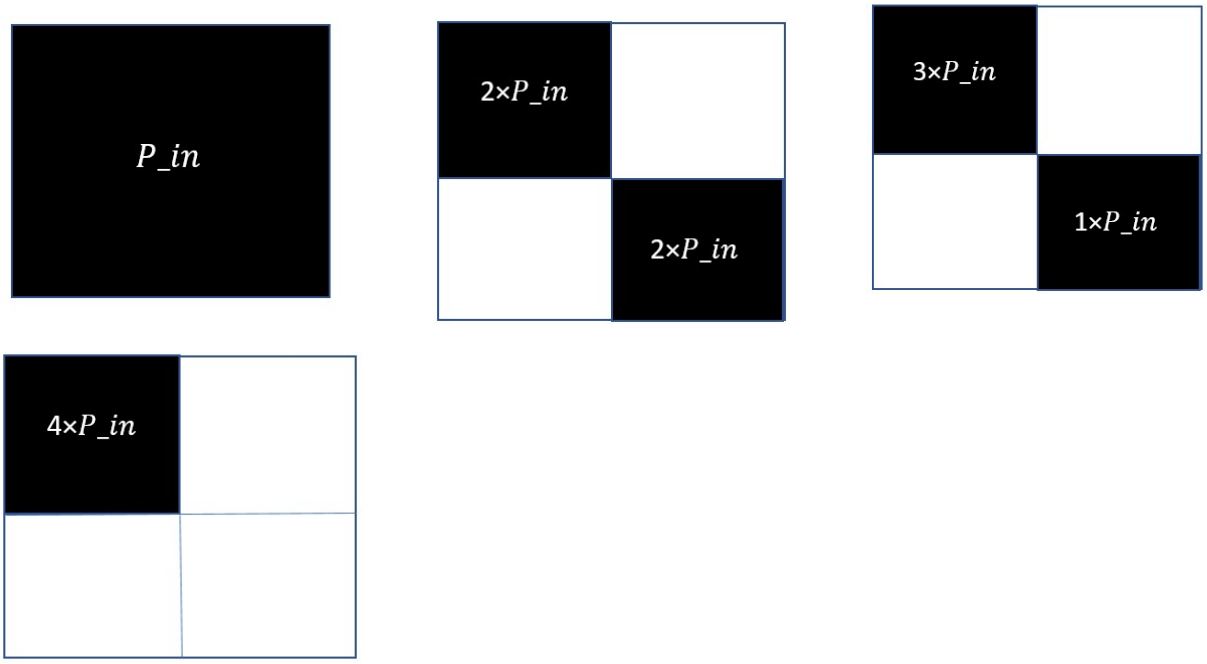


Figure 43: matrices structure 1 ,2-2,3-1,4-0 (in order)

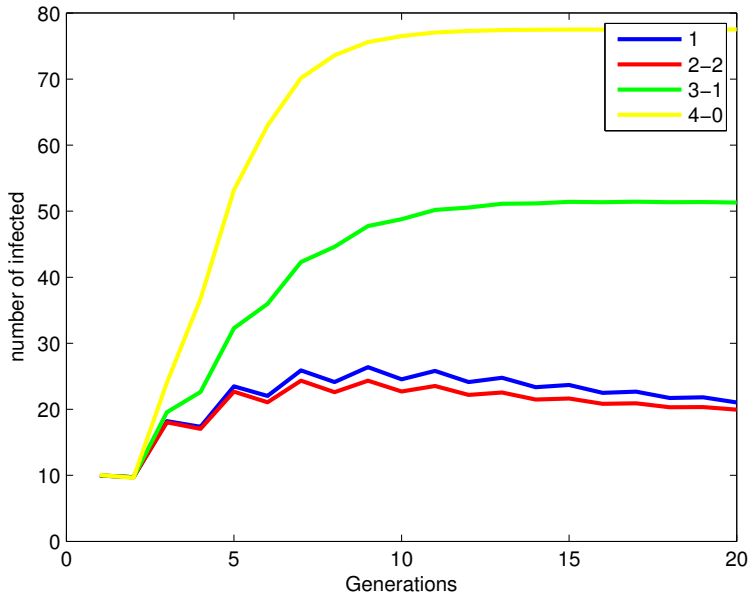


Figure 44: evolution of infected for 20 generations using Monte Carlo method for $n = 1000$

The simulation presented in figure 44 is made for the case when each node in the graph present one individual for global density $P = 0.01$ in a adjacency matrix of size 200×200 with different block structure. Therefore the vectors v_i are all binary during the whole simulation. One observes that the block structure $2-2$ and block structure

1 almost share the same evolution curve while the block structure $3-1$ and $4-0$ are much higher. These results are in coherence with the study of dominant eigenvalues and therefore proves the validation of the spectral radius as R_0 when the adjacency matrix of connection is available.
The commented Matlab code of simulation (only for structure 1-1) is shown below.

```
n=1000; % monte carlo simulation
p_out = 0;
B1 = [1 0 ; 0 1];%structure of reduced graph
N = [100,100];%distribution of nodes in each block
p=2;%infectious time
node=sum(N);%number of nodes in network
evolution2=zeros(1,20);

 p_in=0.01;

for j=1:n
        [A Sim] = graphUD (B1,N,p_in,p_out);%simulate erdos-renyi graph
        A=full(A);
        V0=binornd(1,0.05,1,200);%set initial V_0
        V=zeros(200,20);% V is a matrix where each colmun present V_i
        V(:,1)=V0;

        for u=1:19
            start=max(u-p,0)+1;
            vv=zeros(200,1);
            for i=start:u
                vv=vv+V(:,i);

            end

         V(:,u+1)=A*vv;
         V(:,u+1)=min(V(:,u+1),1);% each component of V_i should be binary (in the case
             when each node present one individual)
        end
       S=sum(V);
         evolution2= evolution2+S;
end
 evolution2= evolution2/n;
```

We would like compute the R_0 value defined by next-generation matrix(ie.$K_{k,m}$) of the four differently structured network to see if it is coherent with the result of simulation. We than obtain the estimation of dominant eigenvalues by making a Monte Carlo simulation for 1000 iterations in each case.

matrix structure	dominant eigenvalue of next-generation matrix
4-0	3.9275
2-2	1.9743
1	1.9686
3-1	2.7990

One observes that the order of R_0 defined by next-generation matrix is in coherence with the curves in figure 44. However all the values obtained are much higher than 1, but we can see in figure 44 the evolution curves of the structure 1 and 2-2 remains stable without an evident explosion. In this example, the R_0 defined as the dominant eigenvalue of the next-generation matrix is not perfect in terms of being the threshold of disease explosion.

However we discover an interesting point is that the R_0 defined by the next-generation matrix is very close to the highest average degree of two communities. As the size of adjacency matrix is 200×200 and we keep the global density $P = 0.01$, it is easy to calculate the highest average degree among two communities. The results are presented in the table below:

matrix structure	highest average degree among two communities
4-0	$100 \times 0.04 = 4$
2-2	$100 \times 0.02 = 2$
1	$200 \times 0.01 = 2$
3-1	$\max(100 \times 0.03, 100 \times 0.01) = 3$

Compare the values in the two tables above, we see they are similar. We have done other simulations and the two groups of values obtained are always close enough to each other. Thus, the R_0 obtained by the next-generation matrix is nearly equivalent to the average degree of the most condensate community in the network when the graph is simulated using Erdos-Renyi method.

To compare the results with **SIS** dynamical system presented at the beginning of the chapter, we have constructed a system of equal size (ie. n=200) with $\beta = 2$ and $\gamma = 1$ in figure 45. Thus the reproductive ratio R_0 is equal to two.

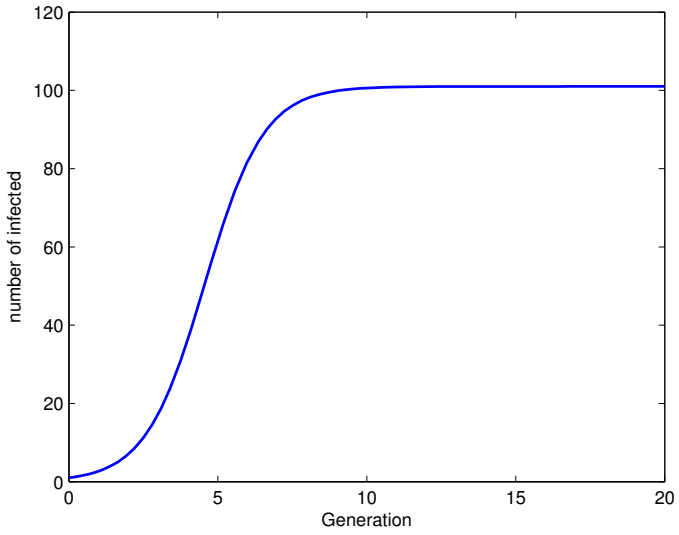


Figure 45: evolution of infected simulated by dynamical system for n=200 and $R_0 = 2$

One observe that the curve simulated by **SIS** dynamical system is always above the four curves of simulation in figure 44. In fact, as we have forced the vectors V_i to be binary, we have eliminated a certain number of infected by doing $V(:, u+1) = min(V(:, u+1), 1)$ (line 27 in Matlab code) these results are not literally compatible.

In order to properly compare the results with the dynamical system , we remove the condition that vectors V_i are binary. One can imagine that each node in the network present the number of infected in a city, therefore it

could be much larger than one. In this case, the connection network present geographical human transport. We also eliminate the upper bound of population (original 200) in the dynamical system by setting $N = 1000000$.

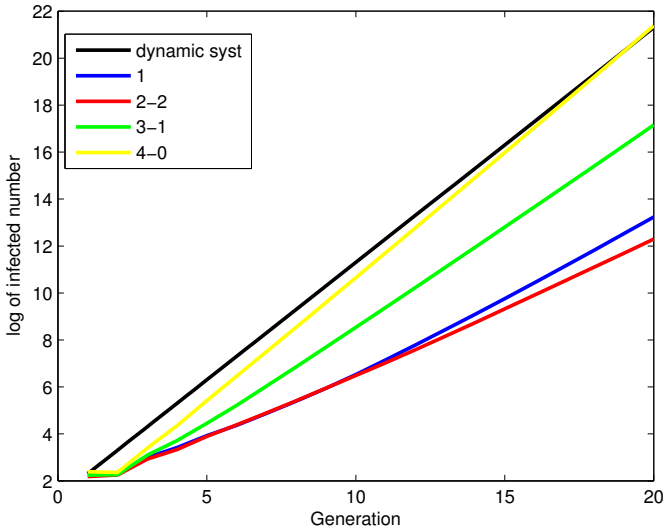


Figure 46: comparison of evolution curve on logarithm

From figure 46, we see that the curve simulated when $R_0 = 2$ almost coincide with the structure $4-0$. These results are in coherence with the values of dominant eigenvalues simulated by Monte Carlo methods for 10000 experiments as presented in table below.

structure	spectral radius
4-0	1.9951
2-2	1.1329
1	1.0097
3-1	1.4896

6.3 Random contact model

Now we focus on the random contact model where the adjacency matrix is not constant during the epidemic propagation. How will this assumption influences the behavior of diffusion?
The main difference from the fixed contact model is that one has no longer the power of matrices. Instead we get a product of several different matrices in a row ie.

$$V_{k+1} = A_1 A_2 (\sum_{i=(k-p)_+ +1}^{k} \sum_{j=(i-1-p)_+ +1}^{i} V_j) = A_1 A_2 A_3 (\sum_i \sum_j \sum_q ...)$$

In this model, analyzing the eigenvalues is not useful anymore as the dominant eigen-vector of matrices A_i are not co-linear. Therefore, the convergence of product $A_1 A_2 .... A_n$ will not depend on the spectral radius of each matrix.
As for the fixed contact model, we start by making a numerical experiment to have a direct vision.
Similar to figure 38, we draw the curve of infected number after 20 generations when the adjacency matrices following the structure presented in figure 37. We remind here that in the simulation we re-simulate adjacency matrices identically with same block structure at each generation.

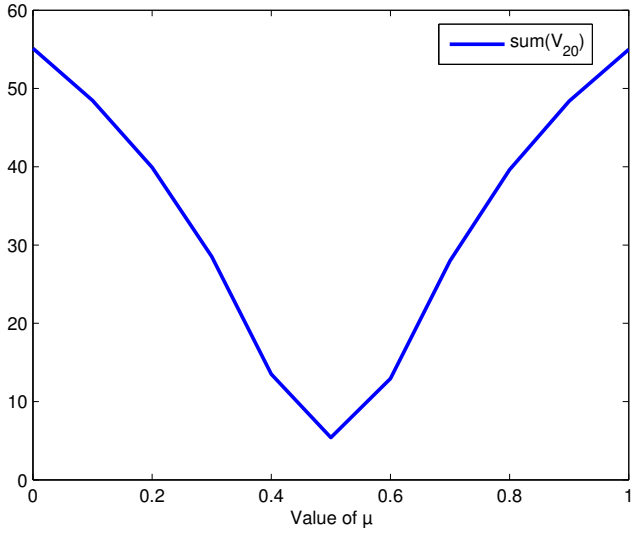


Figure 47: infected after 20 generations in function of μ obtained by monte carlo simulation for $n = 1000$

Comparing the curves in figure 47 to figure 38, we see that in the random contact model the function of final infected on parameter μ stays convex. That means this function also reaches its minimum when the density is homogeneously distributed in two diagonal blocks. However, it is obvious that the difference created by the change of parameter μ is more evident in the random contact model (ie. figure 47). It shows that the structure of non-homogeneous blocks own more advantage in terms of accelerating the diffusion when we are in the random contact model.
To make further analysis, we have to study concretely the product of block structured matrices simulated by Erdos-Renyi method.

Property 6.3.

$A_1, A_2, ..., A_m$ are m matrices of size $n \times n$ simulated by Erdos-Renyi method for one block of a fixed density P_{in} (as presented in figure 48). We further assume that $m = o(n)$ and all these matrices are sparse which means the number of edges in the network should be in the same quantity of number of nodes m. Therefore $P_{in} = O(\frac{1}{n})$.
Denote $P_{in} = \frac{\lambda}{n}$, with $\lambda = o(n)$.
Then we must have for matrix $P = A_1 A_2 ... A_m$

$$E(||P||_F) = n\lambda^m$$

Where $||P||_F$ design the Frobenius norm of matrix P. This result shows that when the number of generation m is small(for example set $m < 20$), the matrix P is still sparse as $\lambda^m = o(n)$

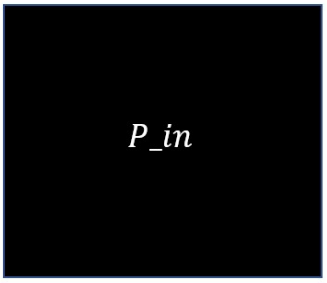


Figure 48: structure 1

Proof. We make this proof by induction.
Starting with the case of production of two matrices

$$P_2 = A_1 A_2 \Rightarrow P_{2,(i,j)} = \sum_{i=k} A_{1,(i,k)} A_{2,(k,j)}$$

As the density of A_i is distributed homogeneously $\quad \forall i, j, P(A_{1,(i,k)} = 1) = P(A_{2,(k,j)} = 1) = \frac{\lambda}{n}$

One can deduce that when $n \longrightarrow +\infty \quad P(P_{2,(i,j)} = 0) = (1 - \frac{\lambda^2}{n^2})^n = exp(n \times ln(1 - \frac{\lambda^2}{n^2})) \approx exp(-\frac{\lambda^2}{n}) \approx 1 - \frac{\lambda^2}{n}$

$$\text{Thus} \quad P(P_{2,(i,j)} \geq 1) = \frac{\lambda^2}{n}$$

$$\text{In fact} \quad P(P_{2,(i,j)} \geq 2) = x = \binom{n}{2} (\frac{\lambda^2}{n^2})^2 (1 - \frac{\lambda^2}{n^2})^2)^{n-2} = o(P_{2,(i,j)} \geq 1)$$

$$\text{Similarly} \quad (P_{2,(i,j)} \geq 3) = o(P_{2,(i,j)} = 2), \quad (P_{2,(i,j)} \geq k) = o(P_{2,(i,j)} = k-1) \quad \text{for } k = 2...m$$

$$\text{Therefore,} \quad E(||P_2||_F) = P(P_{2,(i,j)} \geq 1) \times n^2 = \frac{\lambda^2}{n} \times n^2 = n\lambda^2$$

