UNIVERSITÀ DEGLI STUDI DI TORINO

DIPARTIMENTO DI MATEMATICA GIUSEPPE PEANO

SCUOLA DI SCIENZE DELLA NATURA

Corso di Laurea in Matematica

Tesi di Laurea Triennale

THRESHOLD MODELS OF DIFFUSION ON NETWORKS

Relatore: Paolo Cermelli

Candidato: Michele Ciruzzi

ANNO ACCADEMICO 2017/2018

Chapter 1

Introduction

The topic covered in this thesis is a class of models in network theory called threshold models of diffusion.
Network theory is a wide, and now really popular, field of research which uses graphs and their properties to model real phenomena.
Graphs are a really flexible tool to describe relationships in the real world: we can define a graph to mirror a road network, the connections on a social network, an electric grid or the flow of goods between firms.
On graphs, we can define dynamical models that describe the evolution of systems and in particular how some properties evolve.
Sometimes we are able to solve analytically the dynamics of these models but, more often, we can only get a probabilistic description of the dynamics.
The class of models we will study describes the spreading of a property from a group of vertices of a graph to the others along the graph's edges.
In particular, each vertex will get the property of interest if a certain fraction of its neighbours has already gotten it. This fraction, in fact, is a threshold used to determinate whether a vertex will get the property of interest or not, given the state of its neighbours.
Because of these features, these models are called threshold models of diffusion and they can be used to model the spreading of innovations by imitation, like a new product or a viral meme on the web, or chain failures, like blackouts or internet servers' overload.
In the next chapters we will develop a formal framework to study these models under different sets of hypotheses.
In chapter 2 we will define the mathematical objects used in the following chapters and state some conventions about notation.
In chapter 3 we will give a rigorous definition of threshold model of diffusion and prove a first general property of this class of models.
In chapters 4, 5 and 6 we will study three models proposed by famous scholars using the formal framework developed in the previous chapters.
These models will be presented from the least to the most general in hypotheses

and results.
Because of this, we will be able to find an exact solution only for the first one, while for the other two we will find a probabilistic approximation that we will compare with numerical simulations.
The programs used for numerical simulations, graph visualization and data plot in this thesis have been written in python; their source code, with the list of all the libraries used, are available online as specified in appendix A.1.

Chapter 2

Preliminary definitions

Sometimes one of the biggest difficulties that can be met reading a mathematics book is understanding the exact meaning of each symbol that appears.
In this chapter we will try to solve this problem.
In the first section we will introduce some of the notation used in this thesis, after which we will give a non-formal introduction to probability and some related properties and functions and finally we will formalize the concept of graph and network.

2.1 Notations

Tuples and Sets

We will use round brackets $(\cdot)$ for ordered sets and curly brackets $\{\cdot\}$ for unordered sets.
We will use $|\cdot|$ for set's cardinality.
Intervals of $\mathbb{R}$ will be written as $(\cdot,\cdot)$ if open and $[\cdot,\cdot]$ if closed. In case of semi-closed intervals the square bracket will stand for the closed endpoint and the round bracket for the open one.
The complement of A will be written as $\overline{A}$.
If B is a subset of A we will write $B \subseteq A$ and if we want to exclude the case $B = A$ we will write $B \subsetneq A$

Sequences

We will write a_i for the i-th element of a sequence and $\{a_i\}_i$ to refer to the sequence labelled by the index i.
We will assume that $i \in \mathbb{N}$ if not otherwise specified.

Rounding

For the rounding operator we will use the following notation: $\lceil x \rceil = \min_{n \in \mathbb{N}} n \geq x$ and $\lfloor x \rfloor = \max_{n \in \mathbb{N}} n \leq x$

2.2 Probability

As stated above, we don't aim to give a formal definition of probability, which could be the topic of a thesis by itself.
Instead, we will try to explain the intuitive idea of probability, which will be used to solve models in the following chapters. We will also propose a notation for it and, after that, we will list some of its useful properties.
An extensive discussion and a rigorous formalization of the concepts introduced in this section can be found in any undergraduate probability text, for example (Buonocore, Di Crescenzo, & Ricciardi, 2011).

General concepts

In our experience, a lot of processes don't have a deterministic outcome. Instead, there is a set of possible outcomes and if the process is replicable (and we will call this case experiment), we can find the realisation's frequencies of the different possible outcomes.
In probability theory, the set of possible outcomes is called sample space and every subset of this, i.e. each group of outcomes, is an event.
For example, the set of outcomes of rolling a six-face dice is $\{1, 2, 3, 4, 5, 6\}$ and examples of events are "the result is even" (i.e. the subset $\{2, 4, 6\}$), "the result is 5" ($\{5\}$) or "the result is strictly greater than 2" ($\{3, 4, 5, 6\}$).
If we can define a measure $\mathbb{P}(\cdot)$ that relates each event X to the frequency of its realization on a huge (potentially infinite) number of repetitions of our experiment we have found the function which expresses the probability of the events.
For how we have defined it, the probability measure has value in $[0, 1]$ (since it is a frequency), the probability of the event "the result is any element of the sample space" is 1 and if two events are mutually exclusive (i.e. in the same realization of the experiment they can't both happen), the probability that either one or the other happens is the sum of the probabilities that each one of the events happens.
When it will not be cause of ambiguity, we will write the probability of the event "the result of the experiment is one of the elements of the subset X of the sample space" as $\mathbb{P}(X)$ and the probability of the event "the result of the experiment is one of the elements of the sample space having the property y" as $\mathbb{P}(y)$ instead of writing down the whole sentence.
Finally, we can describe the map that associates each element of the sample space to its probability and call it distribution of probability.

Conditional probability

Often, we have to describe not only the probability that something happens, but also the probability that something happens given that we know that something else has happened.
We call this conditional probability and we write $\mathbb{P}(X|Y)$ to express the probability that X happens given that Y has happened.
This is equivalent to considering the probability that X happens in a new sample space formed by the elements that compose the event Y, which is still a probability function and so we have that it must be $\mathbb{P}(Y|Y) = 1$.
We can find a definition for conditional probability that satisfies this condition if we consider that the probability that both an event and itself occur is simply the probability of the event itself, and so we find that if we use as definition $\mathbb{P}(X|Y) = \frac{\mathbb{P}(X \wedge Y)}{\mathbb{P}(Y)}$, this agrees both with our intuition (X and Y both have to happen) and the condition stated above.
From this, follows that $\mathbb{P}(X \wedge Y) = \mathbb{P}(X|Y)\mathbb{P}(Y)$ and that if we can find a set E_i of mutually exclusive events such that the union of the E_i is the sample space we can state that $\mathbb{P}(X) = \sum_i \mathbb{P}(X \wedge E_i) = \sum_i \mathbb{P}(X|E_i)\mathbb{P}(E_i)$, which is called "law of total probability".
More details on conditional probability and its applications can be found in chapter 3 of (Ross, 2006).

Probability-generating function

The last topic about probability that we need to cover is the probability-generating function (or PGF), which will be extensively used in chapter 6 to solve a model proposed by Watts (Watts, 2002).
The results presented in this section are taken from a paper written by Watts itself with Newman and Strogatz (Newman, Strogatz, & Watts, 2001), in which the proofs that we will omit here can be found.
We start by defining the PGF which is, in fact, a power series, whose coefficients are the values assumed by a probability distribution p_k, defined on a discrete sample set.

Definition 2.1 (Probability-generating function)**.** Let p_k be a probability distribution defined on a discrete sample set, we define its probability-generating function as the convergent sum

$$G(X) = \sum_k p_k x^k$$

with $x \in (0, 1]$.

The name comes from the fact that, knowing $G(x)$, we can find the probability distribution that has generated it only by derivation and other elementary operations.
Now we list some properties we will use in the solution of the Watts' model mentioned above, but we will omit their proofs.

Definition 2.2 (Property of PGF)**.** Probability-generating functions have the following properties:

1. $G(1) = 1$
2. The expectation (or mean) of p_k is $\sum_k kp_k = G'(1)$
3. The PGF of the sum of m independent realizations of an experiment characterized by a probability distribution p_k is $[G(x)]^m$

2.3 Graphs

Since our goal is to describe the spreading of something from one place or person to another, it will be necessary to know along which routes it can happen.
In other words, it is necessary to know which couples of agents (or places or something else) could exchange information (or resources or are connected together).
Graphically, we can draw one point for each agent and then link together the agents with a line when appropriate, creating a net.
World airlines connections, the electric grid of a city or the friendship on social networks are clear examples which can be drawn in this way.
This intuitive idea is formalized with the mathematical notion of graph, which is a set of elements with a relation defined on them (in our examples: airport and flight routes between them, substation and electric cables or people and their friendship).

Definition 2.3 (Graph)**.** Let V be a set and $\sim$ be a symmetric and not reflexive relation on V.
We define the graph $\mathcal{G}$ as the couple (V, E) where $E = \{\{x, y\}|x \sim y\}$.
The elements of V are called vertices.
The elements of E are called edges and we will denote them by xy in place of $\{x, y\}$.

Remark. Note that for the symmetry of $\sim$ we have that xy is equivalent to yx and that they are the same element in E.
Two vertices linked by an edge are called adjacent.
To avoid ambiguity sometimes we will use $(V(\mathcal{G}), E(\mathcal{G}))$ to specify the nodes and the edges of $\mathcal{G}$.

Some scholars who use graphs as a modelling tool, particularly in networks theory, use the terms node instead of vertex and link instead of edge.
The next definition gives a name to the number of points (the order) and lines (the size) with which we can represent our graph.

Definition 2.4 (Order and size of a graph)**.** We define $p(\mathcal{G}) = |V|$, which is called order of $\mathcal{G}$, and $q(\mathcal{G}) = |E|$, which is called size of $\mathcal{G}$.

Subgraphs

Sometimes we are only interested in studying a part of a graph, maybe because we want to study the effects of diffusion on a single community or because we want to focus only on the vertices where diffusion has already happened.
In order to do so, we introduce the notion of subgraph and define a rule to create a subgraph starting from a selection of vertices, keeping only the edges of the graph between two of these selected vertices. This is called inducted subgraph.

Definition 2.5 (Subgraph)**.** We say that a graph $\mathcal{G}' = (V', E')$ is a subgraph of $\mathcal{G} = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$.