This result shows that P_2 is almost binary of density $\frac{\lambda^2}{n}$ homogeneously distributed (ie.$\forall i,j \quad P_{2,(i,j)}$ are equally distributed)
Now suppose $E(||P_{m-1}||_F) = n\lambda^{m-1}$ with $P_{m-1} = A_1A_2....A_{m-1}$ so $P_m = P_{m-1}A_m$

$$\text{Using similar reasoning} \quad P(P_{m-1,(i,j)} = 0) = (1 - \frac{\lambda^m}{n^2})^n = exp(n \times ln(1 - \frac{\lambda^m}{n^2})) \approx exp(-\frac{\lambda^m}{n}) \approx 1 - \frac{\lambda^m}{n}$$

$$\text{We can easily deduce} \quad E(||P||_F) = \frac{n}{2} \times \lambda^m \times 2 = n\lambda^m$$

□

As the density of matrices' product growing with the power of λ, **we may define the reproductive ratio of the system as $R_0 = \lambda$ where λ is the average degree of nodes in the network**. However this definition is only valid for matrices in form of figure 48

How about if we split the population into two isolated communities of equal size as in figure 37?

$$\text{Denote} \quad A_i = \begin{pmatrix} AA_{i,1} & 0 \\ 0 & AA_{i,2} \end{pmatrix}$$

$$\text{Here} \quad dim(AA_{i,1}) = dim(AA_{i,2}) = n \quad \text{with} \quad density(AA_{i,1}) = 4\mu P \quad \text{and} \quad density(AA_{i,2}) = 4(1-\mu)P$$

As we keep the global density as a constant P correspond to an average degree λ, the average degree of nodes in sub-matrix $AA_{i,1}$ is $4\mu\lambda$ while the one in $AA_{i,2}$ is $4(1-\mu)\lambda$.

$$\text{Using the fact that} \quad A_iA_j = \begin{pmatrix} AA_{i,1} & 0 \\ 0 & AA_{i,2} \end{pmatrix}\begin{pmatrix} AA_{j,1} & 0 \\ 0 & AA_{j,2} \end{pmatrix} = \begin{pmatrix} AA_{i,1}AA_{j,1} & 0 \\ 0 & AA_{i,2}AA_{j,2} \end{pmatrix}$$

$$\text{we can easily deduce that} \quad P = \begin{pmatrix} P_{1,1} & 0 \\ 0 & P_{2,2} \end{pmatrix} = \begin{pmatrix} \prod_{i=1}^m AA_{i,1} & 0 \\ 0 & \prod_{i=1}^m AA_{i,2} \end{pmatrix}$$

Applying directly the property 5.3, we get $E(||P_{1,1}||_F) = \frac{n}{2}(4\mu\lambda)^m$ and $E(||P_{2,2}||_F) = \frac{n}{2}(4(1-\mu)\lambda)^m$.

$$\text{Therefore,} \quad E(||P||_F(\mu)) = \frac{n}{2}(4\mu\lambda)^m + \frac{n}{2}(4(1-\mu)\lambda)^m$$

This function is obviously convex, it reaches the minimum when $\mu = \frac{1}{2}$. **We recall this result is only valid under the assumption that the adjacency matrices A_i are sparse.**

Using the result above, we could explain the phenomenon observed in figure 47. The R_0 of the system should be the largest average degree in two diagonal blocks. Therefore, the more homogeneous the two blocks are, the smaller R_0 it owns.

Similar to the example for constant contact model in figure 44, we have also drawn the evolution curves according to different block structures presented in figure 75.We reminder that these four block structures share the same global graph density. We have chosen $n = 200$(number of nodes), $sum(V_0) = 10$(number of initial patients randomly distributed),$P = 0.01$(global density) as parameters and initial values.

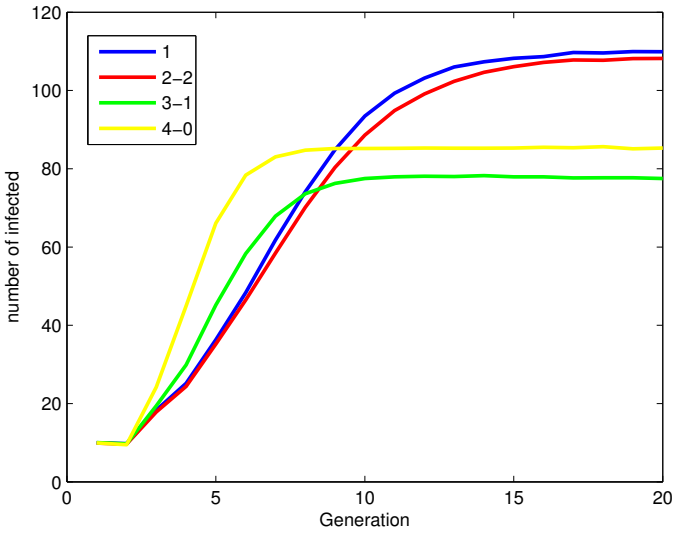


Figure 49: evolution of infected for 20 generations

From figure 49, one observes that same as the fixed contact model the less homogeneous structures $4-0$ and $3-1$ actually accelerate the diffusion speed. However the equilibrium level of these two structures are relatively low. The reason could be the weak number of infected in the community of small density.This result proves the validity of defining R_0 as the average degree in the network when using the random contact model.

In figure 51, we remove the constraint of binarities for V_i, thus each node in the network present a community instead of an individual. We would like to compare the curves obtained by the four different matrix structures with the one simulated by the dynamical system shown in section 3 of this chapter.

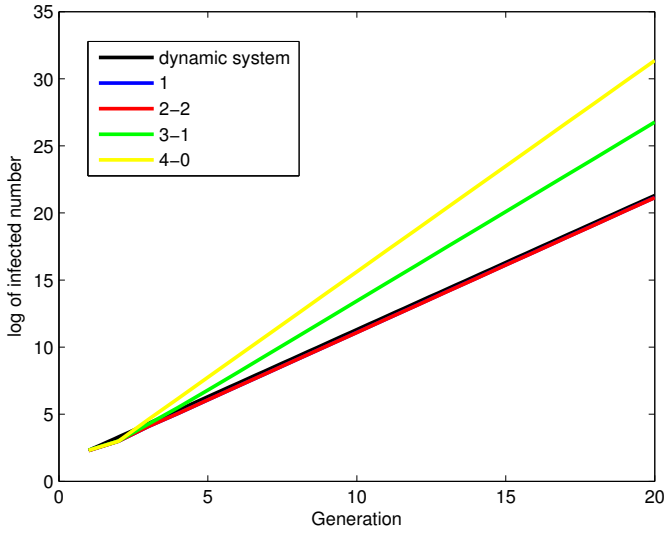


Figure 50: comparison of evolution curve on logarithm

One observes that the curve of dynamical system coincide with the ones of structure $2-2$ and 1. We explained in the section 5.2 that the reproductive ration R_0 could be translated as the average degree of nodes in the network. In the structures $2-2$ and 1, the density are homogeneously distributed in each community and as we re-simulate the adjacency matrix at each step, all nodes in these two graphs are symmetric thus we find the same evolution curve as the dynamical system.

6.4 Comparison of two models

In this section, we would like to compare the behavior of diffusion using the fixed contact model and random contact model. To do so, we fix all the other parameters in the **SIS** system such as $n = 200$, $P_{in} = 0.05$, infectious time $p = 2$, same graph structure (structure 1). We have used the Monte Carlo method to simulate the evolution of infected numbers in both models with the same initial condition.The Matlab code of simulation is available in Appendix.
We recall that the reproductive ratio in these two models are different.

$$\begin{cases} R_{0,\text{fixed contact}} \in [d_{\text{average}}, d_{\max}] \\ R_{0,\text{random contact}} = d_{\text{average}} \end{cases}$$

Here d_{average} denotes the average degree of nodes in the network while $d_{\max}$ is the maximum degree.

By intuition, we have the feeling that the fixed contact model should spread more rapidly than the random contact model as the value of R_0 is relatively higher under same condition.

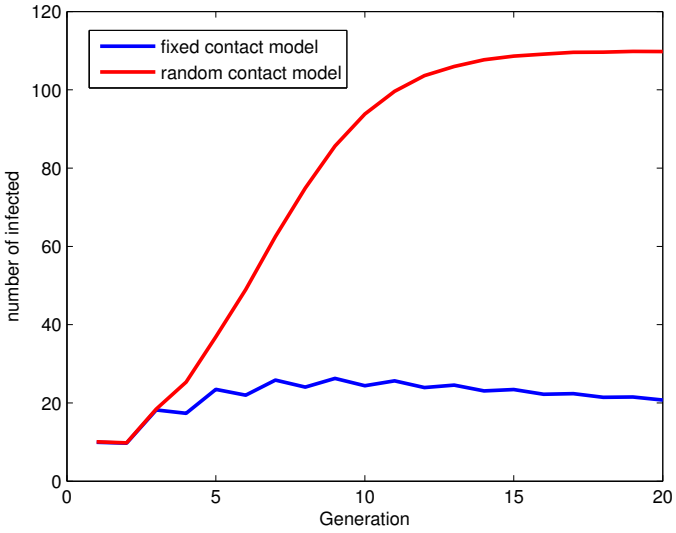


Figure 51: evolution curve of two models

However in figure 51, one observes the opposite of our prediction, why does this happen?
We then draw the distribution of final infected number of 20 generation in both models as presented in figure 52 and figure 53. We see that the one simulated by fixed contact model owns a much larger variance. In fact the rate of disease extinction is very large (almost 30 %)

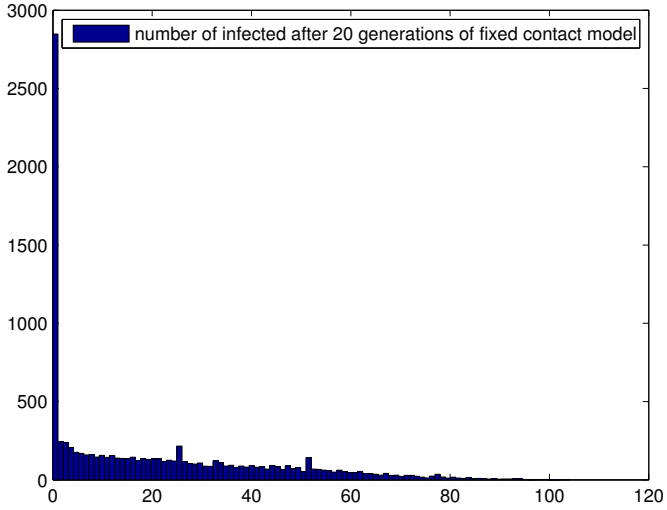


Figure 52: Distribution of infected number in fixed contact model, the horizon axis represent the number of infected after 20 generations the vertical axis represent number of simulations reach this final infected number

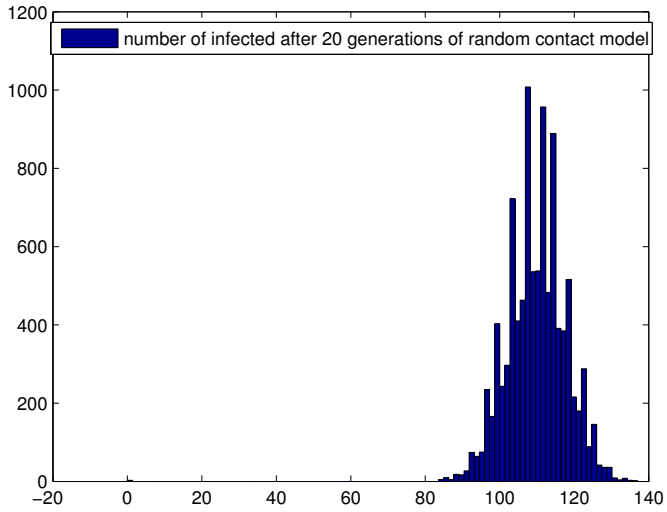


Figure 53: Distribution of infected number in random contact model,the horizon axis represent the number of infected after 20 generations the vertical axis represent number of simulations reach this final infected number

To explain this phenomenon, we will need the next theorem invented by [Harold M. Hastings, 1982].

Theorem 2.

The MAY-WIGNER stability theorem for connected matrices

Let A be the $n \times n$ adjacency matrix of a graph with n^2C $(0 < C < 1)$ binary entries, each chosen independently. Then the graph is asymptotically almost surely connected if

$$C \geq (1+\epsilon)\frac{log(n)}{n}$$

and asymptotically almost surely not connected if

$$C \leq (1-\epsilon)\frac{log(n)}{n}$$

The proof of this theorem is based on the famous May-Wigner stability theorem.
In fact, as we suppose the network is always sparse (ie. $C = P_{in} = O(\frac{1}{n}) << \frac{log(n)}{n}$), the graph is asymptotically almost surely not connected after the MAY-Wigner stability theorem. This result shows that when the fixed contact model is used, we are almost sure that there exist sub-communities inside the network simulated by Erdos-Renyi method.
To illustrate this phenomenon, we take a 200×200 network of block structure $1-1$ simulated by Erdos-Renyi method.

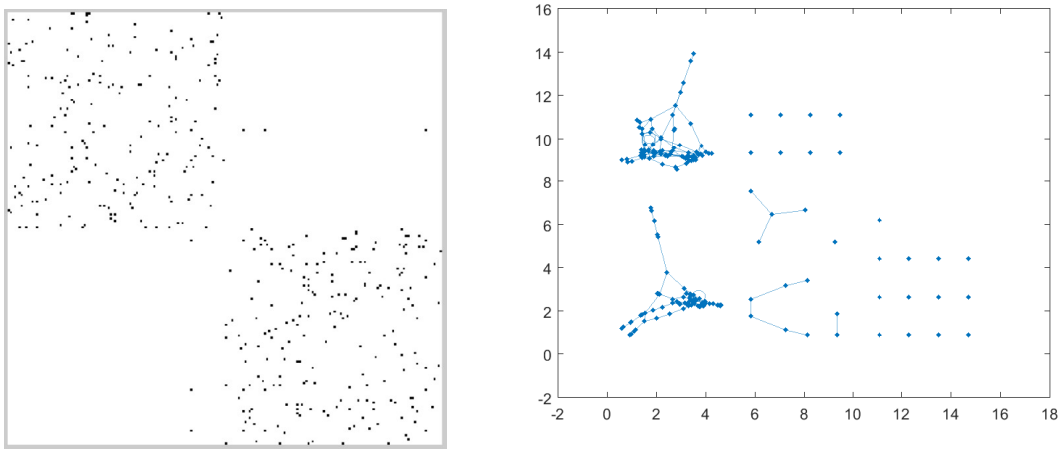


Figure 54: example of a graph simulated by Erdos-renyi

Figure 65 presented a network of diagonal block structure with $P_{in} = 0.02$ and $P_{out} = 0$. However using the graph processing pakage in Matlab, one can see there actually exists more than two communities in the network. The the existence of sub-community could influence heavily the disease spread. For example, if there is no initial patient in one sub-community there will never be any infected in this community as it is isolated from the outside.

On the other hand, the existence of sub-community is no longer possible in the random contact model as we re-simulate the adjacency matrix at each step

We have also simulated the case when each node in the graph present one community instead of one individual. As have mentioned several times in this chapter for doing this, it is enough to remove the condition that V_i are

binary vectors. Here the V present the number of infected in each community. To simplify the model, we put no population upper bound at each community. The other parameters and initial conditions remain the same.

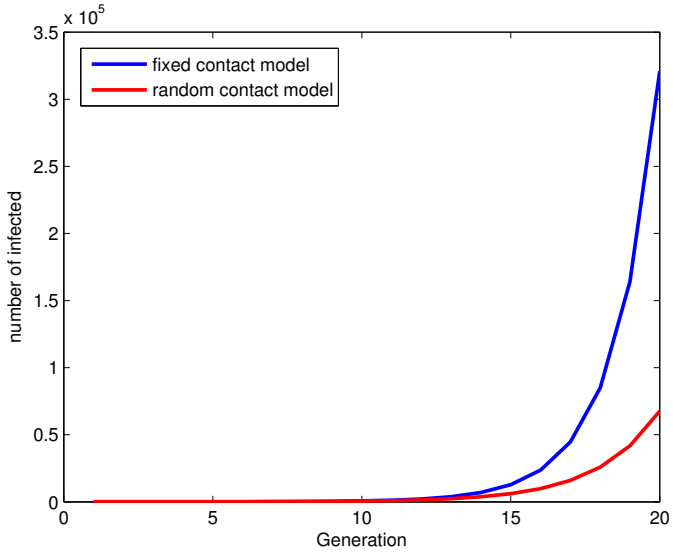


Figure 55: number of infected in random contact model

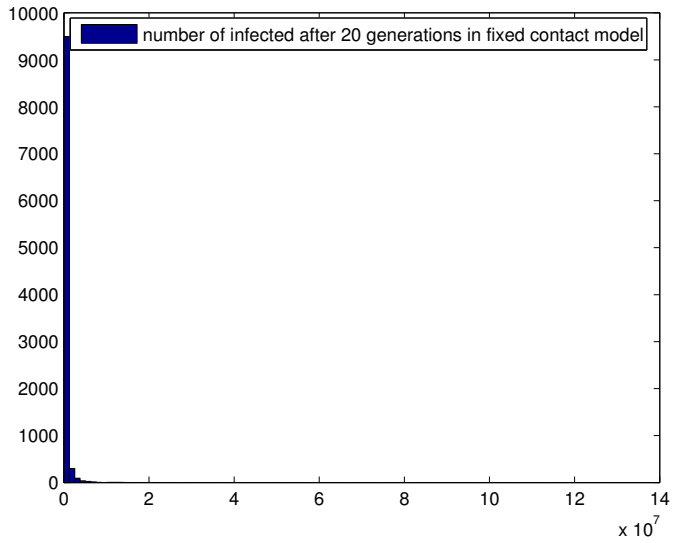


Figure 56: number of infected in random contact model

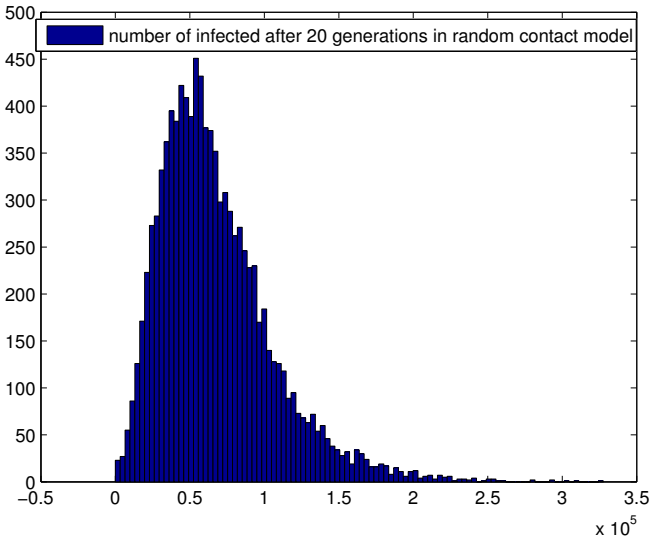


Figure 57: number of infected in random contact model

From figure 55 one observe that in this case the diffusion in the fixed contact model is much faster than in the random contact model. This result is coherent to the value of R_0. **Why the fixed contact model goes above the random one when we remove the condition of binarity ?**
To answer this question, we have drawn the distribution of final infected numbers (only strict positive) in both fixed and random contact models when each node present one individual or one community.

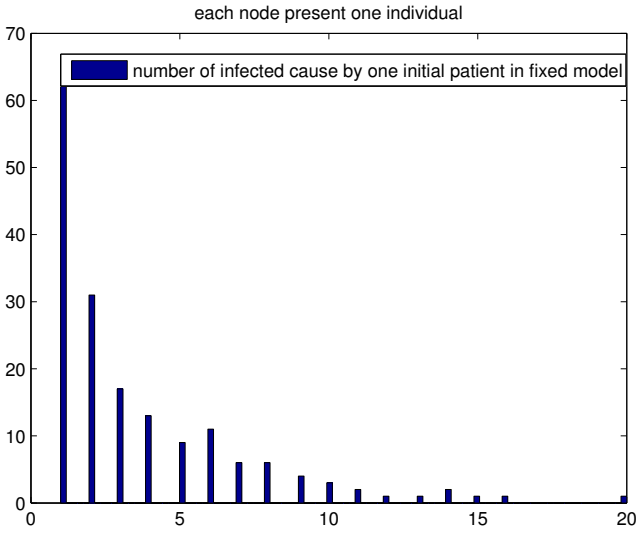


Figure 58: distribution of nonzero infected number after 20 generations in fixed contact model when each node present one individual

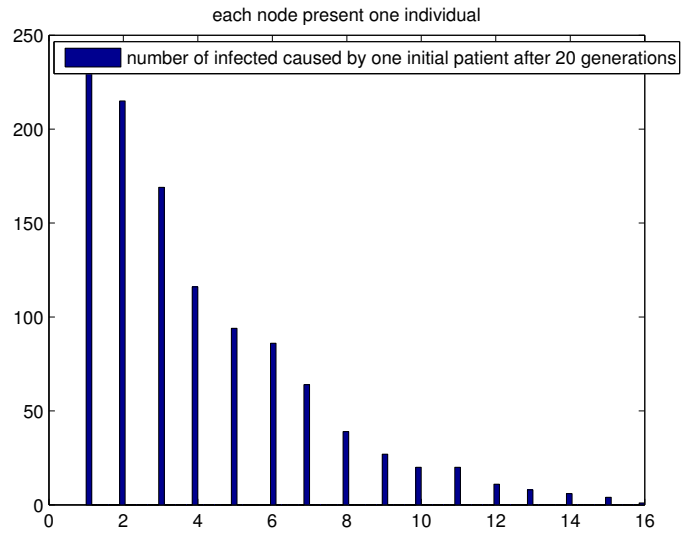


Figure 59: distribution of nonzero infected number after 20 generations in random contact model when each node present one individual

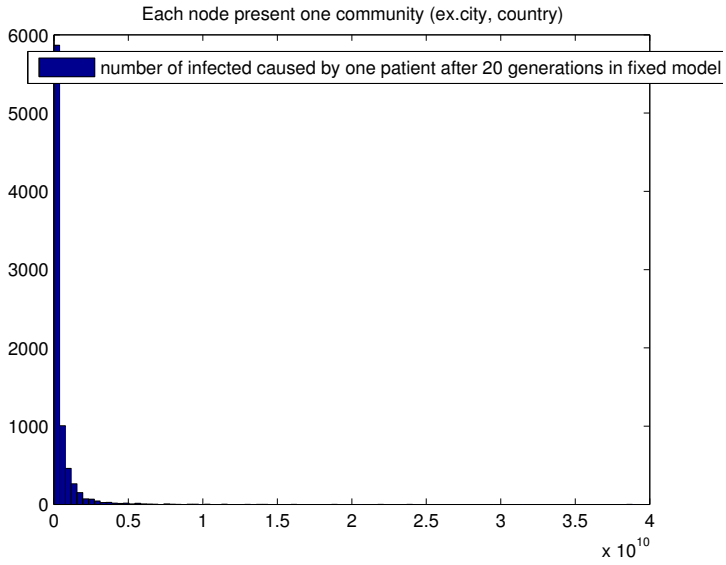


Figure 60: distribution of nonzero infected number after 20 generations in fixed contact model when each node present one community

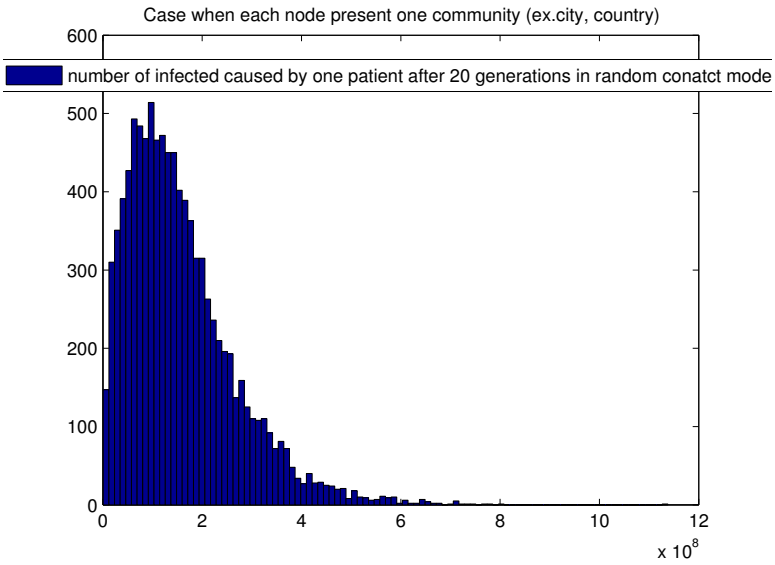


Figure 61: distribution of nonzero infected number after 20 generations in random contact model when each node present one community

From the four figures above, one can easily see that in each simulation model, the distribution obtained by random contact model is much centralized (alomst like a gaussien distribution) than the one of fixed contact model. The fact that the curve of fixed contact model going above the curve of random model in figure 55 is caused some extremely high value among 10000 simulations.
This fact naturally makes us think about the Galton-Watson process presented in this chapter. By comparing figure 35 and 59, we find that the simulation made by Galton-Watson process shares a very similar distribution form as the fixed contact model.

Because $R_0 = 2$ in the simulation of contact model, in order to compare the two processes more precisely we re-define a Galton Watson process for each node to have two roots in average. Which means

$$X_0 = 1$$
$$X_{n+1} = \sum_{j=1}^{X_n} \zeta_j \quad \text{for} \quad E(\zeta_j) = 2$$

The distribution of final value of X_{20} for 1000 simulations is presented in figure 63.

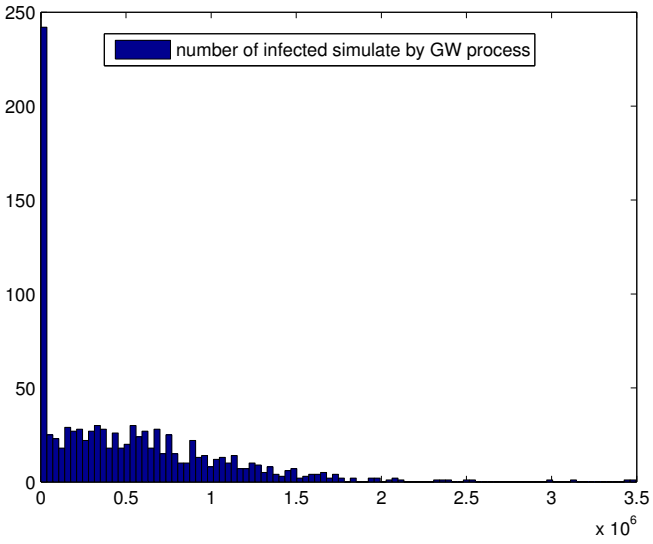


Figure 62: distribution of nonzero infected number after 20 generations simulated by Gaton Waltson process of reproductive ration 2

The statistic values such as standard deviation, percentage of disease survive etc of both two models and the Galton Watson process has also been collected and presented in table 6.

Each node present	contact model	zeros(%)	average (of non-zero)	median (of non-zero)	standard-deviation (of non-zero)
individual	fixed	98.26%	3.5805	2	3.4594
	random	88.88%	3.9910	3	2.9563
community	fixed	19.16%	4.4807e+08	1.4869e+08	1.0629e+09
	random	2.05%	1.5738e+08	135014750	1.1003e+08
Galton Watson		23.3%	6.5576e+05	559814	5.0002e+05

Table 6: stastic of 10000 simulations

In fact, the disease spread processing could be seen as a graph-based Galton Watson process. As the connection of network has been taken into account, it is in fact a 2 dimensional diffusion problem while the classical Galton Watson is only 1 dimensional. From table 6, it seems that the effect of large variance is even more obvious in spread simulation. However we remind that in our simulation model nodes are no longer independent to each other because of the existence of adjacency matrix. Exceptionally, in the random contact model two nodes belong to one community could be considered equivalent as we re-simulate the adjacency matrix for each generation. However in the fixed contact model once the adjacency matrix is constructed, we actually fix the "diffusion characteristic" as well. Some nodes therefore become more important than others for spreading, it explains the variance in fixed contact model is bigger than the one in random contact model.
In one word, the experiments above show that most part of the infected are caused by a few initial or inter-media patients in the network even in the case we observe an explosion of disease after a long period. It is more than important to control and protect these key individuals in a social network to slow down disease spreading.**How to find these key individuals when the connection matrix is given?** we will argue about this problematic in the next chapter.

7 Study on real cases

7.1 Office contact data

In order to validate our methods in a real data base of human contact. We have employed the contact data in a workplace available at **Social Patterns**: `http://www.sociopatterns.org/datasets/contacts-in-a-workplace/`. This data set contains the temporal network of contacts between individuals measured in an office building in France, from June 24 to July 3, 2013 (two weeks). It is very useful for modelling the infection of disease diffuse by weak contact such as influenza. This network was described and analyzed in the publication [Mathieu Génois et al. 2015].

Concretely a symmetric weighted (not binary) adjacency matrix of 92×92 with 1510 edges with total edge weight 19654 is provided, it is easy to calculate the density of graph:

$$P = \frac{1510}{92 \times 92} = 17.84\%$$

Which is much higher than what we suppose in our examples.

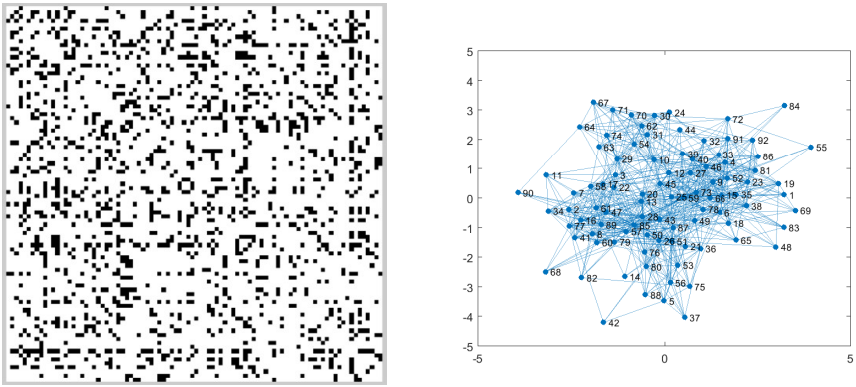


Figure 63: data set of office contact

From figure 63 one can see it is connected graph with no isolated parts. This means one patient in the company could possibly infect all the co-workers. The edge weight in the matrix present number of face-to-face contacts between two individuals in these two weeks.To well present the data set, we present in figure 64 the degree of weighted(left) and unweighted (right) degrees of nodes. We see that despite the unweighted degrees are quite homogeneous, some individuals in the network own a much higher weighted degree than others. It is caused by the frequency of contact between each pairs of individuals.

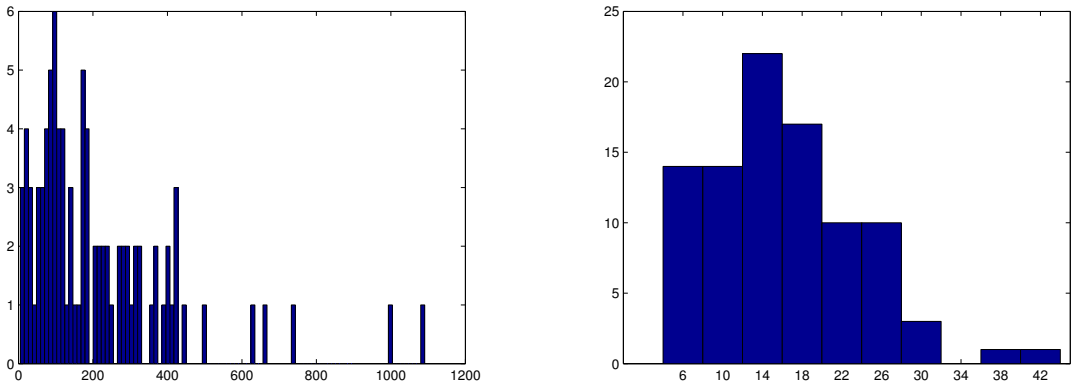


Figure 64: distribution of degrees in network

We use a binomial distribution to simulate the infection when one of the two nodes in a edge has been infected. Apparently, the more contacts they make, the more likely they could infect the disease.This assumption is specially suitable for diseases spread by relatively weak possibilities as the chance of infection is almost linear to the number of contacts during a small period.
For example for two co-workers nodes i and j where i is infected and j is not at the moment, if $A_{i,j}$=5 we will then simulate a binomial variable $x \sim binom(5,\lambda)$.By doing this we are actually mixing the fixed contact model and the random one. In fact, the contact is supposed to be fixed for every period (two weeks in our case) but each contact produce an infection by certain probability.