Definition 2.6 (Induced subgraph)**.** Let U be a set such that $U \subsetneq V$, we can define $E' = \{xy \in E | x, y \in U\}$.
So, we define the subgraph induced by U on $\mathcal{G}$ as $\mathcal{G}' = (U, E')$.

Neighbours and Connectivity

As we mentioned above, in order to study diffusion we are interested to know which points are connected to a given one.
The following definitions take care of it, giving us the properties to identify which and how many are the vertices that can be reached from a chosen one following a single edge, called neighbours, or with a longer walk on the graph.
If not stated otherwise, while solving our models, all vertices of graphs will be reachable from each of the others following a finite number of edges. We call those graphs connected.

Definition 2.7 (Neighbours and degree of vertices)**.** For $v \in V$ we define the set of the neighbours of v as

$$N(v) = \{w \in V | vw \in E\}$$

We define the degree of v as

$$deg(v) = |N(v)|$$

Definition 2.8 (Degree distribution)**.** We define the degree distribution function $D : \mathbb{N} \to [0, 1]$ as $D(n) = \frac{|\{v \in V | deg(v) = n\}|}{p}$.
Note that $D(n)$ is well defined as probability distribution for a random variable, and for this reason we will use the notation $p_n = D(n)$.

Definition 2.9 (Walk)**.** We call a walk between $x, y \in V$, or xy-walk, on $\mathcal{G}$ a sequence $\{v_i\}_{i \leq k}$ such that $v_0 = x$, $v_k = y$ and $\forall i < k - 1$ $v_i v_{i+1} \in E$.

Definition 2.10 (Connected graph)**.** A graph is connected if $\forall x, y \in V$ exists a xy-walk.

In the following chapters we will meet some disconnected graphs, particularly some inducted subgraphs, for the study of which it is important to be able to split the graph in as few connected subgraphs as possible.
Graphically when we draw with points and lines a disconnected graph we are drawing a set of connected graphs which have no edges between them.
With the help of the graphic intuition, it is evident how the connected subgraphs we are looking for are the maximal ones respect to the property of being connected.
So we now define the maximal connected subgraphs of a disconnected graph as connected components.

Definition 2.11 (Connected Component)**.** A connected component C_i of a disconnected graph $\mathcal{G}$ is a connected subgraph of $\mathcal{G}$ which is not subgraph of any other connected subgraph of $\mathcal{G}$.

Remark. We can extend the notion of connected components to a connected graph considering as the only connected component the graph itself.

Noteworthy graphs

Some graphs are rather famous among scholars, because they are easy to study, have specific properties or simulate well enough some empirical case studies.
Now, we will define those that will be of our interest in the following chapters.
The first one is the complete graph, in which each vertex is adjacent to all others: the graph which represents a group of friends in which everyone knows each other is an example of it.

Definition 2.12 (Complete graph)**.** A graph of order p is complete and called K_p if $\forall x, y \in V \; x \sim y$ or equivalently $xy \in E$.

The second one is indeed a wide family of graphs, the random graphs, in which we know how many the vertices are, but the edges are randomly chosen.
In fact, for each couple of vertices we define the probability that the edge which joins them belongs to the graph, and then we get a realization (for example rolling a dice) of the random graph, i.e. one particular graph of the family.
A pretty important random graph is the one in which every edge has the same probability to belong to the graph, which is called binomial random graph or Erdös-Renyi graph.
Random graphs are really important in graphs and networks theory because, if we are trying to model a real case study (like the relations between individuals in an animal population or the flow of goods between firms), we can infer the probability distribution of the edges by using statistical tools on our data, and so get a graph which is a realistic representation of our problem.

Definition 2.13 (Random graph)**.** Let V be a set of vertices and let be $x \sim y$, $x, y \in V$, with a given probability π_{xy} (which can depend on x and y).
The graph defined on V and a realization of $\sim$ is called random graph.

Definition 2.14 (Erdös-Renyi graph)**.** A random graph in which each couple of vertices $\{x, y\}$ has the same probability π to be related, i.e. be $x \sim y$, is called binomial graph or Erdös-Renyi graph.

Other examples of noteworthy graphs and a deeper discussion of their properties can be found in (van Steen, 2010).

2.4 Networks

As seen before, graphs are powerful tools to model the shape of a system and to take care of the number of agents and their relations. Though, it's really hard to use them to describe the evolution of the system through time.
All we can do, without adding in some way other information to a graph, is to add or remove vertices or edges at each step of a simulation or at each iteration of a dynamical system.
For example, if the edges represent roads, we can assign to each one its length or we can assign to each vertex representing an airport the number of passengers waiting for their flights.
We can, in this way, link some values to a graph and so we can keep the system we are studying, which is represented by the graph, constant throughout time and let only change these values.
We will focus on the functions which describe an attribute of the model by assigning a value (often called weight) to either all vertices or all edges.
Finally, we will call network a graph with one or more attributes which can change trough time, generating a dynamical system.

Definition 2.15 (Attribute function)**.** Let $\mathcal{G}$ be a graph, so we say that $f_{\mathcal{G}}$ is an attribute function of $\mathcal{G}$ if $f : V(\mathcal{G}) \to A$ or $f : E(\mathcal{G}) \to A$, with A a generic set.

Definition 2.16 (Network)**.** We define a network $\mathcal{N}$ as the couple $(\mathcal{G}, \{f_{\mathcal{G}}\})$, where $\mathcal{G}$ is a graph and $\{f_{\mathcal{G}}\}$ is a set of attribute functions on $\mathcal{G}$.

Chapter 3

Threshold models of diffusion

In this chapter, we will start to talk about the main topic of this thesis.
Firstly, we will try to explain what is a model on a network, and then we will specify which subset of models we will study, giving a rigorous definition and demonstrating a first general property.

3.1 Models on networks

Epidemic diffusion of diseases, widespread blackouts on electric grids and road congestions are different problems with a lot in common: each of these can be represented by a network that changes throughout time.
If we want to describe an epidemic disease, we create a vertex for each person and an attribute on vertices to determinate if the person is infected or not; then we draw the edges between the couples of people who can infect each other.
As another example we represent a road network, associating the roads to the edges, the crossroads to the vertices and associating to each edge the number of cars on it.
In this way, we get a mathematical object for which we are able to write some equations that describe the evolution of attributes, and so of the system, as function of time (either as continuous or discrete variable). In this way we get a model (defined) on a network.
We will only consider models in which time is discrete and the dynamics of the process, i.e. the following step, is determined only by the current state of the system, i.e. the system has no memory of the past.

Definition 3.1 (Model on a network)**.** Let $\mathcal{N} = (\mathcal{G}, \{f_{\mathcal{G}}\})$ be a network; a model on a network is a sequence $\{\mathcal{N}_t\}_t$, where $\mathcal{N}_t = (\mathcal{G}, \{f_t\})$, for which $f^i_{t+1} = g(\{f_t\}, \mathcal{G})$, with g the function which describes the evolution of the model.

3.2 Threshold models of diffusion

Now we will describe the class of models we will study in this thesis: the so called threshold models of diffusion. They aim to describe the spreading of something on a network in which each node has a threshold, whose role will be explained below. With these models, we can represent, for example, the appearance of a new technology progressively adopted by more and more agents.
In this type of process, a small group of innovators spreads the innovation throughout their social or economical network to new agents.
Another example is a chain failure on an electric grid, in which the overload of a substation could cause the failure of the other substations linked to it.
The first hypothesis we made is that once an agent has changed his state (be active or not, be innovative or not), this doesn't change again.
The idea behind this is that the change is either an advantage that is big enough not to want to renounce to it, or it is something irreversible (like a failure).
The second hypothesis is that the spreading happens because of social pressure, imitation or another form of diffusion which involves only first neighbours.
In order to formalize the second hypothesis, we assign to each node a number between 0 and 1, which is called the threshold of the node, and let a node change its state only if the fraction of its neighbours which have already changed is greater than its threshold.
These models represent a lot of phenomena in which the change is caused by overload (i.e. the node can manage only the failure of a little fraction of its neighbours) or imitation (i.e. the node tries to copy the advantage gained by its neighbours).
The fact that both the graph and the thresholds are arbitrary allows us to create a lot of different threshold models of diffusion, which can account for different relation structures, by changing the graphs, and for different behaviours of agents, by changing individual thresholds.
For example, if the nodes have a constant threshold their behaviour is influenced only by their relations, i.e. by the graph. On the other hand, we will see that different sets of thresholds (i.e. different behaviours or characteristics of nodes) produce different dynamics on the same graph.

Definition 3.2 (Threshold model of diffusion)**.** Let $s_t : V \to \{0, 1\}$, called state of nodes, and $\phi : V \to [0, 1] \subseteq \mathbb{R}$, called threshold of node.
Then the sequence $\{\mathcal{N}_t\}_t$ with $\mathcal{N}_t = (\mathcal{G}, s_t, \phi)$, a given s_0 and the evolution function

$$s_{t+1}(v) = \begin{cases} 1 & if\ s_t(v) = 1 \\ 1 & if\ \frac{|\{u \in N(v) | s_t(u) = 1\}|}{deg(v)} \geq \phi(v) \\ 0 & otherwise \end{cases}$$

is called threshold model of diffusion.

We will call initial seed the group of vertices active at the beginning of the diffusion process, i.e the set $\{v \in V | s_0(v) = 1\}$.

The evolution function defined above states that an already active node (i.e. which has already failed or innovated) remains active and an inactive node activates itself only if enough neighbours are already active, i.e. a fraction greater than the threshold.
Now we try to give a rigorous and formal proof of an intuitive fact: since an active node will never become inactive, if the nodes are a finite number, after a certain amount of time, i.e. after a finite number of steps, no more nodes change their state.
This can happen because all nodes are active or because the inactive nodes have a high enough threshold and not enough active neighbours to cause their state to change.

Theorem 3.3. *In a threshold model of diffusion if the order of $\mathcal{G}$ is finite (i.e. $|p(\mathcal{G})| < \infty$) then $\exists \mathcal{N}_\infty = \lim_{t\to\infty} \mathcal{N}_t$.*

Before proving this, we observe that only s_t depends on the time, so it is sufficient to demonstrate that it has finite limit to prove the theorem.
Since V is finite s_t is bounded from above, given that it counts vertices.
Note that the theorem is about the existence of the equilibrium and not about its uniqueness, which often doesn't stand unless we specify s_0.

Proof. Let $A_t = \{v \in V | s_t(v) = 1\}$, we have that $A_t \subseteq A_{t+1}$ because $s_t(v) = 1 \Rightarrow s_{t+1}(v) = 1$.
Also $A_t \subseteq V$ and so $\{A_t\}_t$ is a monotonic increasing sequence of set bounded from above by V and these are sufficient conditions for the existence of the limit $\lim_{t\to\infty} A_t = A_\infty$. Then we found

$$s_\infty(v) = \begin{cases} 1 & if\ v \in A_\infty \\ 0 & otherwise \end{cases}$$

□

In the next chapters we will try to characterize this equilibrium finding both exact results (in chapter 4 for models defined on complete graphs) and approximations (chapters 5 and 6).
Moreover, we will investigate the role of communities in diffusion dynamics (chapter 5) and we will prove another very important general property of threshold models of diffusion (chapter 6).

Chapter 4

Granovetter's model

In this chapter, we will analyse a model proposed by Mark Granovetter (Granovetter, 1978) to describe social systems in which each person is influenced by all the others.
Thanks to its simplicity, this is the only model for which we will be able to find exactly the equilibrium.
The example he proposed as application of this model is a protest during which a group of people starts to riot.
The idea is that some people will start to riot spontaneously and, as the riot becomes bigger, more people will join it, because joining a large group reduces the risk of a police intervention, the risk to be accused of some crimes or the social stigma.
We traduce this two observations considering a complete graph (see definition 2.12), in which each person sees all the others, and assigning a different threshold to each person. In particular, we will assign a threshold equal to 0 to those who start the riot; we will assign threshold equal to 1 to each person that never joins the riot, unless he is the last not to have joined it.

Definition 4.1 (Gravovetter's model)**.** We will call Granovetter's model a threshold model of diffusion with $\mathcal{G} = K_p$ complete graphs of order p and $s_0(v) = 0 \ \forall v \in V(\mathcal{G})$.

The second part of the definition accounts for the fact that the riot has still to arise. We will see in the section 4.2 how to generalize this hypothesis.

4.1 Equilibrium in Granovetter's model

We have already proven (theorem 3.3) that every threshold model of diffusion has at least one equilibrium and, for Granovetter's ones, we are able to find an exact solution for the problem.
What simplifies the problem is considering complete graphs because, in this way, every node knows the system globally.

First of all, we will define the function $T(n)$ which counts how many nodes will be active if there are other n nodes active:

$$T(n) : \mathbb{N} \to \mathbb{N},\ T(n) = |\{v \in V | (p-1)\phi(v) \leq n\}| \tag{4.1}$$

which considers the threshold ϕ and the $p-1$ neighbours (since the order is p) of each vertex.
Now, to study the dynamics of the diffusion, we will study the number of active nodes at the time t

$$a_t = |\{v \in V | s_t(v) = 1\}| \tag{4.2}$$

and since $a_0 = 0$ we look for the smallest equilibrium, i.e. for the smallest number for which $T(n) = n$ which means that all nodes that could be activated are already active (note that to get this result it is fundamental that in this model all nodes start inactive and activate themself only as effect of the activation of the other nodes).

Theorem 4.2. *In a Granovetter's model* $s_\infty(v) = 1 \iff (p-1)\phi(v) \leq a^*$ *with* $a^* = \min\{x \in \mathbb{N} | T(x) = x\}$

Remark. From theorem 3.3 we know that the model has at least one equilibrium, which means that T has at least one fixed point (i.e. a value x for which $T(x) = x$). In fact, because $\phi \leq 1 \Rightarrow (p-1)\phi(v) \leq p-1 < p\ \forall v$ we have that $T(p) = p$, i.e. if all nodes are active each node has exceeded its own threshold because all its neighbours are active.
Our proof will show that under the hypothesis $s_0(v) = 0\ \forall v$ the equilibrium reached could be different, i.e. T could have more than one fixed point.

Proof. $\Leftarrow$ We need to demonstrate that all the nodes with $(p-1)\phi(v) \leq a^*$ are active at the equilibrium.
We start by proving that $T(a)$ is monotonic non decreasing in a which it's equivalent to say that if $(p-1)\phi(v) \leq \bar{a}$ then $(p-1)\phi(v) \leq a,\ \forall a \geq \bar{a}$.
Because all nodes are inactive at $t = 0$, and so $a_0 = 0$, we have that at time $t+1$ the active nodes are those with $(p-1)\phi(v) \leq a_t$ which is equivalent to say $a_{t+1} = T(a_t)$.
In every threshold model of diffusion $a_{t+1} \geq a_t$ and so $T(a_t) \geq a_t\ \forall t$.
If $T(0) = 0 \Rightarrow a^* = a_0 = 0$, which is for sure the smallest fixed point.
But $T(0) = 0 \Rightarrow \nexists v$ s.t. $\phi(v) \leq 0$ and $a^* = 0 \Rightarrow \nexists v$ s.t $s_\infty(v) = 1$, which is the thesis.
Otherwise since $a_t \leq p$ by definition it must exist at least one t^* for which $a_{t^*} = a_{t^*+1}$ and since $\mathcal{N}_{t+1}$ depends only on $\mathcal{N}_t$ this is an equilibrium, and a_{t^*} is a fixed point.
Let a^* be the smallest a_{t^*}, and so we have $a_t < a_{t+1}\ \forall a_t < a^*$.
Since $a^* = T(a^*)$, by the definition of T only the a^* nodes with $(p-1)\phi(v) \leq a^*$ will be active at the equilibrium, which is the thesis if a^* is the smallest fixed point of T.
But if another fixed point $\bar{a} < a^*$ exists, it must exist a $\bar{t}$ such that $a_{\bar{t}} < \bar{a} < a_{\bar{t}+1}$

and by the monotonicity of T we have that $a_{t+1} = T(a_t) < T(\bar{a}) = \bar{a}$, which is a contradiction.
$\Rightarrow$ Let be $A_\infty = \{v \in V | s_\infty(v) = 1\}$, we have that $\forall v \in A_\infty \ \phi(v) \leq \frac{|A_\infty|-1}{p-1} < \frac{|A_\infty|}{p-1}$.
So we note that, since we have reached the equilibrium, $|A_\infty| = T(|A_\infty|)$ and so if we choose $a^* = |A_\infty|$ we have the thesis.
Note that a^* must be the least fixed point because if another smaller fixed point exist not all the a^* nodes in A_∞ will be active at the equilibrium, which is a contradiction. □

Now, with the help of an example proposed by Granovetter himself, we highlight how the equilibrium of the model may not be very stable.
In fact, there are situations in which the change of the threshold of a single vertex causes a big change in the equilibrium.

Example 4.3. Let us define a bijection $l : V \to \{n \in \mathbb{N} | n < p\}$ which labels the nodes and define $\phi(v) = \frac{l(v)}{p-1}$ to assign the thresholds $\{0, \frac{1}{p-1}, \ldots, \frac{p-2}{p-1}, 1\}$ to the vertices.
We see that with these thresholds $a_t = t$ for $t \leq p$ after which we reach the equilibrium.
Now, if for the vertex labelled with n we substitute its threshold with an higher value the equilibrium is reached when n vertices will be activated (the vertices from 0 to $n-1$).
And if this vertex is the one with label 0 the spreading doesn't start at all.
We have represented this situation with the two graphs in figure 4.1 changing the threshold of the vertex 0 to $\frac{1}{p-1}$ in the right plot.
The fixed points, i.e. the possible equilibria, can be obtained as the intersections between the function and the line $y = x$.
We see on the left plot that the only equilibrium is p while on the right plot, where an agent has increased his threshold only by $\frac{1}{p-1}$ (which is very small in a large graph), the minimum fixed point is 0.
This is obviously a pathological example, but it is useful to highlight that little changes of the threshold for a little fraction of agents could cause a big change of the equilibrium.

4.2 Generalized Granovetter's model

Granovetter's model doesn't account for all threshold diffusion models defined on a complete graph, but only for those that are characterized by all inactive nodes when $t = 0$.
Now we will show how to transform any threshold diffusion model defined on a complete graph in a Granovetter's model.

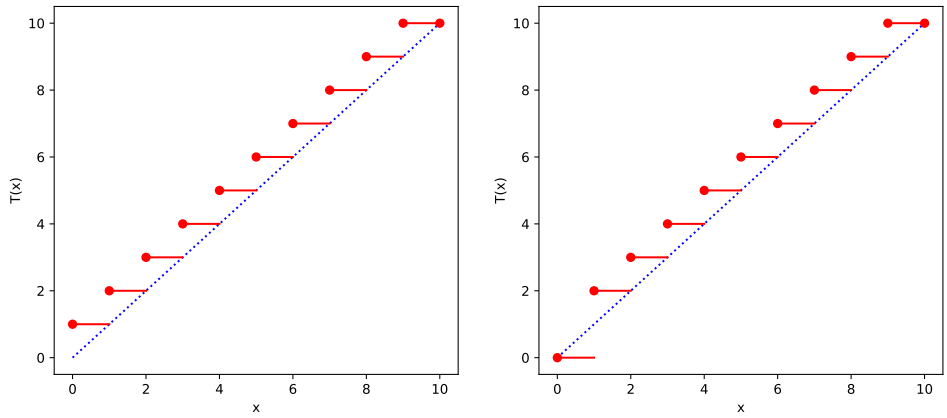

Figure 4.1: On the left, there is the plot of the function T for a graph of order 10 with the thresholds defined in example 4.3 and, on the right, there is the same plot for the same graph with the thresholds obtained changing the individual threshold of the vertex 0 to $\frac{1}{p-1}$. The dotted line represents $y = x$, needed to find fixed points, and the red dots are the included endpoints of intervals.