For our first experiment, we fix $\lambda = 1\%$ that makes the expected number of edges in the adjacency matrix become 138.9630 (simulated by Monte Carlo method of 1000 experiments). This leads to an average degree of nodes $d_{average} = \frac{138.9630}{92} = 1.5105$ in the network.

As having explained in **5.3**, the average degree of nodes could be seen as a potential reproductive ratio R_0. We would like to compare our simulation with the dynamical system for $R_0 = 1.5105$.
For this simulation we have employed the **SIS** model with infectious period $\frac{1}{\gamma} = 2$. The results are presented in figure 65 and the Matlab code of this simulation can be found in Appendix.

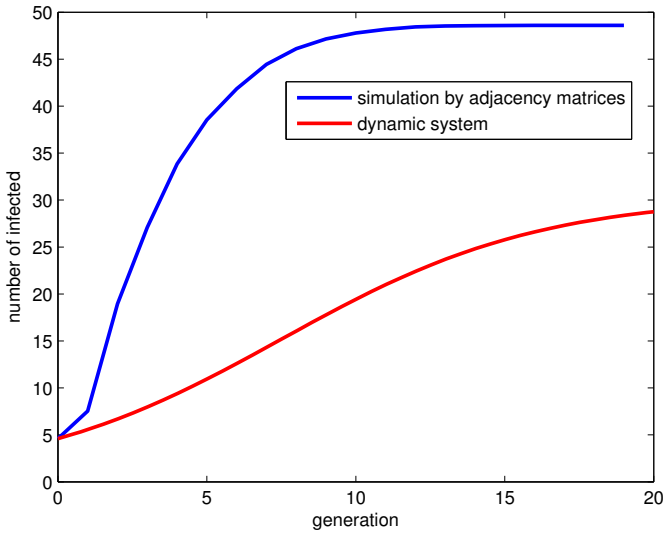


Figure 65: **SIS** spread simulation using office contact data

One observes that the diffusion curve simulated using the adjacency matrices stays always above the one made by dynamical system despite they share the same "average degree". This fact confirm the fact that in a real case the non-homogeneity of network plays a role in terms of accelerating the disease spread especially in our case when the non-homogeneity is significant (see figure 64). In fact, the nodes with highest degree in the network is more probably to be infected and once they are, it creates a huge risk for the spread to continue.
To have a more clear vision of this phenomenon, we have presented the distribution of average time for individuals stay infected (attention: it is different from the infected times as an individual stay infected during the infectious time $\gamma = 2$).

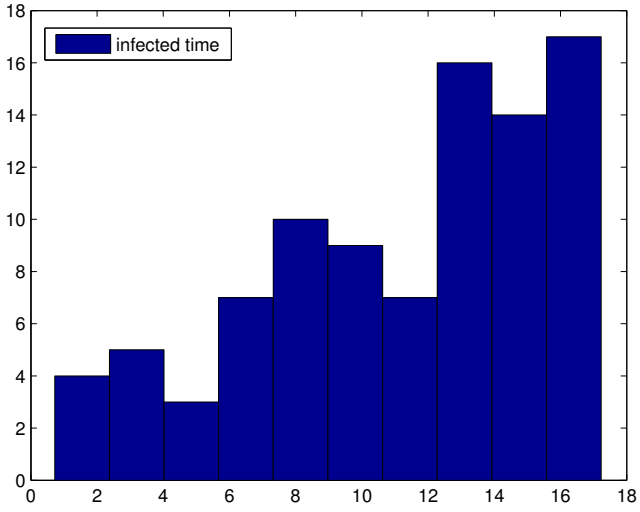


Figure 66: distribution of time for individuals stay infected

We can see from figure 66 that some nodes stays almost infected (17/20) during the period. These nodes cor-

respond to the individuals who are most active in the network.

What happens if we work with a SIRS model? If our previous explanation about the disease spread relying on the active nodes is valid then the fact that these nodes could be recovered after infectious time should slow down the diffusion suddenly.
Figure 67 shows a numerical experiment for **SIRS** system with recovered period equals to infectious period $\frac{1}{\gamma} = r = 2$ for 50 generations.

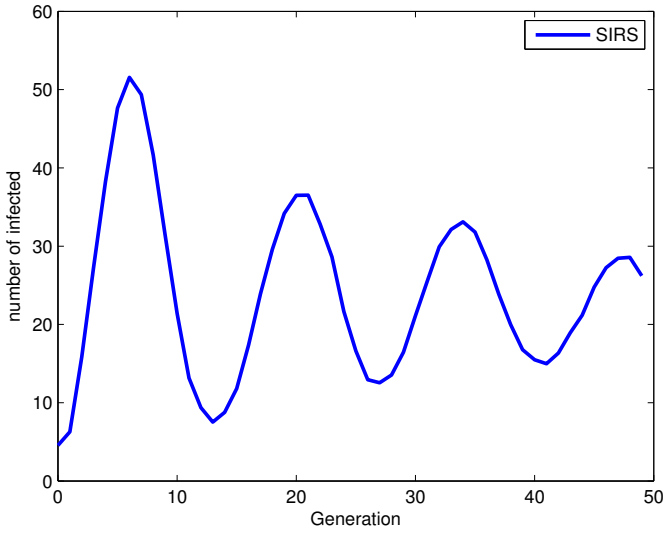


Figure 67: Different methods of caculating R_0

Under this conditions, the evolution of infected curve is almost a compelled vibration which is very interesting. In fact, at the beginning of diffusion, most part of active individuals are infected within a few generations. These active individuals are then recovered which cause a sudden drop of infected number. After certain moment when most part of population become susceptible again, the number of infected grows once more. As time goes, we get closer and closer to the convergence level. Actually, the damped oscillation curve is a nature characteristic for **SIRS** model which has been described in [O.E.Aièlo et al,2000] vy using a dynamical system. In our simulation, the oscillation stay much longer before convergence. The main reason of that could be the non-homogeneity brought by the weighted contact matrix and most importantly when the most active nodes in the network are recovered the paths of infection is cut off.

7.2 Airport transport data

Now we are interested in the case when each node present a community. For the following simulation, we employed the data set of global daily air flights of 227 airports with 4597 international airlines as presented in section **2**, figure 31 and in figure 77 in Appendix.One can see like the office contact data, it is a totally connected graph.
The average degree of the directed and weighted adjacency matrix is 128 but it is very nonhomogeneously distributed.
The busiest airlines are presented in figure 70, most of them are among the countries in Europe and north America.

Source_airport	Destination_airport	Routes
Canada	United States	399
United States	Canada	399
United States	Mexico	341
Mexico	United States	339
Spain	United Kingdom	290
United Kingdom	Spain	289
Spain	Germany	236
Germany	Spain	232
Taiwan	China	165
China	Taiwan	165

Figure 68: busiest airlines between countries, data from `https://github.com/gsmanu007/Complex-network-analysis-of-Airport-network-data/blob/master/airport_CnToCn.csv`

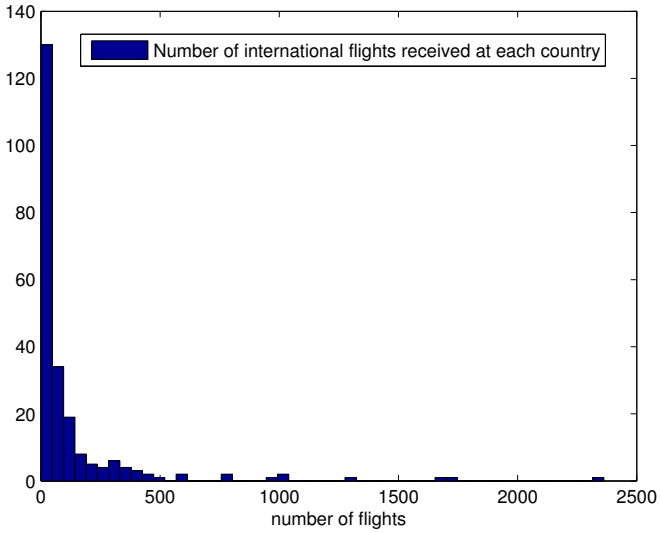


Figure 69: distribution of received international flights for each country in one day

From the histogram in figure 69, one could easily find that the distribution of international flights in each country is extremely non-homogeneous. The busiest country (the USA) receives 2362 flights from other countries every day while the median number is 38 for all countries.
As for the office contact data, we use a mixed model of fixed and random contact to simulate the evolution of infected number for $\lambda = 0.001$. The curve below is made by a Monte Carlo method for 1000 simulations for 10 initial patients.

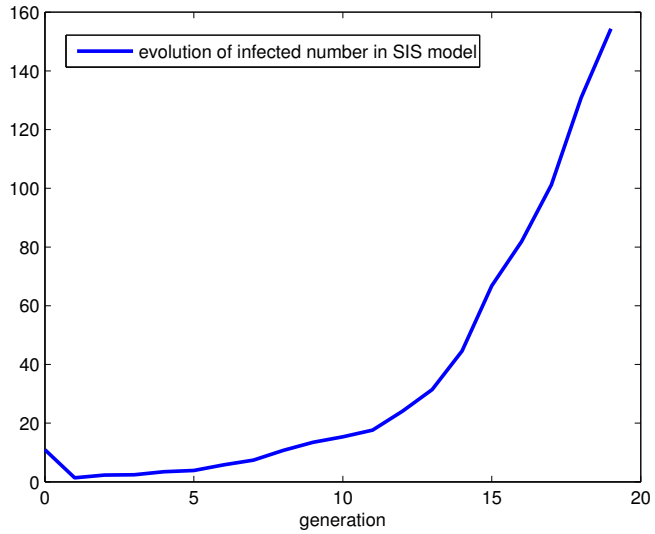


Figure 70: **SIR** model using airport transport data for $\lambda = 0.001$

As there is no limit for the infected number, we observe a explosion of disease under these conditions. An interesting point can be found for the decrease between the first and second generation. In fact as the 10 initial are randomly distributed they may not be found in the countries with most international airlines (like U.S.A, Canada, European countries China etc). Thus few of them could have a secondary infection. After several generations the most active countries mentioned above will be probably be engaged and once the infected appears it will be hard to wipe out as there exist a very frequent contact among people from these counties.
We have also made a simulation of **SIRS** model in Figure 71 for a recovered period $r = 2$.

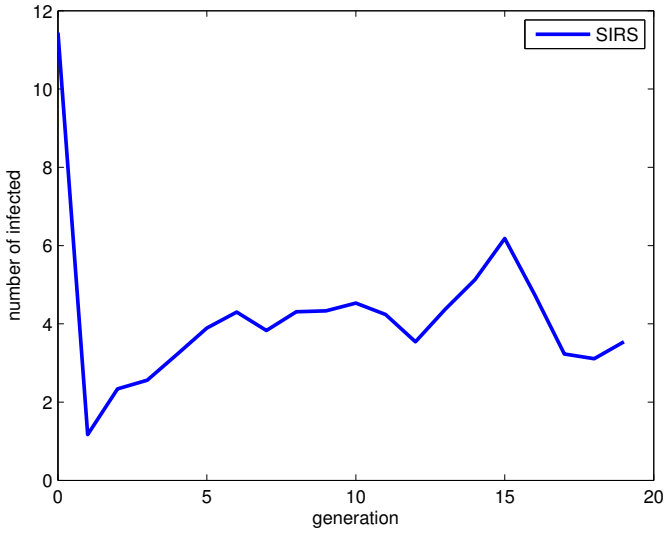


Figure 71: **SIRS** model using airport transport data for $\lambda = 0.001$

Compare with the simulation data of office contact data, it is easy to see that the oscillation phenomenon is less obvious in the case when each node presents a country. In fact, only the patients will be recovered after

infectious time the rest of the population in this country will stay susceptible which means the disease could still transfer by this country (node). Unlike the case when each node presents one individual, the paths of infection will not be cut off because of the recovery. On the other hand, the evolution is also very sensible on the value of λ. Several other experiments (obtained by 100 simulations) have been shown in the Appendix from figure **??** to **??**. The critical value for the **SIS** model is around 0.0007. For λ under than this value we are almost sure to have an extinction of disease, otherwise the explosion of infected seems inevitable. The critical value of **SIRS** model obtained by numerical experiments is around

8 Conclusion

We have analyzed different modelling of epidemic diffusion in this chapter especially the fixed contact model and random contact model. To our knowledge, these two methods which based on the Erdos-Renyi graph haven't been created before this paper. However, some similar or even more advanced (but harder to analysis) methods have been introduced for example in [Donald S. Burke, M.D, 2003]. The two methods created in this paper are easy to compute and most importantly once the initial parameters and type of graph (ie. if one node present an individual or a community) being set the result is only depend on the structure of adjacency matrix.
After all simulations and analysis in this chapter, we are convinced that the diffusion depend not only on the density (proportional to the average degree of nodes) of graph but actually heavily on the structure of connection.

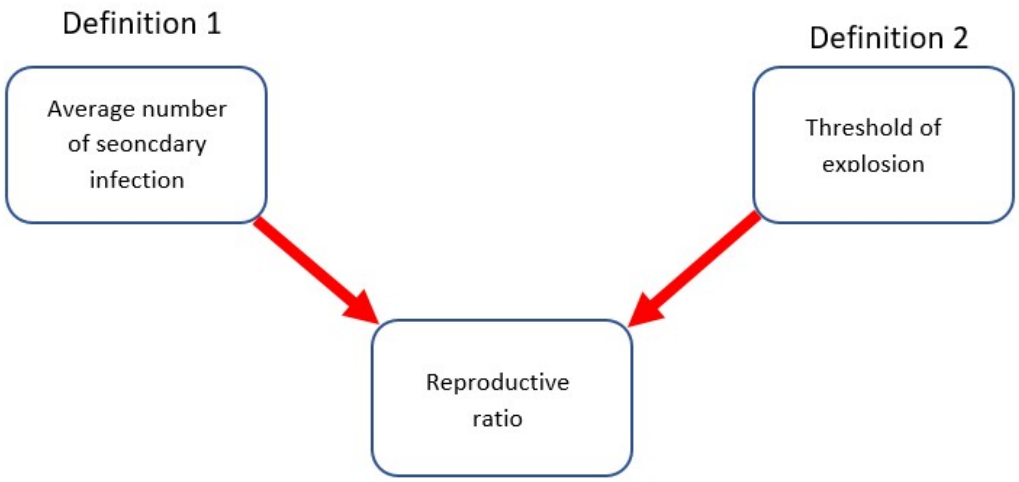


Figure 72: Different methods of caculating R_0

As have mentioned at the beginning of this chapter, several methods has been developed to define the reproductive ratio R_0. Most part of them is based on two concepts: the average number of secondary infected and the threshold of explosion. We have proved that in the case where each node are equivalent in terms of infection probability (for example, in dynamical systems or random model using one block structure matrix as presented in figure 48) these two quantities are equal. However once the connected graph brought some non-homogeneity among the nodes, the situation become more complicated. In fact as proved in examples the fact that the number of secondary infection of one patient is smaller than one is not enough any more to guarantee the no explosion of disease. When the connected graph is available, it is commonly adapted to use the dominant eigenvalue of adjacency matrix to define this threshold. Nevertheless, it doesn't stand for a perfect solution as well. For example, if

this dominant eigenvalue comes from a highly condensed sub-matrix with small size it will not decide if an disease explosion will take place as the community is isolated from outside.

As shown in figure 74, using the numerical experiments and theoretical analysis described in this chapter, we make some suggestions to define the value of R_0 based on the concepts of average infection number and threshold of explosion.
furthermore, by transferring the idea of Galton-Watson process into a graph based diffusion problem it seems we could explain why most part of infected population is caused by a few primary patients. Unlike the classical Galton-Watson process, because of the non-homogeneity in contact matrix it seems we may find the nodes who play a key role in disease spreading. It may provide a huge advantage and convenience in disease preventing. Our study in the next chapter is mainly inspired by this fact.
We remind that all the results obtained in this chapter are based on the assumption that the adjacency matrix of infection are always sparse(attention: the adjacency matrix of social contact is not necessarily sparse, it can be a dense contact matrix with low probability of infection).