Theorem 4.4. *Each threshold model of diffusion on a complete graph K_p with threshold $\phi(v) : V \to (0, 1]$ is equivalent to a Granovetter's model.*

Proof. If our model is described by $\{\mathcal{N}_t\}_t$ with $\mathcal{N}_t = (K_p, \phi, s_t)$ we are looking for a Granovetter's model $\{\mathcal{M}_\tau\}_\tau$ with $\mathcal{M}_\tau = (K_p, \psi, r_\tau)$ such that we can find a bijection between τ and t such that $s_t(v) = r_\tau(v)\ \forall v$.
We observe that $\tau = t + 1$ and

$$\psi(v) = \begin{cases} 0 & if\ s_0(v) = 1 \\ \phi(v) & otherwise \end{cases}$$

solves our problem because at $\tau = 0$ all nodes are inactive, at $\tau = 1 \to t = 0$ only the nodes initially active in $\mathcal{N}$ are active, and for $\tau > 1$ the two systems are indistinguishable because the threshold of active nodes does not influence the dynamics of the model. □

Remark. The hypothesis that no vertex has threshold equal to 0 is necessary to create a complete equivalence.
Otherwise, a vertex with threshold 0 which doesn't belong to the initial seed should become active at $t = 1 \to \tau = 2$ but, under our transformation, it becomes active at $\tau = 1$.
It is possible anyway to demonstrate that, without the additional hypothesis on the threshold, the equilibrium of $\mathcal{N}$ and $\mathcal{M}$ is the same, working by contradiction on the set of vertices active at the equilibrium of $\mathcal{N}$ but not at the one of $\mathcal{M}$, which is obviously empty.

Chapter 5

Nematzadeh's model

The next model we will study was proposed by Nematzadeh and other scholars (Nematzadeh, Ferrara, Flammini, & Ahn, 2014) to study the role of communities in the diffusion processes.
With the term community we mean a group of nodes with a lot of edges amongst them and with few edges between them and other nodes of the graph.
In this chapter we will introduce a graph with only two communities, but in the appendix of the paper cited above is stated that the results found hold also with more than two communities.
This chapter will be divided into three sections: in the first one we will characterize the graph proposed by Nematzadeh, in the second one we will find a probabilistic expression for the number of active nodes at equilibrium in the diffusion process and in the third one we will simulate some graphs and compare them with the result found in the second section.

5.1 Nematzadeh's graph

The goal of this section is to define a random graph which has two well recognisable communities.
To obtain this, after having set the size and the order of the graph, we divide the vertices in half. Then, we use a small fraction μ of the edges to link vertices that belong to two different halves, and all the other edges to link vertices belonging to the same half.

Definition 5.1 (Nematzadeh's graph)**.** We call Nematzadeh's graph a connected graph of order p even and of size q for which $\{A, B\}$ is a partition of vertices with $|A| = |B| = \frac{p}{2}$ and $\exists \mu \in (0, 1)$ such that $|\{xy \in E | x \in A, y \in B\}| = \mu q$.

In figure 5.1 we see the effects of the parameter μ on the graph's structure: with very low values of μ (i.e. less than 0.1) almost every edge links nodes in the same community, while with μ near or above $\frac{1}{3}$ the community structure is lost since there is at least the same number of edges between communities and in each of the

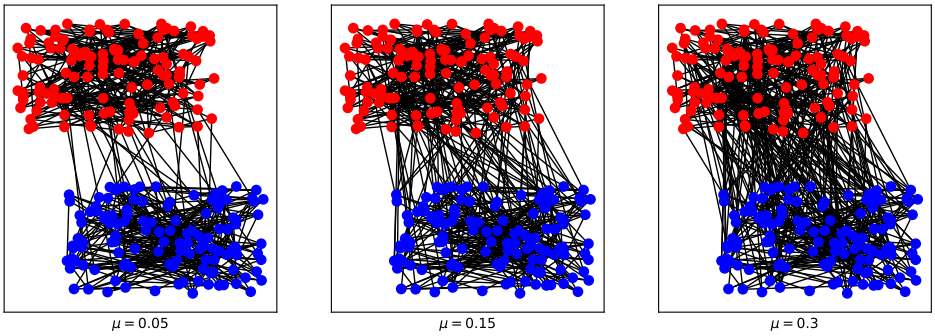

Figure 5.1: The three graphs represented here are Nemaztadeh's graphs of order 250 and size 500 with μ respectively, from left to right, 0.05, 0.15 and 0.30. It is clear that low values of μ imply a low connectivity between communities, while high values of μ make the communities unrecognisable.

two communities.
Now we will obtain the degree distribution of this type of graph, which will be useful to compare theoretical results with simulations in the next sections.
Our argument is based on the one used by van Steen to find the degree distribution of an Erdös-Renyi graph (van Steen, 2010, section 7.2.1).
First of all, we can break down the degree of a vertex v into the sum of the number of edges from v to another vertex in the same community (and we will call this number $deg_{in}(v)$) and the number of edges from v to a vertex in the other community (and we will call this number $deg_{out}(v)$). So $deg(v) = k \Rightarrow deg_{in}(v) + deg_{out}(v) = k$.
From this follows that

$$\mathbb{P}(deg(v) = k) = \sum_{n=0}^{k} \mathbb{P}(deg(v) = k | deg_{out}(v) = n)\mathbb{P}(deg_{out}(v) = n) =$$
$$= \sum_{n=0}^{k} \mathbb{P}(deg_{in}(v) = k - n)\mathbb{P}(deg_{out}(v) = n)$$

Now we consider the two subgraphs composed by all the vertices of $\mathcal{G}$ and either the intra-community edges (and we will call this subgraph $\mathcal{G}_{in}$) or the inter-community edges ($\mathcal{G}_{out}$).
These two subgraphs are random graphs obtained by selecting only a fraction of the possible edges. Because of this, we can use the same argument used by van Steen for the Erdös-Renyi graph to get the degree distribution.
In $\mathcal{G}_{in}$ we want to select $(1-\mu)q$ edges from the $\frac{1}{2}p(\frac{p}{2}-1)$ possible ones, because for each vertex we can choose one of the other vertices in the same community but we have to consider only half of the edges to account for symmetry (i.e. consider either xy or yx and not both).
In $\mathcal{G}_{out}$ we want to select μq edges from the $(\frac{p}{2})^2$ possible ones, because for each

vertex in one of the two communities we can choose all the vertices in the other one.
So we find that the probabilities π to choose a given edge in each subgraph are

$$\pi_{in} = \frac{4(1-\mu)q}{p(p-2)} \qquad \pi_{out} = \frac{4\mu q}{p^2} \tag{5.1}$$

Finally, to obtain the degree distribution, we need to remember that the probability to choose k elements with a certain property from a set of n elements in which a fraction π has the properties of our interest can be described with a binomial distribution, which is expressed by

$$\binom{n}{k}\pi^k(1-\pi)^{n-k}$$

Given a vertex, we look for n neighbours in $\mathcal{G}_{out}$ amongst the $\frac{p}{2}$ vertices in the other community and for $k-n$ neighbours in $\mathcal{G}_{in}$ amongst the others $\frac{p}{2}-1$ vertices in the same community.
And so we get

$$\begin{aligned}\mathbb{P}(deg(v)=k) = \sum_n \Bigg[&\binom{\frac{p}{2}}{n} {\pi_{out}}^n (1-\pi_{out})^{\frac{p}{2}-n} \cdot \\ &\cdot \binom{\frac{p}{2}-1}{k-n} {\pi_{in}}^{k-n} (1-\pi_{in})^{\frac{p}{2}-1-(k-n)} \Bigg]\end{aligned} \tag{5.2}$$

In figure 5.2 we see how our theoretical prediction fits the mean degree distribution of 1000 different Nematzadeh's graphs with constant order 100, size 1000 and $\mu = 0.20$.

5.2 Equilibrium in Nematzadeh's model

Now we can define threshold model of diffusion on the graph studied above.
We are interested in studying how the variations of μ (and the initial seed) change the dynamics of the diffusion.
To simplify the study of the dynamics, we fix the threshold of each vertex to the same value θ and we choose the initial seed as a subset of the community A (or equivalently all vertices of B are inactive at $t = 0$).

Definition 5.2 (Nematzadeh's model)**.** We call Nematzadeh's model a threshold diffusion model defined on a Nematzadeh's graph with threshold $\theta \in (0, 1)$ constant for each vertex and $s_0(v) = 0\ \forall v \in B$.

As we have already done in section 4.1 for Granovetter's models, we are interested in finding the fraction of active nodes at equilibrium.
Since we are studying the influence of communities, we want to find an expression for both the fraction of active nodes in A and the fraction of active nodes in B.

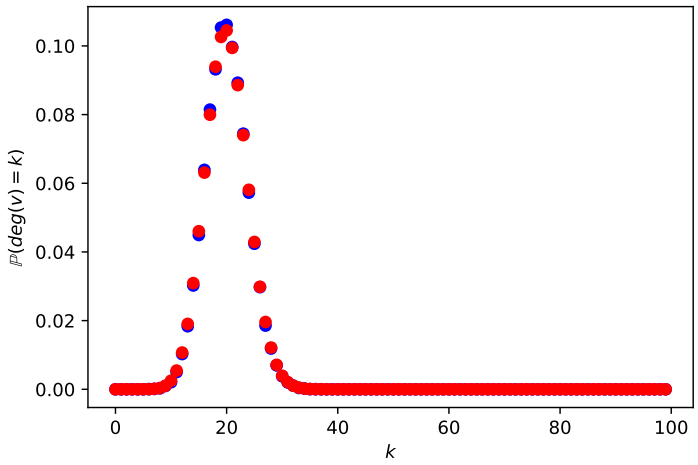

Figure 5.2: The red dots represent the theoretical degree distribution of a Nematzadeh's graph with $p = 100$, $q = 1000$ and $\mu = 0.20$ obtained with equation 5.2. The blue dots represent the mean of the degree distributions of 1000 simulated Nematzadeh's graphs with the same parameters.

We will try to find an approximate solution using Mean Field approximation, i.e. we will try to express the probability that a vertex is active at equilibrium as a function of itself, and we will look for stable solutions, in particular the least stable solution.
We will start by considering a generic graph, then we will adapt the equation found to a Nematzadeh's graph, taking into account the two-components structure.
So we define the fraction of nodes active at $t = 0$ $\rho_0 = \frac{|\{v|s_0(v)=1\}|}{p}$, the fraction of nodes active at equilibrium $\rho_\infty = \frac{|\{v|s_\infty(v)=1\}|}{p}$ and the empirical degree distribution $p_k = \frac{|\{v \in V|deg(v)=k\}|}{p} \simeq \mathbb{P}(deg(v) = k)$.
We compute ρ_∞ as

$$\begin{aligned}
\rho_\infty &= \mathbb{P}(s_\infty(v) = 1) = \\
&= \mathbb{P}(s_0(v) = 1) + \mathbb{P}(s_0(v) = 0)\mathbb{P}(|\{u \in N(v)|s_\infty(u) = 1\}| \geq \theta deg(v)) = \\
&= \mathbb{P}(s_0(v) = 1) + \\
&\quad + \mathbb{P}(s_0(v) = 0) \sum_k \mathbb{P}(deg(v) = k)\mathbb{P}(|\{u \in N(v)|s_\infty(u) = 1\}| \geq \theta k)
\end{aligned}$$

and we find

$$\rho_\infty \simeq \rho_0 + (1 - \rho_0) \sum_k p_k \sum_{m=\lceil \theta k \rceil}^{k} \binom{k}{m} {\rho_\infty}^m (1 - \rho_\infty)^{k-m} \tag{5.3}$$

This approximation gets better when the empirical degree distribution is close to the theoretical one: that is when p and q become bigger, even if it remains an approximation and not an exact solution.

Another problem is that we have found a closed form, which is not analytically solvable, and so we can find solutions only by numerical approximation.
To generalize the result to Nematzadeh's graphs we introduce the fraction of active nodes in A at $t = 0$ $\rho_0^A = \frac{|\{v \in A | s_0(v)=1\}|}{p/2}$, the fraction of active nodes in A at equilibrium $\rho_\infty^A = \frac{|\{v \in A | s_\infty(v)=1\}|}{p/2}$, the probability that a vertex adjacent to a chosen one is active at equilibrium $q^A = (1-\mu)\rho_\infty^A + \mu\rho_\infty^B$ and the correspondent ones on B.
Note that ρ_∞ is the mean of ρ_∞^A and ρ_∞^B.
Equation 5.3 becomes, on a Nematzadeh's graph,

$$\rho_\infty^A = \rho_0^A + (1-\rho_0^A)\sum_k p_k \sum_{m=\lceil \theta k \rceil}^{k} \binom{k}{m} {q^A}^m (1-q^A)^{k-m} \tag{5.4a}$$

$$\rho_\infty^B = \rho_0^B + (1-\rho_0^B)\sum_k p_k \sum_{m=\lceil \theta k \rceil}^{k} \binom{k}{m} {q^B}^m (1-q^B)^{k-m} \tag{5.4b}$$

Figure 5.3 shows the solution of equations 5.4 for a graph of order 100, size 1500, threshold 0.4 and $\rho_0^A = 0.4$ using the theoretical degree distribution 5.2.
We see that the change of μ causes two phase transitions (i.e. two fast and big changes in model's dynamics): low values of μ cause an almost complete diffusion in A but no diffusion in B (i.e. there is not enough connectivity between communities to spread the diffusion); intermediate values of μ cause an almost complete diffusion in both communities; high values of μ cause little diffusion in A and no diffusion in B, probably because there is not enough connectivity in either community to start the diffusion.

5.3 Simulations of Nematzadeh's model

Our goal is to highlight the phase transition and the range of μ and ρ_0 for which it happens.
To achieve this, we have fixed $p = 100$, $q = 1500$ and $\theta = 0.3$ and we have let μ and ρ_0^A change in $[0, 1]$ with a step of 0.025.
For each combination of μ and ρ_0^A we have both solved numerically the equation 5.4 (the first row of figure 5.4) and simulated the diffusion process on 10 different connected (for $\mu \neq 0$) and randomly generated Nematzadeh's graphs, choosing for each 10 different random initial seeds on each. The second row of figure 5.4 represents the average of these 100 simulations for each different combination of μ and ρ_0^A.
The last plot of each row compares ρ_∞^A, ρ_∞^B and ρ_∞ as functions of μ for a chosen value of ρ_0^A.
First of all, we observe an excellent qualitative and a good quantitative agreement between our model and the results of the simulation, which show a slightly slower

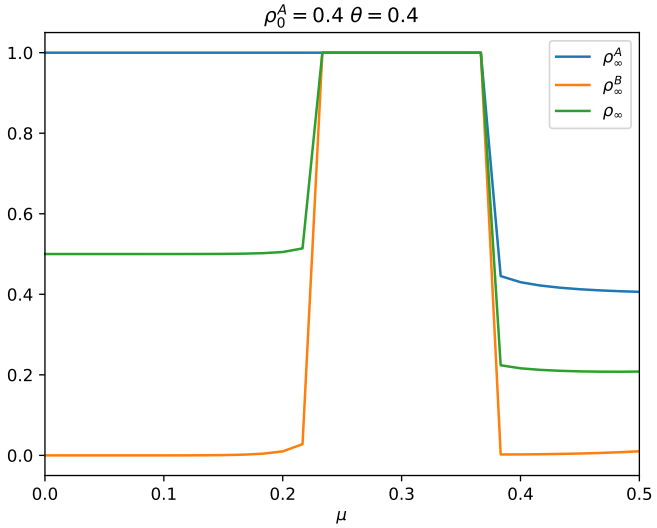

Figure 5.3: This plot represents the solution of equation 5.4 as function of μ, fixed $p = 100$, $q = 1500$, $\theta = 0.4$ and $\rho_0^A = 0.4$. The blue line represents ρ_∞^A, the orange line represents ρ_∞^B and the green line represents ρ_∞, which is the mean of ρ_∞^A and ρ_∞^B.

diffusion.
We can draw some conclusions from these data:

1. for low values of ρ_0^A there aren't enough active vertices to start the diffusion.

2. for low values of μ there aren't enough "bridges" between the two communities to spread the diffusion from one to the other.

3. for high values of ρ_0^A we observe only one phase transition between a state in which A is completely activated and B is inactive and one in which the whole network is active.

4. there are values of ρ_0^A for which as μ changes we meet more than one phase transition: in fact, for low values of μ the diffusion is complete in A and absent in B; then there is a range for μ in which the diffusion is complete on the whole network; then for higher values of μ we lose the two-community structure and the diffusion doesn't happen neither in A nor in B (because is the community structure which trigger and boost the diffusion).

5. for $\mu > 0.5$ the dynamics change and a third phase transition toward a complete diffusion change: $\mu = 1$ represents in fact a bipartite graph (i.e. a two-community graph without intra-community edges) and so in the range $[0.5, 1]$ we go from a chaotic graph (someway similar to an Erdös-Renyi graph) to an organized graph, in which a form of community structure, which helps to trigger diffusion, appears.

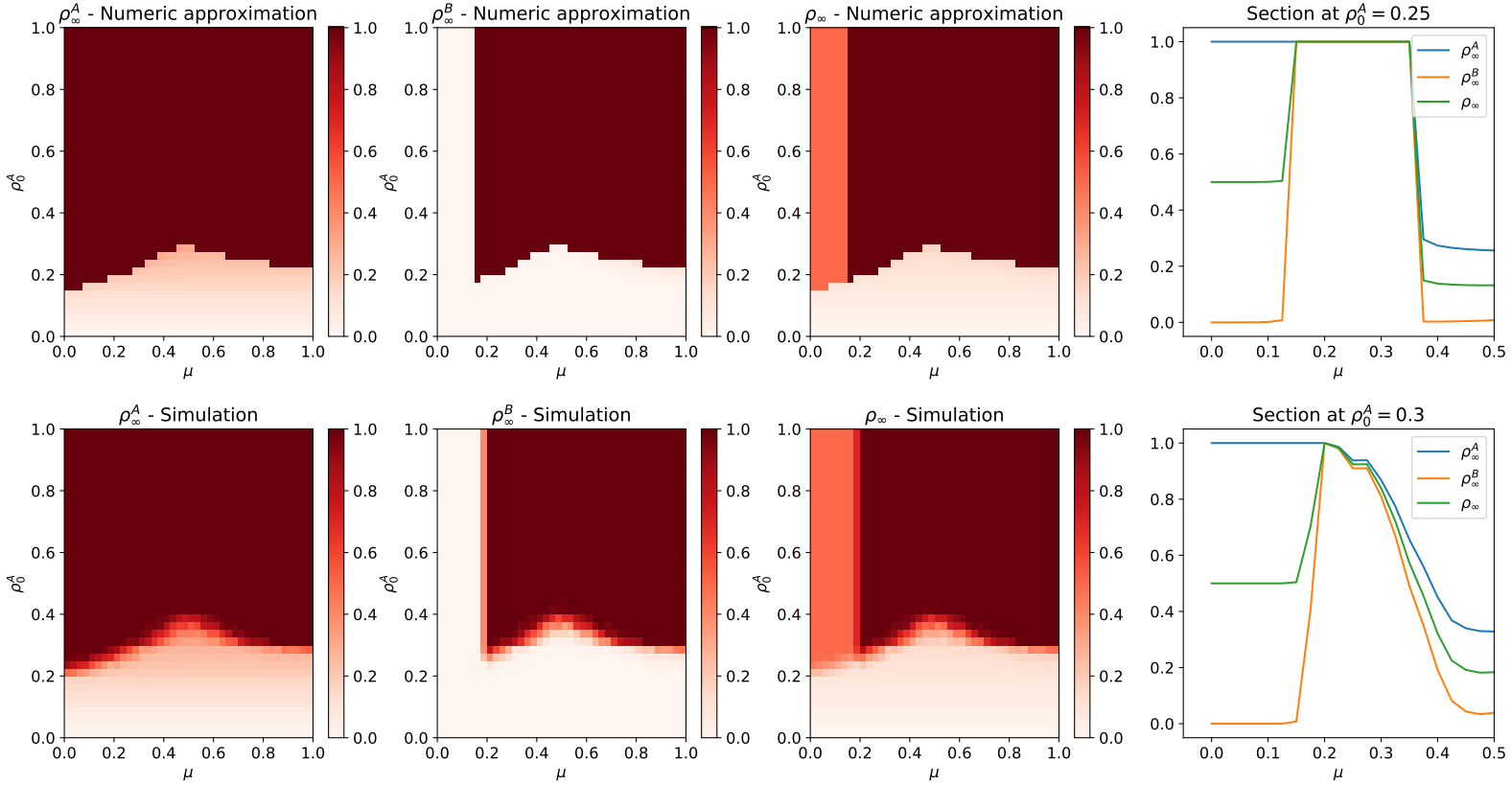

Figure 5.4: The first row has been obtained by solving numerically equation 5.4 and the second row has been obtained by observing for each combination of μ and ρ^A_0 (with a step of 0.025) the diffusion dynamics caused by 10 different samples in A on 10 different Nematzadeh's graphs (i.e. each dot of the plot is the mean of 100 different observations, 10 for each of 10 graphs).
From left to right, the columns show the values of ρ^A_∞, ρ^B_∞ and ρ_∞ (where dark red is 1 and white is 0) and a section of the previous three, which highlights the double phase transition.
For all plots we have fixed $p = 100$, $q = 1500$ and $\theta = 0.3$.