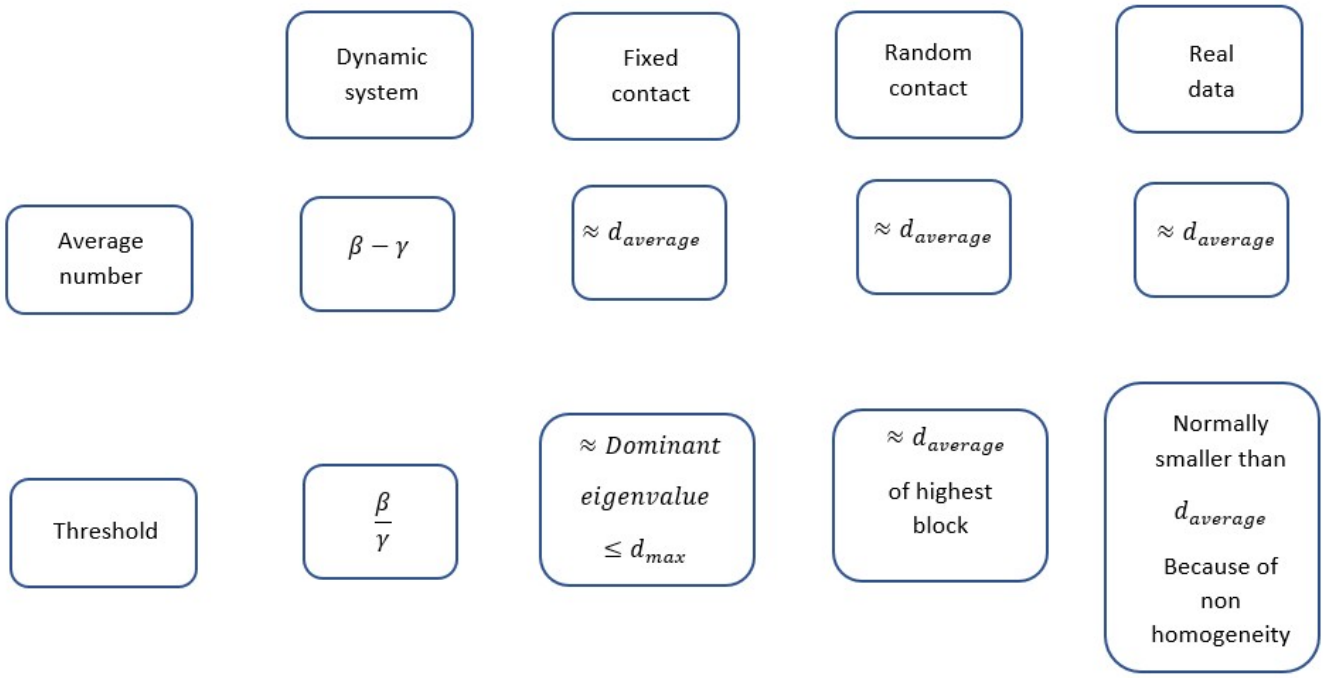


Figure 73: definitions of R_0

We recall that we have proved by simulation in this chapter that the R_0 obtained by the next-generation matrix is almost equivalent to the average degree of the most condensate community in the network.

Appendix of chapter 2

Complementary figures

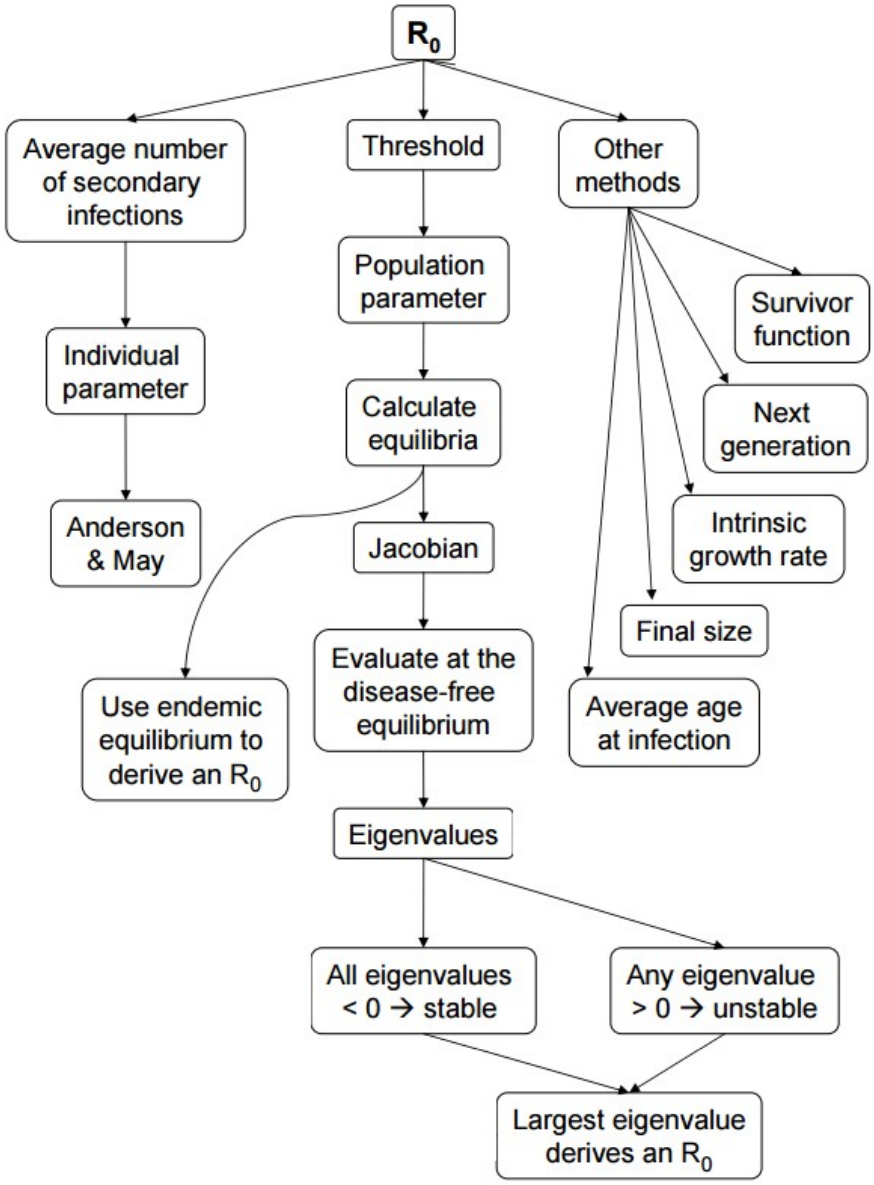


Figure 74: Different methods of calculating R_0 taken from `https://web.stanford.edu/~jhj1/teachingdocs/Jones-on-R0.pdf`

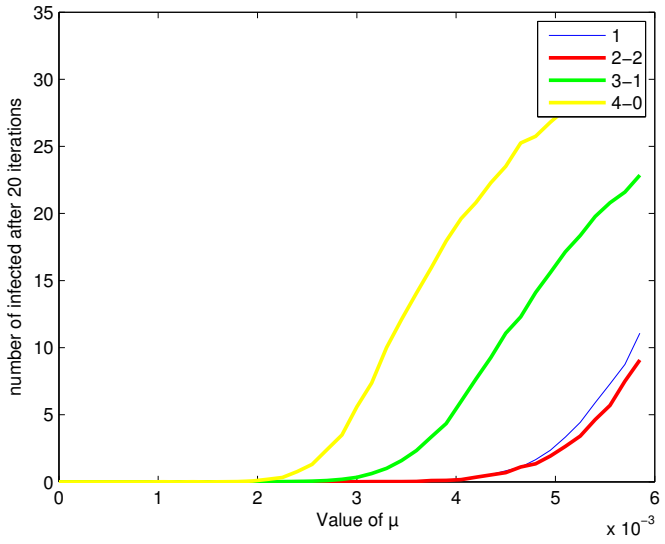


Figure 75: figure
random matrices at each generation

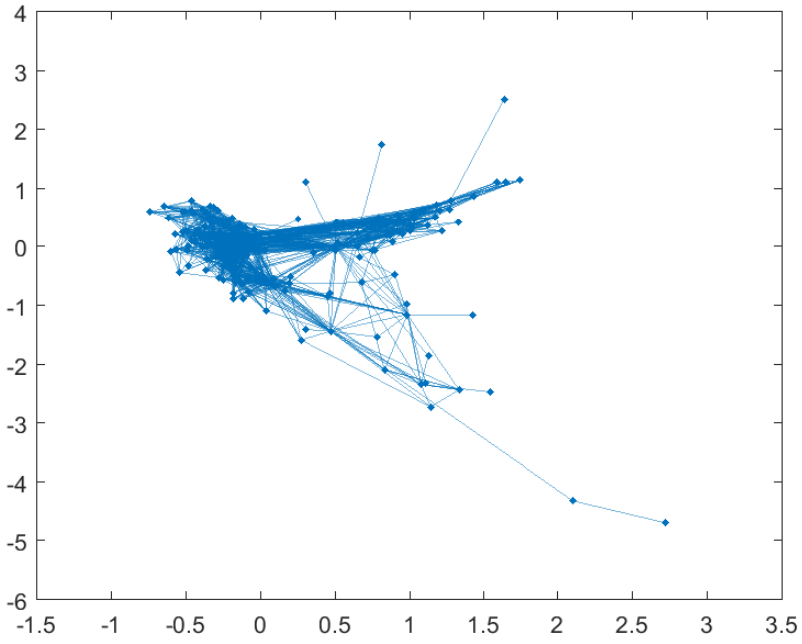


Figure 76: figure
connection graph formed by international airlines

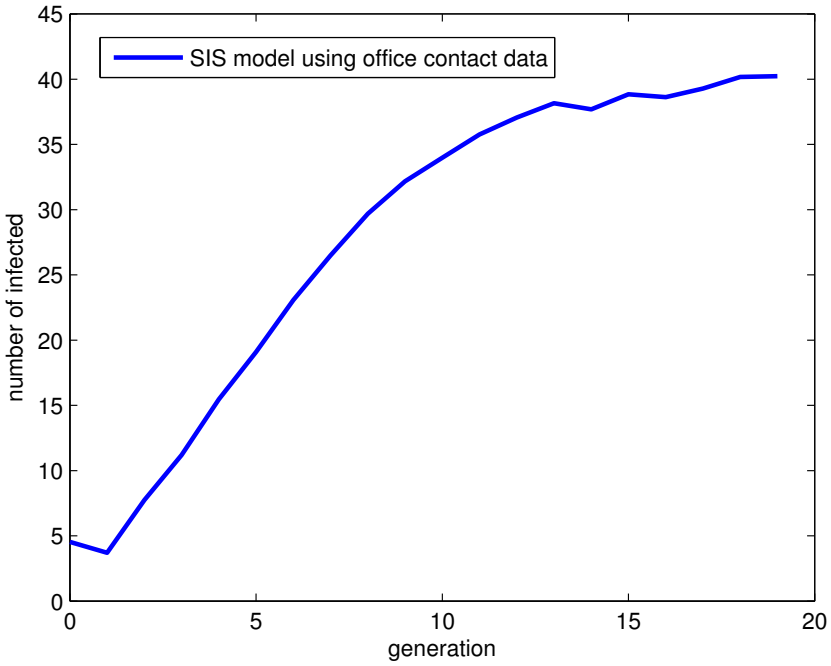


Figure 77: figure
SIS model using office data for $\lambda = 0.005$

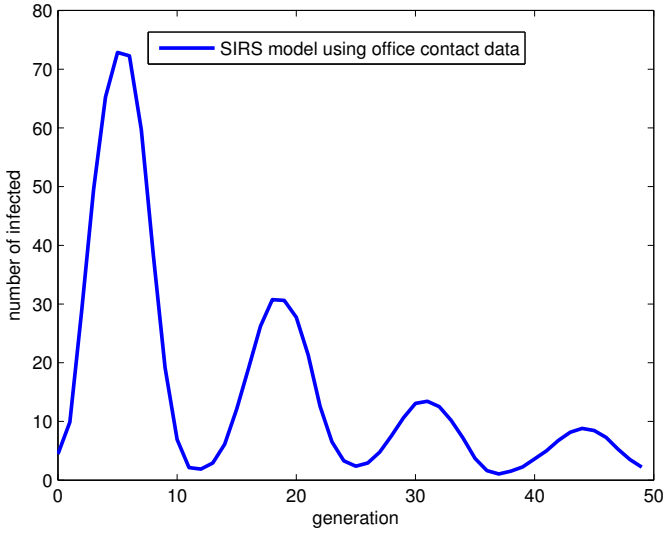


Figure 78: **SIRS** model using office data for $\lambda = 0.002$

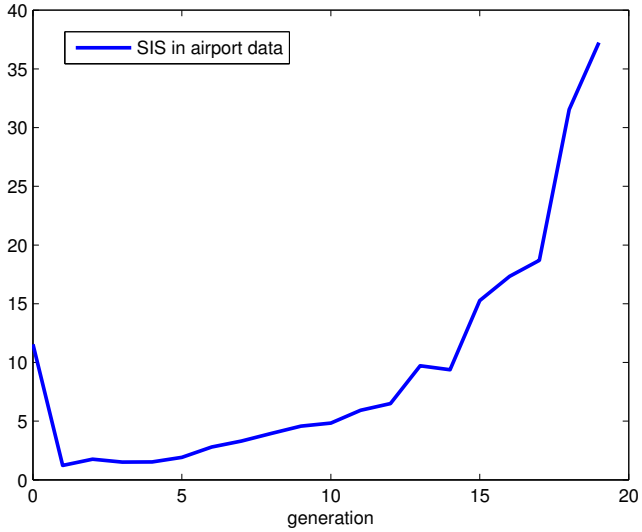


Figure 79: **SIS** model using airport transport data for $\lambda = 0.0008$

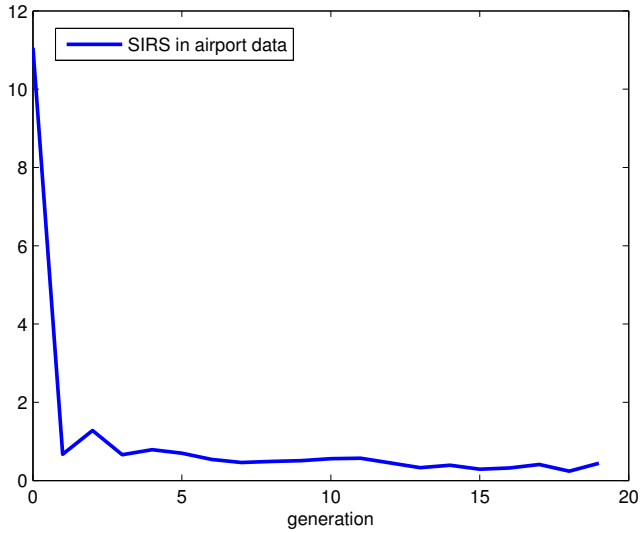


Figure 80: **SIS** model using airport transport data for $\lambda = 0.0006$

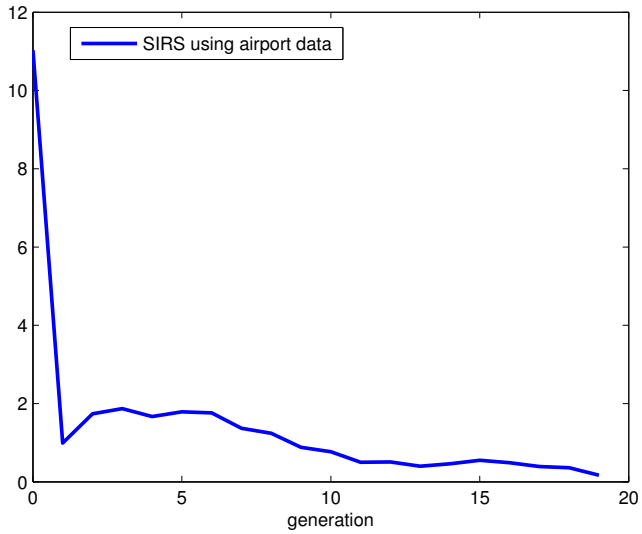


Figure 81: **SIRS** model using airport transport data for $\lambda = 0.0008$

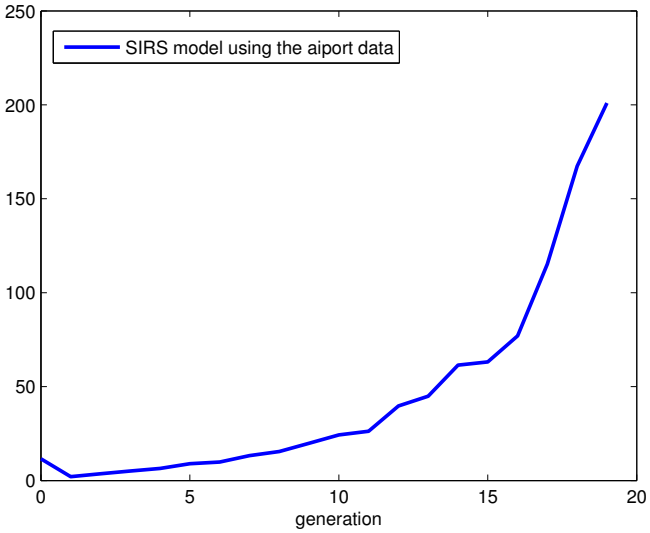


Figure 82: **SIRS** model using airport transport data for $\lambda = 0.0012$

Matlab code

We have chosen to show the most preventative code in Appendix, other simulation were made using the same idea.

Matlab code for generating the next-generation matrix

```
B = [1 0 0 0; 0 1 0 0; 0 0 1 0 ; 0 0 0 1 ];
N = [25,25,25,25];
```

```
[A Sim] = graphUD (B,N,p_in,p_out);
A=full(A);

V=sum(A+A');%degree of nodes
[C,ia,ic] = unique(V);
l=length(V);
 size=max(V);
NG=zeros(size);%next-generation matrix

 for k=1:size
     for m=1:size
         if sum(V==k)*sum(V==m)==0% if there exist nodes of degree m and k
             NG(k,m)=0;

         else
             n_m=sum(V==m);% number of nodes with degree m

             S=0;
             position_m=find(V==m);% index of nodes of degree m
             position_k=find(V==k);
             for i=1:length(position_m)
                 for j=1:length(position_k)

                     S=S+A(position_m(i),position_k(j))
                 end
             end
              NG(k,m)=S*(m-1)/(m*n_m);

         end
     end
 end
```

Matlab code for disease spread simulation using fixed contact model and random contact model

```
n=10000;%number of simulations

final_fix=zeros(1,n);
final_ran=zeros(1,n);% record the number of infected after 20 generations


evolution_fix=zeros(1,20);
evolution_ran=zeros(1,20);% average evolution curve

   B1 = [1,0,0,0;0,1,0,0;0,0,1,0;0,0,0,1];%reduced graph
  N = [50,50,50,50];%number of nodes in each block

 p_in=0.02;%density in blocks (could be several)
for j=1:n
        [A Sim] = graphUD (B1,N,p_in,p_out);%simulate erdos-renyi graph
        A=full(A);
```

```
        V0=binornd(1,0.05,1,200);% simulate initial patients
        V=zeros(200,20);
        V(:,1)=V0;
        evol=zeros(1,20);
        for u=1:19
            start=max(u-p,0)+1;
            vv=zeros(200,1);
            for i=start:u
                vv=vv+V(:,i);

            end

         V(:,u+1)=A*vv;
         V(:,u+1)=min(V(:,u+1),1); % when each node present one individual
        end
       S=sum(V);
         evolution_fix= evolution_fix+S;
       final_fix(j)=sum(V(:,20));
end

for j=1:n
        [A Sim] = graphUD (B1,N,p_in,p_out);
        A=full(A);
        V0=binornd(1,0.05,1,200);
        V=zeros(200,20);
        V(:,1)=V0;
        evol=zeros(1,20);
        for u=1:19
            [A Sim] = graphUD (B1,N,p_in,p_out);
             A=full(A);
            start=max(u-p,0)+1;
            vv=zeros(200,1);
            for i=start:u
                vv=vv+V(:,i);
            end

         V(:,u+1)=A*vv;
         V(:,u+1)=min(V(:,u+1),1);
        end
       S=sum(V);
         evolution_ran= evolution_ran+S;
       final_ran(j)=sum(V(:,20));
end
 plot(1:20,evolution_fix/n,'LineWidth',2); hold on;
 plot(1:20,evolution_ran/n,'r','LineWidth',2);
```

Matlab code for disease spread simulation using office contact data

```
B=tij_InVS(:,[2,3]);% reading the original data of edges in the graph

p=2;%infectious period
r=2;%revovered period

n=max(max(B));
A=zeros(n);
[q,~]=size(B);
% construct the contact matrix
for i=1:q

    t1=B(i,1);
    t2=B(i,2);
    A(t1,t2)=A(t1,t2)+1;
    A(t2,t1)=A(t2,t1)+1;

end


T=sum(A);
 del=find(T==0);
 A(del,:)=[];
 A(:,del)=[];
 [n,~]=size(A);%delete empty nodes in the graph

  lambda=0.01;

 Adj=A;
 % construct the Adjacency matrix with binomial variables
 for i=1:n
     for j=1:n
         Adj(i,j)=min(binornd(A(i,j),lambda),1);

     end
 end

 %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
 account=zeros(1,92);
 evolution1=zeros(1,20);
 n=100;
 m=92;

for j=1:n

        V0=binornd(1,0.05,1,92);% initial patients in the network presented by a binary
            vector V_0
        V=zeros(92,20);
        V(:,1)=V0;
        evol=zeros(1,20);
        for u=1:19

```

```
51
52                  for i =1:m
53                      for j=1:m
54                          Adj(i,j)=min(binornd(A(i,j),lambda),1);% resimulate the adjacency
                                matrix based on the contact matrix for each generation
55
56                      end
57                  end
58
59                 start=max(u-p,0)+1;
60                 vv=zeros(92,1);
61
62                 for i=start:u
63                     vv=vv+V(:,i);
64
65                 end
66 %
67                  start2=max(u-p-r,1);%*
68                  end2=max(u-p-1,0);
69                  R=zeros(92,1);
70                  for i=start2:end2
71                     R=R+V(:,i);
72
73                  end
74                 vv=max(vv-R,0);%  form *: this part only useful for SIRS model
75
76
77              V(:,u+1)=Adj*vv;
78                V(:,u+1)=min(V(:,u+1),1);% each node present on individual
79             end
80
81             account=account+sum(V');
82            S=sum(V);
83               evolution1= evolution1+S;
84 end
85 %account=account/n;
86
87 beta=1.5105;
88
89 N=100;
90 gamma=1;
91
92 f = @(t,x) [-beta*x(1)*x(2)/N+gamma*x(2); beta*x(1)*x(2)/N-gamma*x(2)];
93     [t,xa] = ode45(f,[0 20],[92 92*0.05 ]);% dynamic syst for SIS
94
95     f2 = @(t,x) [-beta*x(1)*x(2)/N; beta*x(1)*x(2)/N-gamma*x(2); gamma*x(2)];
96     [t2,xa2] = ode45(f2,[0 20],[92 92*0.05 0]);% dynamic syst for SIRS
97
98  plot(t,xa(:,2),'b','LineWidth',2); hold on
99  plot(t2,xa2(:,2),'r','LineWidth',2)
100 plot(0:19,evolution1/n,'LineWidth',2); hold on;
```

```
101 %plot(t,xa(:,2),'b','LineWidth',2); hold on
102 plot(0:19,evolution2/100,'r','LineWidth',2);
```

Chapter 3 Transition Capability Measure

1 Summary

In this chapter we are interested in the optimal strategy of vaccination in order to slow down the disease spreading. Suppose that we could only vaccinate a part of the whole population in time, therefore it is important to choose wisely. The following assumptions will be respected in the whole chapter.

- The contact network is totally known ;
- We limte ourselves in the fixed contact model defined in chapter 2;
- For networks shown in this chapter, each node presents one individual;

2 Introduction and description

Following the idea of "graph based 2-D Galton Watson process" presented in the chapter2, most part of infected after a long period of time could probably caused by very few primary patients. Therefore, we would like to identify the nodes which may probably generate more roots in the infection process based on the contact graph structure . We remind that in a classical Galton-Watson process, all individuals are considered equivalent in probability distribution so there are no nodes could be more "important" than others. However, like we have discussed in chapter 2, the introduction of adjacency matrix breaks the homogeneity of all nodes.
Imagine we put ourselves in the situation when a city faces the invade of some deadly influenza virus with high rate

of infection. Unfortunately, we could only vaccinate 10% of population to prevent the disease explosion. Suppose the social behavior network is perfectly known,obviously we would like to use our limited vaccinate on people who own a higher probability of being infected and more importantly higher probability to infect others.

The main problematic in this chapter stands: **How can we predict the nodes who will play an "important" role in disease diffusion?**
The first nature idea that comes to our mind is to take the nodes with highest degree. The reason is very simple,

actually if the probability of all edges appear in the temporary connection graph stays the same, higher degree surely leads to higher number of secondary infected. In fact, there exists a well-known strategy of vaccination based on the idea of choosing people with highest degree. In this method, we will randomly interview n person and ask each of them to name of his or her friends. We will than vaccinate all the friends being mentioned. By doing this we make a random selection where the probability of each individual to be chosen is proportional to its degree in the social network. It is also much easier to carry out because in reality it is almost impossible to get complete knowledge of social network. This method is well described in the book [the economist of epidemiology, Tassier, Troy](section **6.5**)
These nodes with high degree certainly play an important role in terms of virus spreading. But does there exist a more efficient measure? **Yes! We believe so!**

To give a clear vision for readers, we give an example of connection graph in figure 83. It is easy to find that the nodes can be roughly separated into two communities which locate on both sides of the red node. Therefore the red node is the only link between these two communities. The out-degree of the red node is two which is not the highest in the network. (for example, the green one has three as out-degree). However one can not deny that when disease appears in one of two communities, the red node becomes extremely important as it is the only way for virus to travel across communities. Which means the red node will be the front line to prevent half of the

population. Its importance is needless to say. Using this example, we would like to show that the importance of node should not only be decided on its degree but also the exact position it belongs in the network.

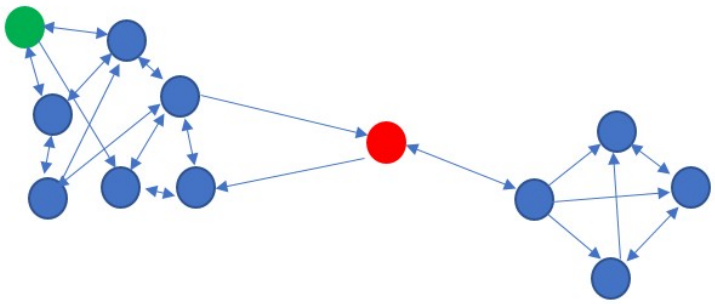


Figure 83: example of connection graph

In this chapter we define a quantity named transition capability of network to score the capability of a node for connecting one community to another. The first task we face is to identify the different communities exist in a connection graph. Many methods have been developed for this objective, a lot of them are based on the similarity measure.

3 Similarity measure

Given a weighted or binary adjacency matrix A, the goal is to obtain a similarity matrix $S_{i,j}$ presenting the measure of similarity between node i and j, which is a value between 0 and 1.The higher the value is , the more similar the two nodes are. The method is to give a relatively high value of similarity for two nodes who are most likely to be in the same community.
For example, let's look at the network presented in figure 83. For two nodes from the same community should have a higher similarity as they own a lot of children and parents in common. On the other hand, for two nodes from two different communities their similarity should be very close to zero. We would like to establish a mathematical expression to quantify this measure. In this paper, we based ourselves on the methods described in [S.Cheng, A.Laurent,P.Van Dooren, 2017]

The main idea to count the common "parents" or "children" of each pair of nodes and then divide it by the c of these two nodes. Concretely suppose A is binary, for nodes i and j, the number of common "children" is : $\sum_{k=1}^{n} A_{i,k}A_{j,k}(A\cdot A^T)_{i,j}$).We then divide this quantity by geometric average of out-degrees then we get:

$$\frac{(AA^T)_{i,j}}{\sqrt{\sum_{k=1}^{n} A_{i,k}}\sqrt{\sum_{k=1}^{n} A_{j,k}}}$$

As the same for common "parent", one only need to replace the out-degrees by in-degrees in the denominator:

$$\frac{(A^TA)_{i,j}}{\sqrt{\sum_{k=1}^{n} A_{k,i}}\sqrt{\sum_{k=1}^{n} A_{k,j}}}$$

We finally get the similarity measure for nodes i and j:

$$S_{(i,j)} = \frac{(AA^T)_{i,j}}{\sqrt{\sum_{k=1}^{n} A_{i,k}}\sqrt{\sum_{k=1}^{n} A_{j,k}}} + \frac{(A^T A)_{i,j}}{\sqrt{\sum_{k=1}^{n} A_{k,i}}\sqrt{\sum_{k=1}^{n} A_{k,j}}}$$

To compute the similarity matrix, one could simply use the row-normalized adjacency matrix C and the column-normalized adjacency matrix D:

$$C(i,:) = \frac{A(i,:)}{\sqrt{\sum_{k=1}^{n} A_{i,k}}} \qquad D(:,j) = \frac{A(:,j)}{\sqrt{\sum_{k=1}^{n} A_{k,j}}}$$

Then S can be written as

$$S = [C|D^T] \cdot [C|D^T]^T$$

By definition, S is surely symmetric.

Remark : There are quite a few other similarity measures as described in [A.Browet, P.Van Dooren, 2015], the one we have chosen owns two important advantages:

- Its complexity of computation is low: n^3 for full adjacency matrix and n^2 for sparse adjacency matrix ;
- It is very efficient in community detection;

In fact an approximation of low rank matrix could also be applied on S, by doing that we can get a linear computation complexity.
To make an example, we construct the similarity matrix for the network shown in figure 83. The adjacency matrix could be written as:

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Using the similarity measure introduced above, we obtain the similarity matrix:

$$S = \begin{bmatrix}
1.0000 & 0.4081 & 0.4268 & 0.6250 & 0.6422 & 0.1443 & 0.5303 & 0 & 0 & 0 & 0 & 0 \\
0.4081 & 1.0000 & 0.5854 & 0.2831 & 0.3679 & 0.4025 & 0.5122 & 0.1581 & 0 & 0 & 0 & 0 \\
0.4268 & 0.5854 & 1.0000 & 0.4268 & 0.5387 & 0.4330 & 0.4268 & 0.1768 & 0 & 0 & 0 & 0 \\
0.6250 & 0.2831 & 0.4268 & 1.0000 & 0.6098 & 0.1443 & 0.1768 & 0.4268 & 0 & 0 & 0 & 0 \\
0.6422 & 0.3679 & 0.5387 & 0.6098 & 1.0000 & 0.4553 & 0.1250 & 0.2041 & 0 & 0 & 0 & 0 \\
0.1443 & 0.4025 & 0.4330 & 0.1443 & 0.4553 & 1.0000 & 0.1443 & 0 & 0.1443 & 0 & 0 & 0 \\
0.5303 & 0.5122 & 0.4268 & 0.1768 & 0.1250 & 0.1443 & 1.0000 & 0 & 0.1443 & 0 & 0 & 0 \\
0 & 0.1581 & 0.1768 & 0.4268 & 0.2041 & 0 & 0 & 1.0000 & 0 & 0.2500 & 0.4082 & 0.5000 \\
0 & 0 & 0 & 0 & 0 & 0.1443 & 0.1443 & 0 & 1.0000 & 0.4541 & 0.4553 & 0.3809 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2500 & 0.4541 & 1.0000 & 0.2041 & 0.8536 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.4082 & 0.4553 & 0.2041 & 1.0000 & 0.4082 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5000 & 0.3809 & 0.8536 & 0.4082 & 1.0000
\end{bmatrix}$$

The row in red color presents the red node in figure 83. It is easy to see that for nodes in the same community the similarity among them is always positive. If two nodes belong to different communities the similarity between them is zero because in our case the two communities will be totally isolated when we remove the red node. We make a remark that the diagonal of similarity matrix is always one because all nodes are 100% similar to themselves.

How can the similarity help us to define the transition capability ?