The observations listed here also hold for different θs. Analogous plots with different θs, but same order and size, are reported in appendix A.2.
The code written for the simulations and to get the numerical solutions is available online as specified in appendix A.1.

Chapter 6

Watts' model

The model proposed by Watts (Watts, 2002) is the most general we will analyse and, because of this, we will only be able to get a lower bound for the number of vertices active at the equilibrium (under some additional hypotheses).
Nevertheless, the results obtained by Watts are really important because they stand for almost every graph and threshold distribution.

Definition 6.1 (Watts' model)**.** We will call Watts' model a threshold model of diffusion defined on an arbitrary graph $\mathcal{G}$, whose degree distribution is known, with threshold distribution known and $s_0 = 0 \; \forall v \in V(\mathcal{G}) \setminus \{v_0\}$.

The assumption that the initial seed is composed by a single vertex is needed to simplify the solution of the model, but is not strictly necessary. In fact in his paper Watts uses a little bit more relaxed hypothesis.
We will focus on vertices for which one active neighbour is enough to become active and we will call them vulnerable vertices.
The idea we will explore is that if we find a connected subgraph (called a cluster) composed entirely by vulnerable vertices, the activation of one vertex in the cluster will cause a cascade activation, after which all nodes belonging to the cluster will be active.
So the dimension of the biggest cluster is a lower bound for the number of the active vertices at the equilibrium if the initial seed belongs to the cluster (which is more likely the bigger the cluster is).
Once a big cluster is active it is likely that there are enough active vertices to also activate some non vulnerable vertices, causing a cascade effect.

6.1 Vulnerable clusters in Watts' model

The goal of this section is to find an approximation for the number of vulnerable vertices in a vulnerable clusters.
In order to do so, we will use the probability generating functions and their properties (see section 2.2) to carry out calculations.

First of all, we define the set $U = \{v \in V | deg(v) \leq \lfloor \frac{1}{\phi(v)} \rfloor\}$ which is the set of the vulnerable vertices, i.e. the vertices which become active if at least one of the neighbours is active.
To simplify the notation we define

$$p_k = \mathbb{P}(deg(v) = k) \qquad \rho_k = \mathbb{P}(v \in U | deg(v) = k) = \mathbb{P}(\phi(v) \leq \frac{1}{k})$$

because if $deg(v) = k$ then $v \in U \Rightarrow k \leq \lfloor \frac{1}{\phi(v)} \rfloor \leq \frac{1}{\phi(v)} \Rightarrow \phi(v) \leq \frac{1}{k}$.
By the properties of conditional probability, we also have that $\mathbb{P}(v \in U) = \sum_k p_k \rho_k$ and $\mathbb{P}(v \in U \wedge deg(v) = k) = p_k \rho_k$ (note that this is a joint probability of a Bernoulli random variable, i.e. the vulnerability, and another random variable, i.e. the degree distribution).
The probability generating function related to $p_k \rho_k$ is, pairing vulnerable with 0 and not vulnerable with 1, $\sum_k p_k \rho_k x^k y^0 + \sum_k p_k \bar{\rho_k} x^k y^1$, with $\bar{\rho_k} = \mathbb{P}(\phi(v) > \frac{1}{k}) = 1 - \rho_k$, and its limit for $x, y \to 1^-$ is 1 for the normalization condition and so we can define

$$G(x) = \sum_k p_k \rho_k x^k \tag{6.1}$$

whose normalization condition is

$$\lim_{x \to 1^-} G(x) = \mathbb{P}(v \in U) = \frac{|U|}{p(\mathcal{G})} \tag{6.2}$$

Note that for the properties of the PGF the mean degree of vulnerable vertices is

$$z_U = \sum_k k p_k \rho_k = G'(1)$$

while the mean degree of all verteices is

$$z = \sum_k k p_k$$

Now we want to find the analogue of $G(x)$ for vulnerable nodes adjacent to a chosen vertex v^*.
So chosen $v \in U$ we need to find $\mathbb{P}(deg(v) = k \wedge v \in U \wedge v \sim v^*)$.
The probability to be adjacent to a vertex must be proportional to the number of neighbours of v (i.e.the degree of v) and must be function of vulnerability only through the degree of v.
So we can write

$$\begin{aligned}
\mathbb{P}(deg(v) = k \wedge v \in U \wedge v \sim v^*) = \\
&= \mathbb{P}(v \sim v^* \wedge v \in U | deg(v) = k) \mathbb{P}(deg(v) = k) = \\
&= \mathbb{P}(v \sim v^* | deg(v) = k) \mathbb{P}(v \in U | deg(v) = k) \mathbb{P}(deg(v) = k) = \\
&= c \cdot deg(v) p_k \rho_k = c \cdot k p_k \rho_k
\end{aligned}$$

where c is a normalization constant.
Because of the same argument used above, by the normalization condition we have

$$\begin{aligned} 1 &= c(\sum_k kp_k\rho_k + \sum_k kp_k\bar{\rho_k}) = c\sum_k kp_k(\rho_k + \bar{\rho_k}) = \\ &= c\sum_k kp_k(\rho_k + 1 - \rho_k) = c\sum_k kp_k \\ &\Rightarrow c = \frac{1}{\sum_k kp_k} = \frac{1}{z} \end{aligned}$$

.
In order to internalize the adjacency hypothesis we can consider the graph induced by $V^* = V \setminus \{v^*\}$ (which is V without the edges starting from v^*), in which we have for the neighbours of v^* $deg_{V^*}(v) = deg_V(v) - 1$ and so the correspondent PGF has to be

$$G_*(x) = \frac{\sum_k kp_k\rho_k x^{k-1}}{z} = \frac{G'(x)}{z} \tag{6.3}$$

and its normalization condition gives

$$G_*(1) = \frac{z_U}{z} \tag{6.4}$$

Now we focus on the subgraph induced by U in which we can identify the sets C_i of the vertices of each connected component and we can define $C(v)$ as the set C_i such that $v \in C_i$.
We define the probability for a random vertex to belong to a vulnerable cluster of dimension n (i.e. composed by n vertices), and its PGF

$$q_n = \mathbb{P}(|C(v)| = n) \qquad\qquad H(x) = \sum_n q_n x^n$$

and the same probability given that the vertex is adjacent to a chosen v^*

$$r_n = \mathbb{P}(|C(v)| = n | v \sim v^*) \qquad\qquad H_*(x) = \sum_n r_n x^n$$

analogously as we did with $G(x)$ and $G_*(x)$.
If we assume that there are no cycles in the U-induced graph (or otherwise in order to get an approximate result) we could "decompose" $H_*(x)$ in a part that accounts for non vulnerable neighbours and a part which is the sum of the dimensions of the vulnerable clusters in $V \setminus \{v^*\}$ to which the neighbours of v^* in V belong (using the third property stated in definition 2.2).
In this way, the sum of the dimension of the clusters, to which the k neighbours of a random vertex v belong, weighted by the probability that v is vulnerable and has k neighbours is (using the PGFs) $\sum_k p_k\rho_k H_*(x)^k = G(H_*(x))$. We can use an analogue argument for the neighbours of a chosen vertex v^*.

So we find (multiplying for x to account for v^* or the random v, analogously as we did to get equation 6.3)

$$H_*(x) = [1 - G_*(1)] + xG_*(H_*(x)) \tag{6.5}$$

and likewise

$$H(x) = [1 - G(1)] + xG(H_*(x)) \tag{6.6}$$

The average number of vertices in the cluster $\langle n \rangle$ is the expectation of q_n, or $H'(1)$ by the properties of the PGF.
Note that $\langle n \rangle$ is a lower bound for the dimension of the biggest cluster.
Using equations 6.6 and 6.5 we find

$$\begin{aligned}\langle n \rangle = H'(1) &= \lim_{x \to 1} H'(x) = \\ &= \lim_{x \to 1^-} G(H_*(x)) + xG'(H_*(x)) \frac{G_*(H_*(x))}{1 - xG^{*\prime}(H_*(x))} = \\ &= G(1) + G'(1) \frac{G_*(1)}{1 - {G_*}'(1)}\end{aligned}$$

where we have used the normalization properties of the PGFs.
From equations 6.2, 6.3 and 6.4 we have

$$\langle n \rangle = \frac{|U|}{p} + {z_U} \frac{z_U / z}{1 - G''(1)/z} = \frac{|U|}{p} + {z_U}^2 \frac{1}{z - G''(1)} \tag{6.7}$$

which diverges when

$$z = G''(1) = \sum_k k(k-1) p_k \rho_k \tag{6.8}$$

Note that a cluster $C^* \in \{C_i\}$ such that $|C^*| \geq \langle n \rangle$ always exists for the definition of mean and so if $\langle n \rangle$ diverges also $|C^*|$ diverges.
Also note that for increasing k we have that $k(k-1)$ is increasing while ρ_k is decreasing: this suggests that $G''(1) - z = 0$ has zero or two solutions.
Each solution represents a phase transition between a configuration with small vulnerable clusters, in which the initial seed is probably not vulnerable and only few vertices will be eventually involved in a cascade activation, and one with big vulnerable clusters, in which initial seed probably belongs to a big vulnerable cluster that will be completely activated at equilibrium.
In the next section we will simulate Watts' model on Erdös-Renyi graphs, aiming to highlight the double phase transition.