It seems once the communities are known, it will be easy to find the nodes who are capable to transfer virus from one community to another. However, in this paper we will not identify those communities directly. There are several reasons. First, it is quite expensive in terms of complexity when we are given an adjacency matrix of large size. Most importantly, unlike the ideal example, in most real cases the way to split communities is not unique. Therefore, we would like to define the transition capability using the similarity measure but without pre-identifying the communities.

4 Transition Capability

The idea is quite straightforward: a node is capable to transfer virus across communities means its parents and children should not belong to one same community. **The less the parents and children are similar to each other the higher transition capability should this node own.**

Thus, we need an "anti-similarity " measure. In order to simplify the notation, we denote E as the $n \times n$ matrix of all 1 and $S' = E - S$ is also a matrix of elements of real numbers between 0 and 1. This matrix can be considered as an anti-similarity measure.

Therefore we browse all the pairs of "parent" and "children" and measures the anti-similarity of each pair.The sum of anti-similarity measure of all pairs will be considered as the transition capability measure of this node.

In order to accentuate the importance of anti-similarity value, we use the convexity of exponential function to define an enlarged anti-similarity matrix S_a.

$$\forall i, j \quad (S_a)_{i,j} = \frac{exp(S'_{i,j}) - 1}{exp(1)} \in [0,1]$$

The transition capability of the network associated to matrix A is denoted as a vector T where T_i presents the value of "transition capability" of node i in the network.

For node i this value can be expressed as:

$$T_i = \sum_{q=1}^{n} \sum_{j=1}^{n} (S_a)_{i,j} A_{q,i} A_{i,j}$$

From where one can simplify the expression

$$T_i = \sum_{j=1}^{n} A_{i,j} \sum_{q=1}^{n} ((S_a)_{i,j} A_{q,i})$$

$$T_i = \sum_{j=1}^{n} A_{i,j}(S_a^T A)_{j,i}$$

$$T = diag(AS_a^T A)$$

One can see that the degree of nodes (number of children and parents) play an important role in this measure but it is not the only factor.

We also define another quantity called "transition characteristic" of one node in this paper to present the percentage of links towards to and receives from other communities. This quantity is denoted as a vector C of length n.

$$C_i = \frac{\sum_{q=1}^{n}\sum_{j=1}^{n}(S_a)_{q,j}A_{q,i}A_{i,j}}{\sum_{q=1}^{n}\sum_{j=1}^{n}A_{q,i}A_{i,j}}$$

$$C_i = \frac{diag(AS_a^T A)_i}{\sum_{q=1}^{n}\sum_{j=1}^{n}A_{q,i}A_{i,j}}$$

The denominator $\sum_{j=1}^{n} A_{q,i}A_{i,j}$ actually present the total number of pairs "parent-children" to be compared. The transition characteristic of a node does not depend on its degree.
Since $(S_a)_{q,j} \in [0,1]$, we can easily deduce $C_i \in [0,1]$. To give an example we are going to construct three vectors

D, T, C which correspond to the out-degree, the transition capability and the transition characteristic of nodes in figure 83.

$$D = \begin{bmatrix} 4 \\ 4 \\ 4 \\ 4 \\ 3 \\ 3 \\ 2 \\ 2 \\ 4 \\ 1 \\ 3 \\ 2 \end{bmatrix} \quad T = \begin{bmatrix} 1.5315 \\ 4.2264 \\ 3.7246 \\ 1.8002 \\ 3.4123 \\ 2.8028 \\ 2.5922 \\ 2.2901 \\ 2.6892 \\ 0.2664 \\ 1.2810 \\ 0.5327 \end{bmatrix} \quad C = \begin{bmatrix} 0.1914 \\ 0.2113 \\ 0.2328 \\ 0.2250 \\ 0.2275 \\ 0.3114 \\ 0.3240 \\ 0.3817 \\ 0.2241 \\ 0.1332 \\ 0.1423 \\ 0.1332 \end{bmatrix}$$

The red node owns a relative high transition capability value with low degree. In fact,unsurprisingly, its value of transition characteristic is the highest among all 12 nodes.

Complexity In the case when the adjacency matrix is dense, as we have to build $diag(AS^T A)$, one has to compute the total production of two $n \times n$ size matrices and that takes n^3 operations. However when the matrix is sparse (defined in **chapter 0** number of edges $m = O(n)$) the complexity could be reduced to n^2 by simply doing a naive multiplication.

5 Betweenness Centrality

Betweenness centrality is a measure of centrality in a graph based on shortest paths. This method introduced by [LC.Freeman ,1977] is widely used in graph theory.The betweenness centrality for each vertex is defined as the number of these shortest paths of two nodes in the network that pass through the vertex.

We give the expression of betweenness centrality of a node v in a network.

$$g(v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where s, t are two other nodes in the network. σ_{st} present the total number of shortest paths from node s to node t and $\sigma_{st}(v)$is the number of those paths that pass through v.

Remark : For weighted graph, the shortest path is defined as the path with smallest sum of edge weight in it.

In disease spreading, the shortest path could be explained as the most efficient way for an infected s to contaminate a susceptible t.
The betweenness Centrality measure of figure 83 is obtained by Matlab:

$$B = \begin{bmatrix} 6.0000 \\ 11.5000 \\ 5.0000 \\ 4.3333 \\ 20.5000 \\ 34.8333 \\ 16.8333 \\ 57.0000 \\ 48.5000 \\ 0 \\ 10.5000 \\ 0 \end{bmatrix}$$

The red node in the network does own a much higher centrality than all the other nodes. In fact if two nodes belong to different communities the shortest path between them will surely pass by the red node. It seems this measure could also be a good solution to accentuate the importance of nodes who are capable to transfer virus from one community to others.

According to [Ulrik Brandes,2001], the complexity to build the betweenness centrality for all nodes in a sparse graph will be $O(n^2 log(n))$ which is higher than n^2 for the transition capability measure.

6 Comparison of three measures

For the three measures mentioned in this chapter (degree, transition capability and betweenness centrality), we would like to compare their performance in terms of overlapping and efficiency on identifying the most "dangerous" individual in disease spreading.

6.1 Overlap ratio

We have already seen that the importance order defined by these measures have a significant difference for the example in figure 83. However, we have made this example in purpose to emphasize the "transition effect". What happens if we apply them to random simulated graph?

We use Erdos-Renyi method(see chapter 2) to simulate block structured random matrices. We also change the inside and outside density P_{in}, P_{out} to see what influence they make on the overlapping rate.
We use two types of block structured matrices for this simulation as presented in figures 84 and 85. The first type of contact graph is formed by four self-connected communities of equal size which may represent for example population in four cities. In this case, the inside density P_{in} presents the contact frequency for people in one same city while the outside density P_{out} presents the frequency of people come from different cities. The second graph structure we have picked is formed by a cycle of four communities. This structure also present the behavior of some diseases. For example for virus spread by sexual contact such as HIV, the contact mainly happens between males and females. Then the inside density P_{in} presents the frequency contact across communities (heterosexuality) while P_{out} presents the contact inside same community (homosexuality).
For each graph, we pick 10% nodes of highest values in each of three measures. We then present the overlapping ratio of each two measures in tables below. All results are obtained by Monte Carlo simulation for 1000 networks simulated independently.The commented Matlab code of this simulation could be found in Appendix.

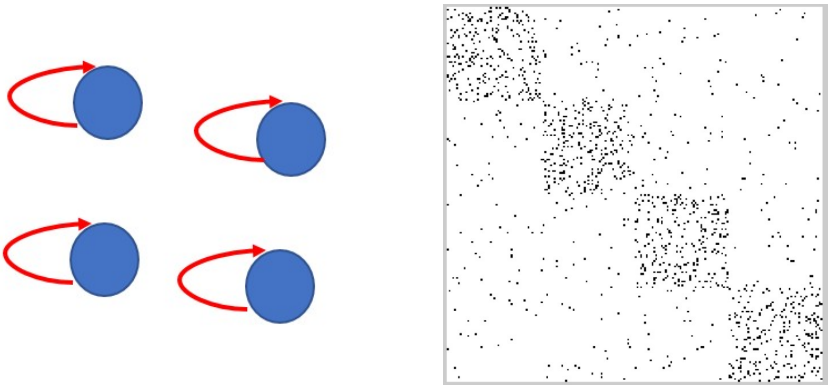


Figure 84: first structure: 4 self-connected communities with noise for 50 nodes each

Overlapping in highest 10%	transition capability	degree	betweenness centrality
transition capability	100 %	90.71%	79.21%
degree	90.71%	100 %	76.96%
betweenness centrality	79.21 %	76.96%	100%

Table 7: $P_{in} = 0.1, P_{out} = 0.05$

Overlapping in highest 10%	transition capability	degree	betweenness centrality
transition capability	100 %	88.69%	73.33%
degree	88.69%	100%	70.54%
betweenness centrality	73.33%	70.54%	100%

Table 8: $P_{in} = 0.1, P_{out} = 0.02$

Overlapping in highest 10%	transition capability	degree	betweenness centrality
transition capability	100%	87.10%	66.38%
degree	87.10%	100%	62.71%
betweenness centrality	66.38%	62.71%	100%

Table 9: $P_{in} = 0.1, P_{out} = 0.01$

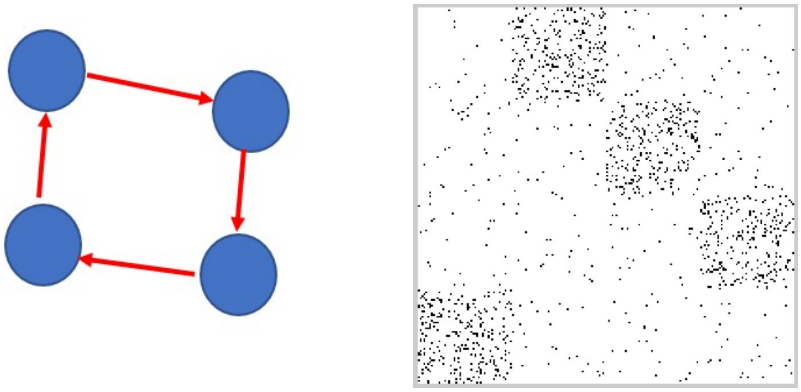


Figure 85: first structure: 4 cycled communities with noise for 50 nodes each

Overlapping in highest 10%	transition capability	degree	betweenness centrality
transition capability	100%	88.98%	75.86%
degree	88.98%	100%	73.67%
betweenness centrality	75.86%	73.67%	100%

Table 10: $P_{in} = 0.1, P_{out} = 0.05$

Overlapping in highest 10%	transition capability	degree	betweenness centrality
transition capability	100%	88.29%	72.45%
degree	88.29%	100%	70.58%
betweenness centrality	72.45%	70.58%	100%

Table 11: $P_{in} = 0.1, P_{out} = 0.02$

Overlapping in highest 10%	transition capability	degree	betweenness centrality
transition capability	100%	89.65%	77.90%
degree	89.65%	100%	75.73%
betweenness centrality	77.90%	75.73%	100%

Table 12: $P_{in} = 0.1, P_{out} = 0.01$

From examples above, one can easily see that these three measures are in fact highly correlated. The overlapping ratio stays always superior than 60%. The transition capability is actually more correlated to the degree than

to the betweenness centrality. Interestingly,in both structures the higher P_{out} is, the more correlated three measure's are. We thought there can be an explanation on this phenomenon. In fact the low value of P_{out} implies the block structure is more clear where we have few nodes link different communities. The transition capability and betweenness centrality measures will surely own a high value at those nodes. However, these nodes don't necessarily have a high degree. We may say the transition capability measure is more "useful" when the contact graph is well clustered.

6.2 Efficiency of vaccination

We now come back to the essential question launched at the beginning of this chapter: if we are only allowed to vaccinate 10% of the population, how can me make our choices wisely to avoid the disease explosion?
Using the simulation methods defined in chapter 2 of this paper, we want to verify the behavior of spreading when vaccinating 10% nodes of highest value in one of the three measures mentioned above.
We will now compare four different strategies of vaccination in the network.

- vaccinate 10% nodes randomly ;
- vaccinate 10% nodes of the highest degree ;
- vaccinate 10% nodes of the highest transition capability ;
- vaccinate 10% nodes of the highest betweenness centrality ;

SIS model

We are going to apply these four strategies on different structured graph randomly simulated by Erdos-Renyi method. The results are obtained by a Monte Carlo simulation for 1000 independent simulations following same community structure.
We start with a 200×200 block structured network including four self-connect communities as shown in figure 84 for $P_{in} = 0.1$ and $P_{out} = 0.01$. We use fixed contact model which means the adjacency matrix stays constant at each iteration and obviously each node presents one individual in this case.

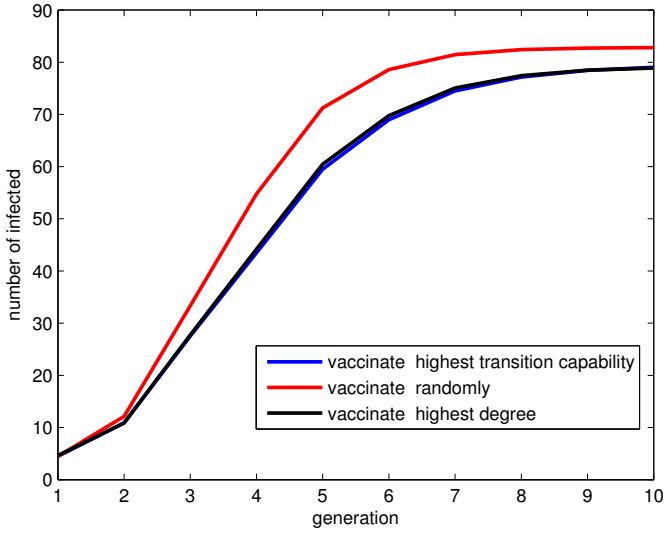


Figure 86: evolution of infected for $P_{in} = 0.1$ and $P_{out} = 0.01$

From figure 86, one observes that by following the strategies of vaccinating the people with the highest transition capability or highest contact degree, we have successfully slow down the spread of disease.It seems in this case, these two strategies have almost the same effect on the evolution of infected number. As have mentioned in **5.2** the difference between two measures can be enlarged when P_{out} is relatively low. Thus we make another simulation only using these two methods of vaccination for exactly the same condition except $P_{out} = 0.001$.

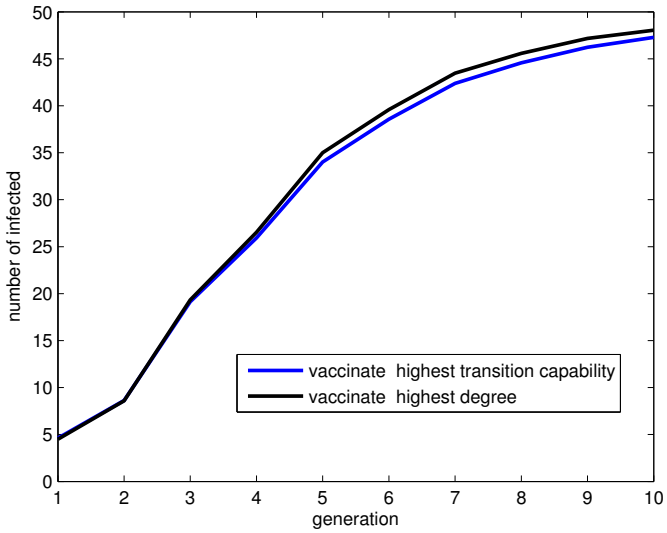


Figure 87: evolution of infected for $P_{in} = 0.1$ and $P_{out} = 0.01$

We can see that when the value P_{out} is small, the transition capability does own a slight advantage in terms of retarding the diffusion.
How about the measure of betweeness centrality?
We have taken the same examples in figure 86 and figure 89 and add the evolution curve (green) of infected number when vaccinating the population with the highest centrality.