6.2 Simulations of Watts' model

To validate Watts' model we have opted to use Erdös-Renyi graphs because they are the simplest random graphs and we know their degree distribution, at least as

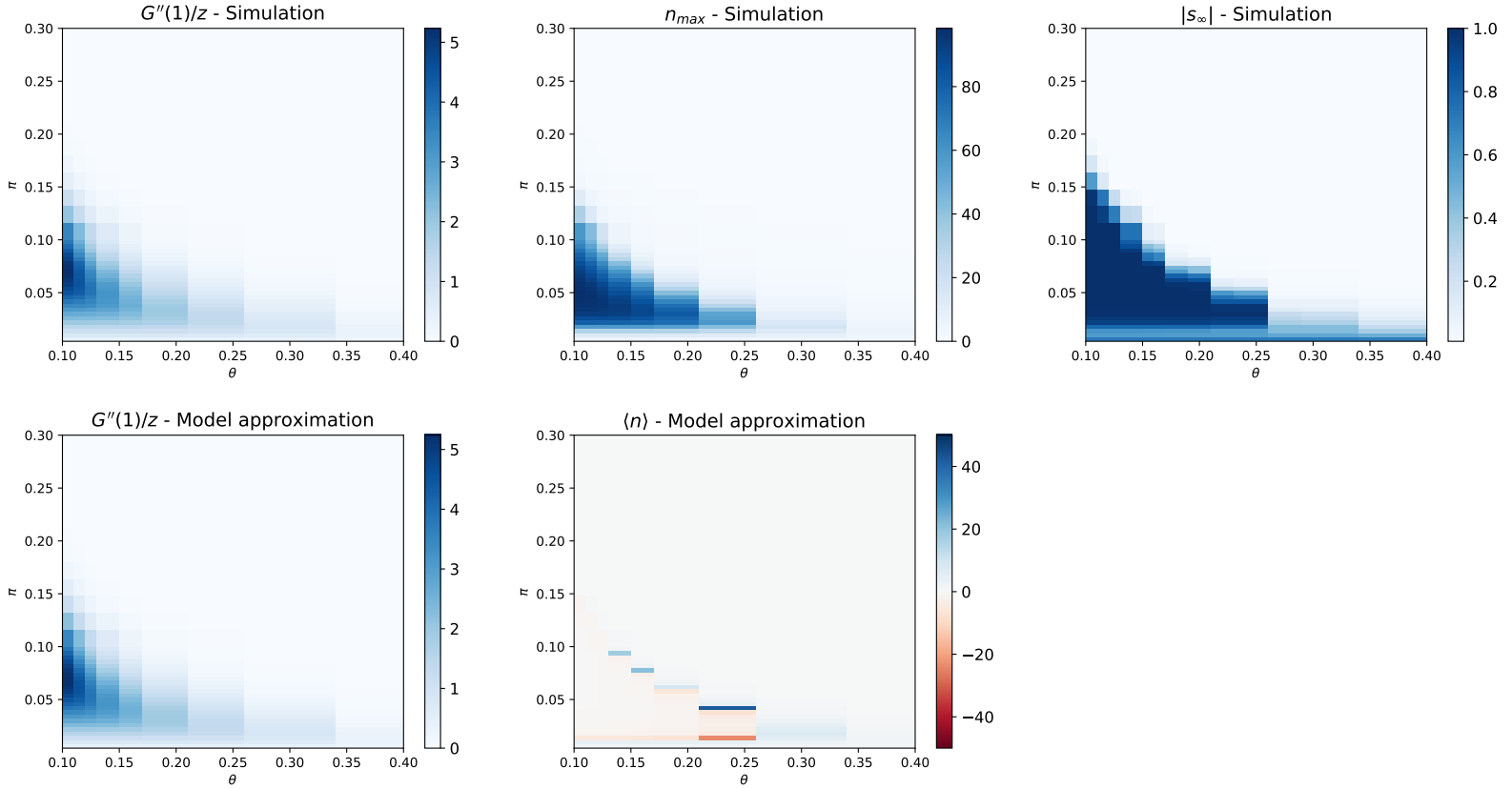

Figure 6.1: The first row has been obtained by simulating the diffusion process on 50 Erdös-Renyi graphs of order 100 for each combination of θ and π choosing as initial seed a vertex in the biggest vulnerable cluster. For θ we have chosen steps of 0.01 and for π we have chosen steps of 0.004 in $(0, 0.1)$ and 0.016 in $[0.1, 0.4]$.
From left to right we have the value of $G''(1/z)$, the dimension of the bigger vulnerable cluster and the number of active vertices at equilibrium, averaged, for each combination of θ and π, on all the 100 graphs.
In the second row we have, on the left, the values of $G''(1/z)$ calculated using the theoretical degree distribution of an Erdös-Renyi graph (van Steen, 2010, section 7.2.1) and, on the right, the values of $\langle n \rangle$ (according to equation 6.7). The values of θ and π are the same of the first row.

the probability to have a certain degree (van Steen, 2010, section 7.2.1).
We have chosen the shape of the graph (i.e. the probability π that a random edge belongs to the graph) and a threshold θ, equal for each node, as variables.
After having fixed the order to 100, we have calculated the ratio $G''(1)/z$, according to equation 6.8, and $\langle n \rangle$, according to equation 6.7, which are displayed in the second row of figure 6.1.
For θ we have used a step of 0.01 along all the range considered, while for π we have chosen a step of 0.004 in $(0, 0.1)$ and of 0.016 for $\pi \geq 0.1$, in order to have sufficient resolution to highlight the lower phase transition.
As benchmark we have simulated 50 Erdös-Renyi graphs for every combination of θ and π.
For each we have measured: $G''(1)/z$ (using the empirical degree distribution), the dimension of the biggest vulnerable cluster and the number of active vertices at equilibrium, having chosen as initial seed a vertex belonging to the biggest vulnerable cluster.
The first row of figure 6.1 represents the average of these measures.
We see that Watts' model on Erdös-Renyi graphs forecast very well the values of $G''(1)/z$ and the critical points for the phase transition.
However, at the same time in the region in which a complete diffusion happens Watts' model forecasts completely wrong values for $\langle n \rangle$ (i.e. a negative value for the average of a set of natural numbers). This could be caused by the violation of the hypothesis about the absence of cycles in the vulnerable clusters, which is unrealistic when a giant vulnerable cluster which includes the majority of vertices (and so the majority of the edges belong to the correspondent inducted subgraph) appears.
Also, the apparent complete diffusion which appears for $\pi \simeq 0$ is likely caused by a numeric artefact, which could be probably removed with a better design of the simulation algorithm.

Chapter 7

Conclusions

In chapter 4 we have obtained the first important result of this thesis: the dynamics of the threshold models of diffusion is not very stable.
Sometimes a little change of the threshold of a single node is sufficient to cause a phase transition.
This behaviour has been confirmed in chapter 5, where we have studied the role of the network's shape.
In particular, we have found that a handful of new edges could increase or reduce significantly the number of the active vertices at the equilibrium.
Finally, in chapter 6 we have taken into account both the threshold distribution and the shape of the graph to find the critical points of the phase transitions (equation 6.8).
The double phase transition induced by the shape of the network looks like the most important feature of threshold models of diffusion, since we have observed it for different couples of variables (even if we probably have observed the same transition under a non trivial change of variables).
Further studies on this topic can review both theoretical and applied aspects.
It is possible to focus on a particular subset of models, for example considering only a family of graphs or a specific threshold distribution, to achieve better approximations for critical points' value or final size of diffusion, either in terms of accuracy or computation time.
It could be interesting to adapt these models on directed graphs and check if the results are similar.
Another possibility is the development of algorithms to check if a network is vulnerable to diffusion. This kind of algorithms would be very useful during the design of real networks (like an electric grid or an intranet).
An applied researcher, instead, could aim to find real situations which can be modelled with a threshold model of diffusion, verify the accuracy of the theoretical predictions and characterize the distributions of degrees and thresholds.

Appendix A

Appendices

A.1 Scripts

The source code of the programs used in this thesis to plot, simulate or calculate is available online at https://github.com/TnTo/threshold-models-of-diffusion.
The *py* folder contains the python scripts.
The *fig* folder contains the figures of this thesis both as pdf and svg.
The *data* folder contains the data obtained and used in sections 5.3 and 6.2. These data are useful to experiment with plot tools or to perform statistical analysis, because the programs which produce the data, in particular the one for Nematzadeh's model, require up to some hours to conclude their execution.
All the code for this thesis has been written using Python programming language, version 3.7 (https://www.python.org/).
Network simulations and analysis have been performed with NetworkX library, version 2.3 (https://networkx.github.io/).
The algorithms use some functions from NumPy library, version 1.16 (http://www.numpy.org/), and SciPy library, version 1.2 (https://www.scipy.org/).
Data manipulation has been carried out with Pandas library, version 0.24.1 (http://pandas.pydata.org/).
Plots have been realized with Matplotlib library, version 3.0.3 (https://matplotlib.org/).

A.2 Additional plots

In the following pages the plots realized in the same way as figure 5.4, but with different thresholds θ, are shown.
Further information are available in section 5.3.
The source code to generate these graphs is available online, as stated in appendix A.1.
These plots are available online in the same Git repository.