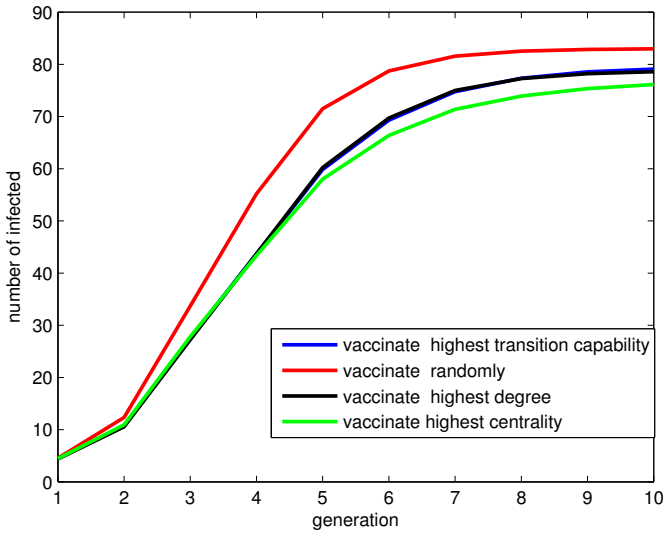


Figure 88: evolution of infected for $P_{in} = 0.1$ and $P_{out} = 0.01$

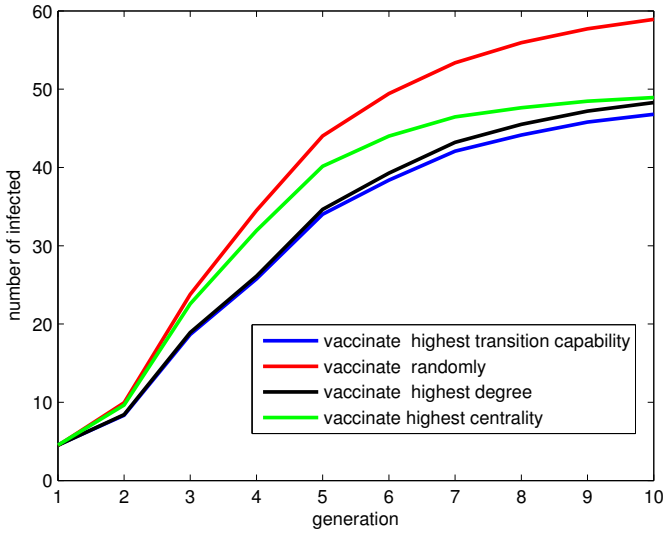


Figure 89: evolution of infected for $P_{in} = 0.1$ and $P_{out} = 0.001$

From the two figures above, we see that the last strategy owns a slight advantage when the density outside the communities is high. However, when we decease the value of P_{out} it seems this strategy is less efficient than the two others.

SIRS model

The examples have also been made for the **SIRS** models for infectious period $\frac{1}{\gamma} = 1$.

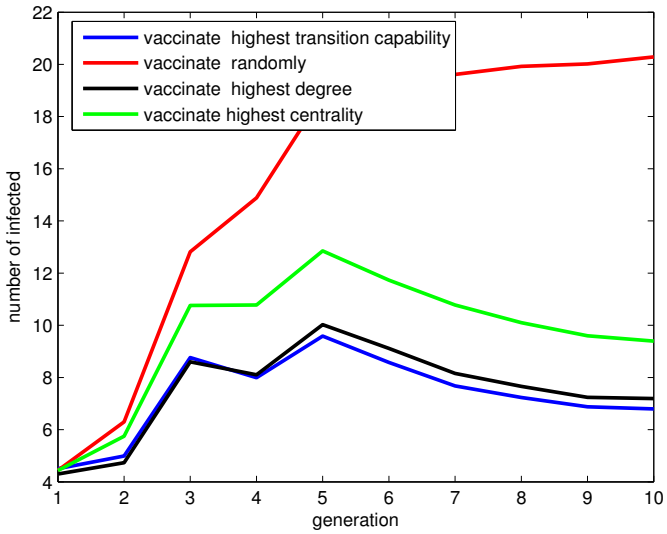


Figure 90: evolution of infected for $P_{in} = 0.05$ and $P_{out} = 0.001$

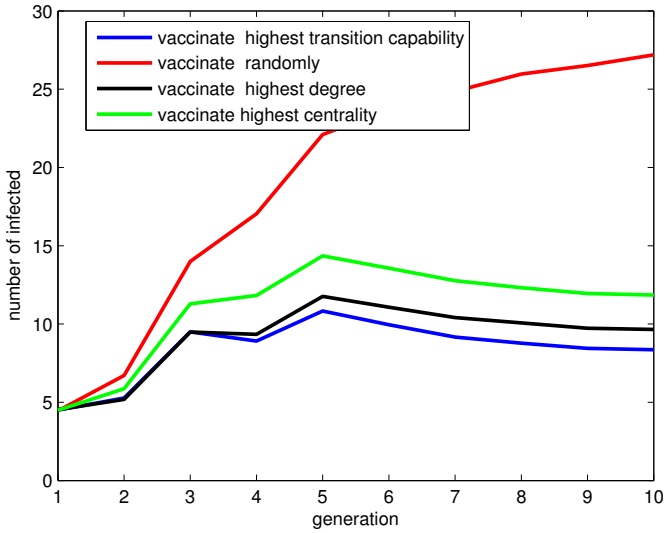


Figure 91: evolution of infected for $P_{in} = 0.05$ and $P_{out} = 0.002$

It seems in **SIRS** model, the choice of vaccination strategy makes a larger difference.

Simulation with real data set

To check the performance of these three methods in a real data set, we use the office contact network introduced in chapter 2.Both **SIS** and **SIRS** models have been employed. As described in the previous chapter, a mixed contact model is employed for $\lambda = 0.05$. Unlike the ones simulated by Erdos-Renyi method, the real networks are much less homogeneous (some nodes are much more active than others).

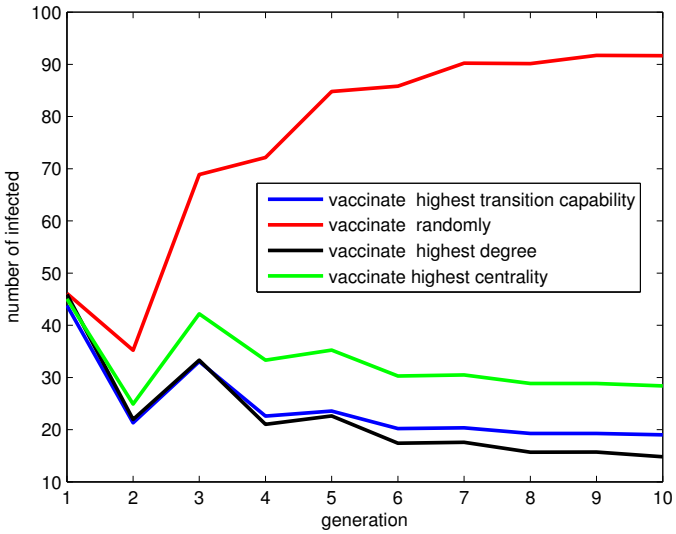


Figure 92: office data simulation with **SIS** model

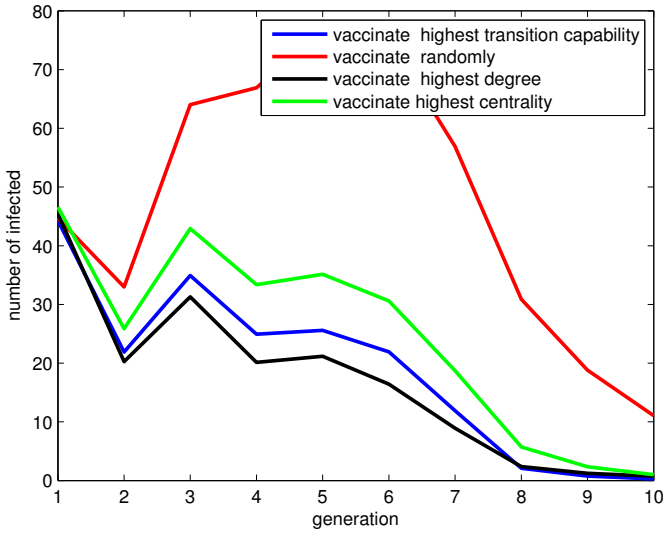


Figure 93: office data simulation with **SIRS** model

From figure 92 and 93, one observes that the shift between the random vaccination and other three other strategies are more evident. This proves the importance of a well-chosen strategy of vaccination when facing disease threat. It looks like the strategy of vaccinating nodes with the highest degree or capability measure is more efficient than using the betweenness centrality measure.

7 Conclusion

In this chapter, we have compared three different measure of nodes' transition importance in a network. Among them the transition capability measure is defined and introduced in this paper for the first time. The complexity of three algorithms in a sparse network is shown in the table below.

degree measure	$O(n)$
transition capability measure	$O(n^2)$
betweenness centrality measure	$O(n^2)log(n)$

Table 13: complexity of measures

It is difficult to announce which measure is better than others. Choosing the nodes with highest degree seems the most easy to compute. However, according to all the numerical experiments above the strategy of vaccinating nodes with the highest transition capability and betweeness centrality can be more efficient than using the degree measure. As having said it is not easy to give a proper winner in this competition, nevertheless the performance of the method using transition capability measure seems to be the most stable. In fact, in almost all numerical experiments, its result stays always in the top two.
In one word, the transition capability measure could be a strong candidate for constructing the optimal strategy of vaccination.

Appendix of chapter 3

Matlab code for testing overlapping

```
  B1 = [1 1 1 1; 1 1 1 1; 1 1 1 1 ; 1 1 1 1 ];
  N = [50,50,50,50];
  p_in=0.05;
  p_out=0.01;

  n_times=1000;% number of simulations

  M1=zeros(4);% matrix stock the results of competition

  for k=1:n_times

  [A Sim] = graphUD (B1,N,p_in,p_out);
  A=full(A);%Erods-Renyi graph construction


   C=normalize_line(A);
   D=normalize_line(A')';
   B = [C,D'];
   S=B*B'/2;%similarity measure
   E=ones(sum(N));

Sa=E-S;%anti-simularity
n=sum(N);

for i=1:n
    for j=1:n
        Sa(i,j)=(exp(Sa(i,j))-1)/exp(1);
    end
end


To=diag(A*Sa'*A);%transition capability
Ao=To;
for i=1:length(To)

    Ao(i)=Ao(i)/max(sum(A(i,:))*sum(A(:,i)),1);
end%  transition characteristic

AA=sparse(A);
 V=betweenness_centrality(AA);%betweenness centrality measure
 D=sum(A'+A);%degree: outdegree+indegree
 [B_tr,I_tr] = sort(To,'descend'); %order the nodes by measured vale
 [B_ar,I_ar] = sort(Ao,'descend');
```

```
[B_d,I_d] = sort(D,'descend');
[B_bc,I_bc] = sort(V,'descend');
fin=floor(n/10);
I_tr=I_tr(1:fin);
I_ar=I_ar(1:fin);
I_d=I_d(1:fin);
I_bc=I_bc(1:fin);

MAT=0.5*eye(4);
MAT(1,2)=1-length(setdiff(I_tr,I_ar))/fin;
MAT(1,3)=1-length(setdiff(I_tr,I_d))/fin;
MAT(1,4)=1-length(setdiff(I_tr,I_bc))/fin;

MAT(2,3)=1-length(setdiff(I_ar,I_d))/fin;
MAT(2,4)=1-length(setdiff(I_ar,I_bc))/fin;

MAT(3,4)=1-length(setdiff(I_d,I_bc))/fin;
MAT=MAT+MAT';

M1=M1+MAT;

  end
M1=M1/n_times
```

Conclusion

The conclusions of each chapter have been given separately. Here we would like to highlight again our main contribution and suggest some points to be developed.

In chapter 1

- We have defined two relations of dominance in order to compare the characteristics of probability distribution in network diffusion.
- Several properties around the new relations have been found and proved either theoretically or numerically.
- We could also suppose the contact between people are not independent. Some past related process such as the Hawkers process can be applied and studied in the future.

In chapter 2

- We have compared different definition of reproductive ratio in both aspects as average secondary infection number and the threshold of outbreak.
- Different connection structure of network has been discussed using simulations of disease spreading.

In chapter 3

- We have defined a new measure called "transition capability" which indicates the power of nodes of carrying virus from one community into another. Using this method, we don't need to identify the difference communities in the network.
- Several strategies of vaccination have been compared numerically in terms of preventing disease explosion.
- We can remove the condition that the social network is totally known. One needs to find an estimation of measures which is feasible for large population.

References

References

[1] Daniel A. Spielman, (2012). Lecture: The Adjacency Matrix and The nth Eigenvalue. , http://www.cs.elte.hu/~lovasz/eigenvals-x.pdf

[2] James Holland Jones . *Notes on R_0* , https://web.stanford.edu/~jhj1/teachingdocs/Jones-on-R0.pdf

[3] Renaud Lambiotte, Jean-Charles Delvenne and Mauricio Barahona.(2015) *Random Walks, Markov Processes and the Multiscale Modular Organization of Complex Networks* , https://arxiv.org/pdf/1502.04381.pdf

[4] Mathieu Génois et al (2015). Data on face-to-face contacts in an office building suggests a low-cost vaccination strategy based on community linkers, http://jmlr.csail.mit.edu/proceedings/papers/v9/telgarsky10a/telgarsky10a.pdf

[5] O.E.Aièllo et al., (2000), Solution of deterministic–stochastic epidemic models by dynamical Monte Carlo method http://www.sciencedirect.com/science/article/pii/S0378437100000807

[6] Nakata, Y. Röst, G. J. Math. Biol. (2015). *Global analysis for spread of infectious diseases via transportation networks* https://link.springer.com/article/10.1007/s00285-014-0801-z

[7] Seyhani Koç, Vildan Durmaz. (2015). *Airport Corporate Sustainability: An Analysis of Indicators Reported in the Sustainability Practices* http://www.sciencedirect.com/science/article/pii/S1877042815031705#

[8] Andrew J. Tatem et al. (2006). *Global traffic and disease vector dispersal* Journal of machine learning research, 3(Dec), 583-617. http://www.sciencedirect.com/science/article/pii/S1877042815031705#

[9] Richard P.Stanley. (1987). *A Bound of Spectral Radius of Graphes with e Edges* Journal of machine learning research, 3(Dec), 583-617. http://www-math.mit.edu/~rstan/pubs/pubfiles/71.pdf

[10] Upham, Paul, M. Janet, R. David, T. ve Callum. (2003). *Towards sustainable aviation* Earthscan Publications Ltd, London.

[11] Roger Guimera et al. (2005). *The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles* Proceedings of the National Academy of Sciences.

[12] Martin Rosvall and Alcides V. Esquivel et al.(2014) *Memory in network flows and its effects on spreading dynamics and community detection* , https://arxiv.org/pdf/1502.04381.pdf

[13] LC.Freeman, (1977).A set of measures of centrality based on betweenness , Sociometry The Journal of Mathematical Sociology http://moreno.ss.uci.edu/23.pdf

[14] LC.Freeman, (1977).A set of measures of centrality based on betweenness , Sociometry The Journal of Mathematical Sociology http://moreno.ss.uci.edu/23.pdf

[15] U.Brandes, (2001).A faster algorithm for betweenness centrality , The Journal of Mathematical Sociology http://www.tandfonline.com/doi/abs/10.1080/0022250X.2001.9990249

[16] S.C. Park,J.P. Draayer, (1992).Fast sparse matrix multiplication, Computer Physics Communications http://www.sciencedirect.com/science/article/pii/001046559290116G

[17] A. d'Onofrio, (2002).Pulse vaccination strategy in the sir epidemic model: Global asymptotic stable eradication in presence of vaccine failures http://www.sciencedirect.com/science/article/pii/S0895717702001772

[18] A.Browet, P.Van Dooren, (2015).Low-rank Similarity Measure for Role Model Extraction https://arxiv.org/abs/1312.4860

[19] Mathieu Génois et al.,(2015).Data on face-to-face contacts in an office building suggests a low-cost vaccination strategy based on community linkers https://hal.archives-ouvertes.fr/hal-01113431/document

[20] W. O. Kermack, A. G. McKendrick,(1927).A Contribution to the Mathematical Theory of Epidemics http://rspa.royalsocietypublishing.org/content/115/772/700

[21] DATASET: Contacts in a workplace,2016 SocioPatterns http://www.sociopatterns.org/datasets/contacts-in-a-workplace/

[22] Harold M. Hastings, (1982).The MAY-WIGNER stability theorem for connected matrices ,AMERICAN MATHEMATICAL SOCIETY https://projecteuclid.org/download/pdf_1/euclid.bams/1183549642

[23] Tassier,Troy.The Economics of Epidemiology (2013)

[24] S.Cheng, A.Laurent,P.Van Dooren, (2017).Role model detection using low rank similarity matrix, https://arxiv.org/pdf/1702.06154.pdf