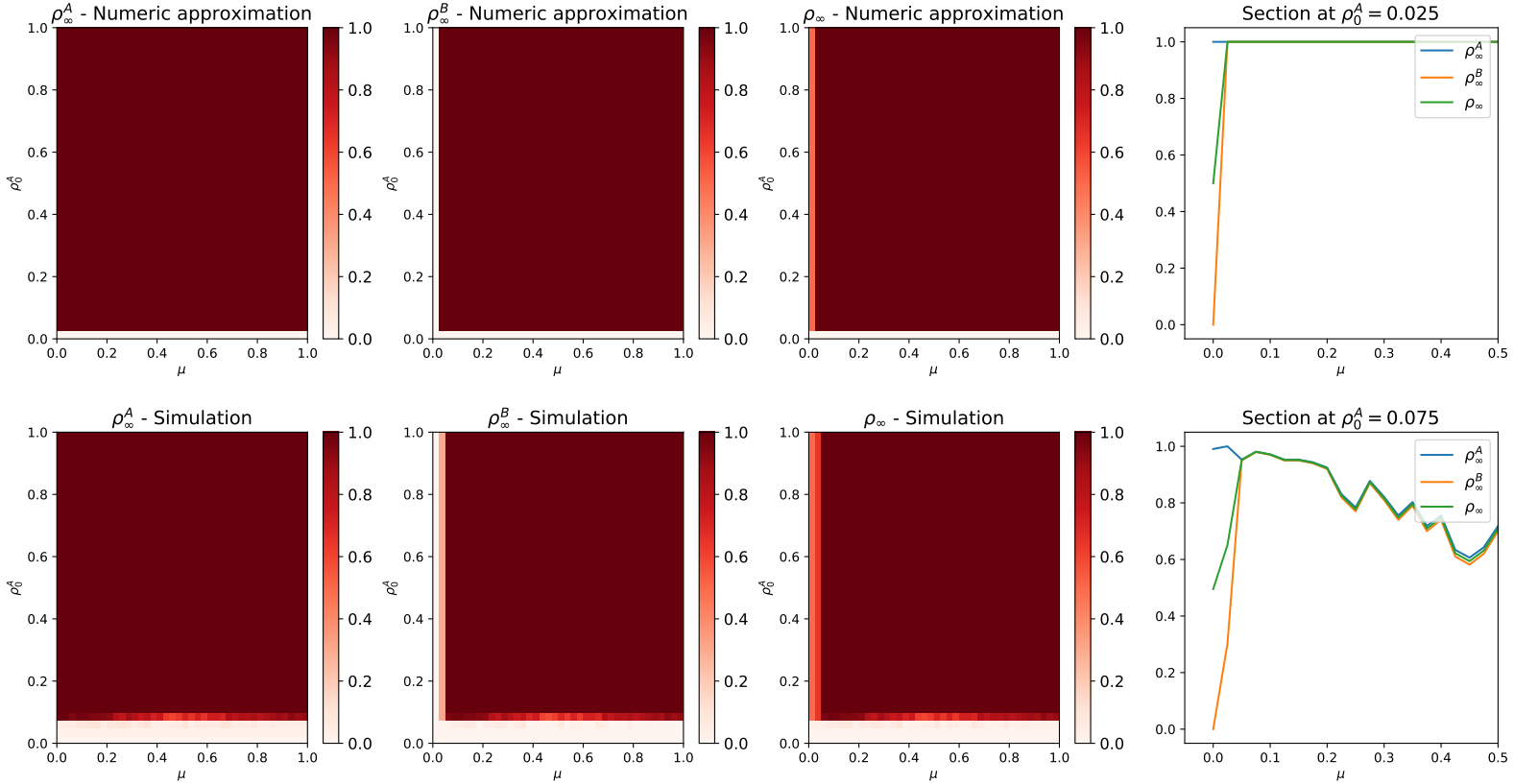

Figure A.1: This is the same plot represented in figure 5.4 but with $\theta = 0.1$.
For all plots we have fixed $p = 100$ and $q = 1500$. μ and ρ_0^A are both incremented by a step of 0.025.
The simulated results are obtained choosing 10 different initial seeds in A on each of 10 different Nematzadeh's graphs.
For further informations see section 5.3.

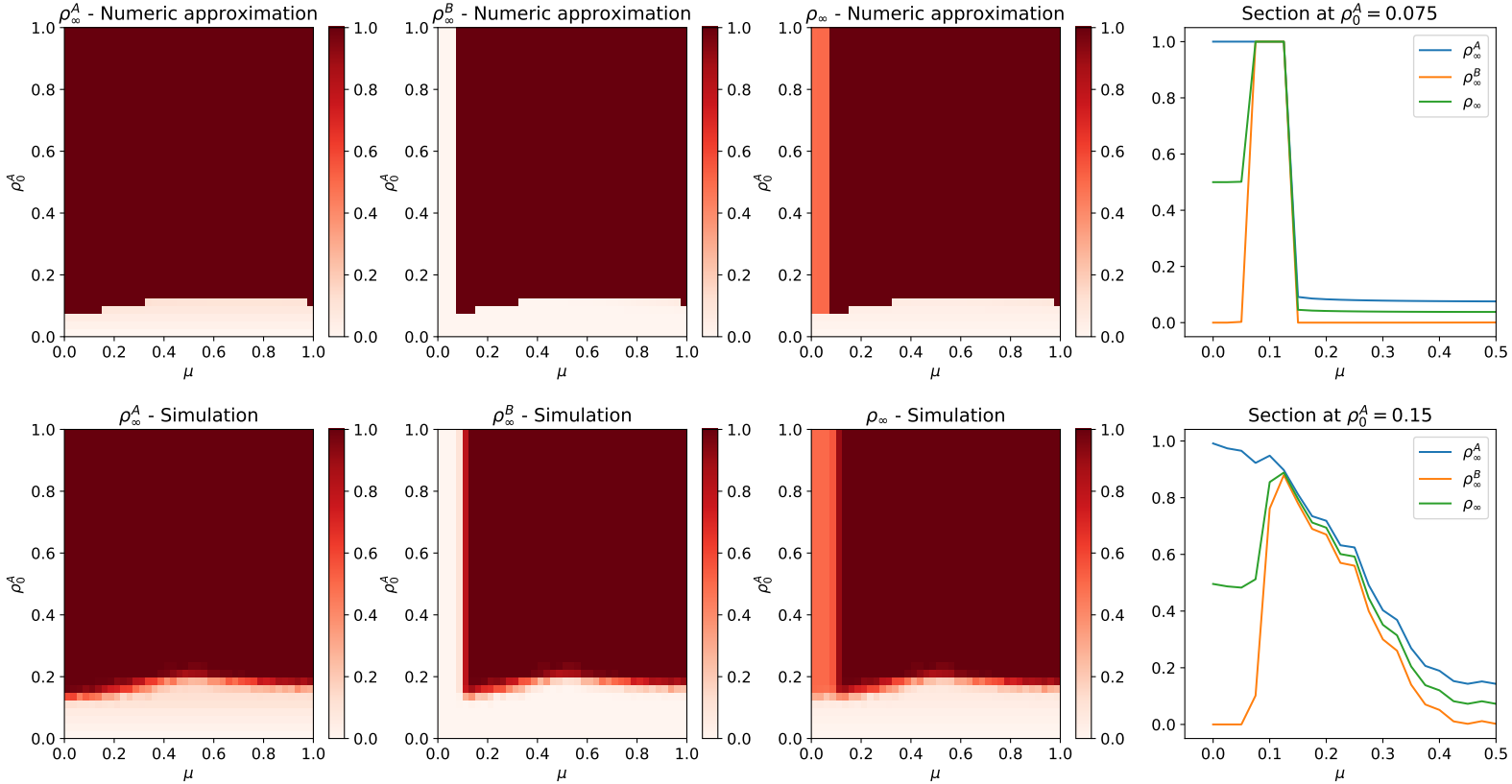

Figure A.2: This is the same plot represented in figure 5.4 but with $\theta = 0.2$.
For all plots we have fixed $p = 100$ and $q = 1500$. μ and ρ_0^A are both incremented by a step of 0.025.
The simulated results are obtained choosing 10 different initial seeds in A on each of 10 different Nematzadeh's graphs.
For further informations see section 5.3.

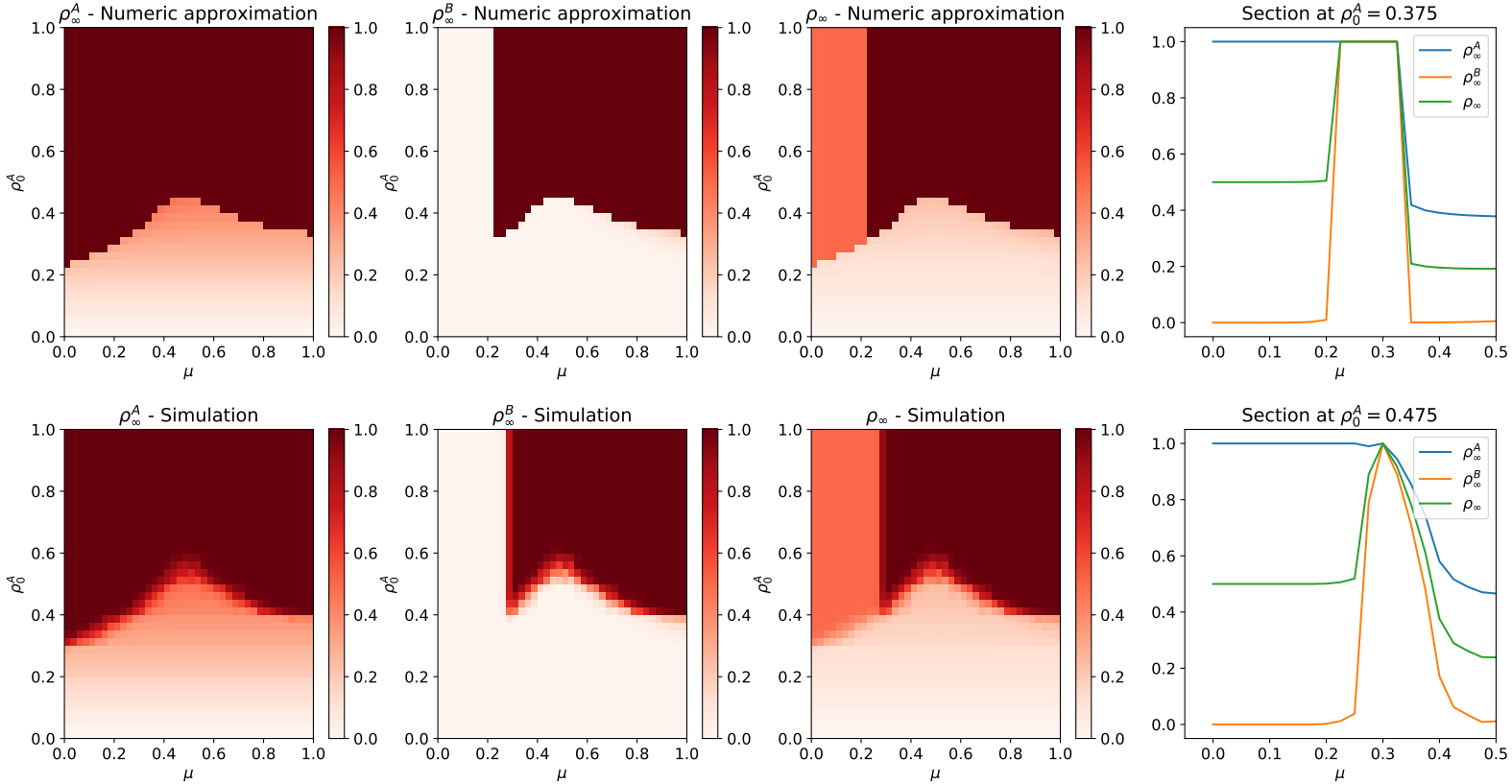

Figure A.3: This is the same plot represented in figure 5.4 but with $\theta = 0.4$.
For all plots we have fixed $p = 100$ and $q = 1500$. μ and ρ_0^A are both incremented by a step of 0.025.
The simulated results are obtained choosing 10 different initial seeds in A on each of 10 different Nematzadeh's graphs.
For further informations see section 5.3.

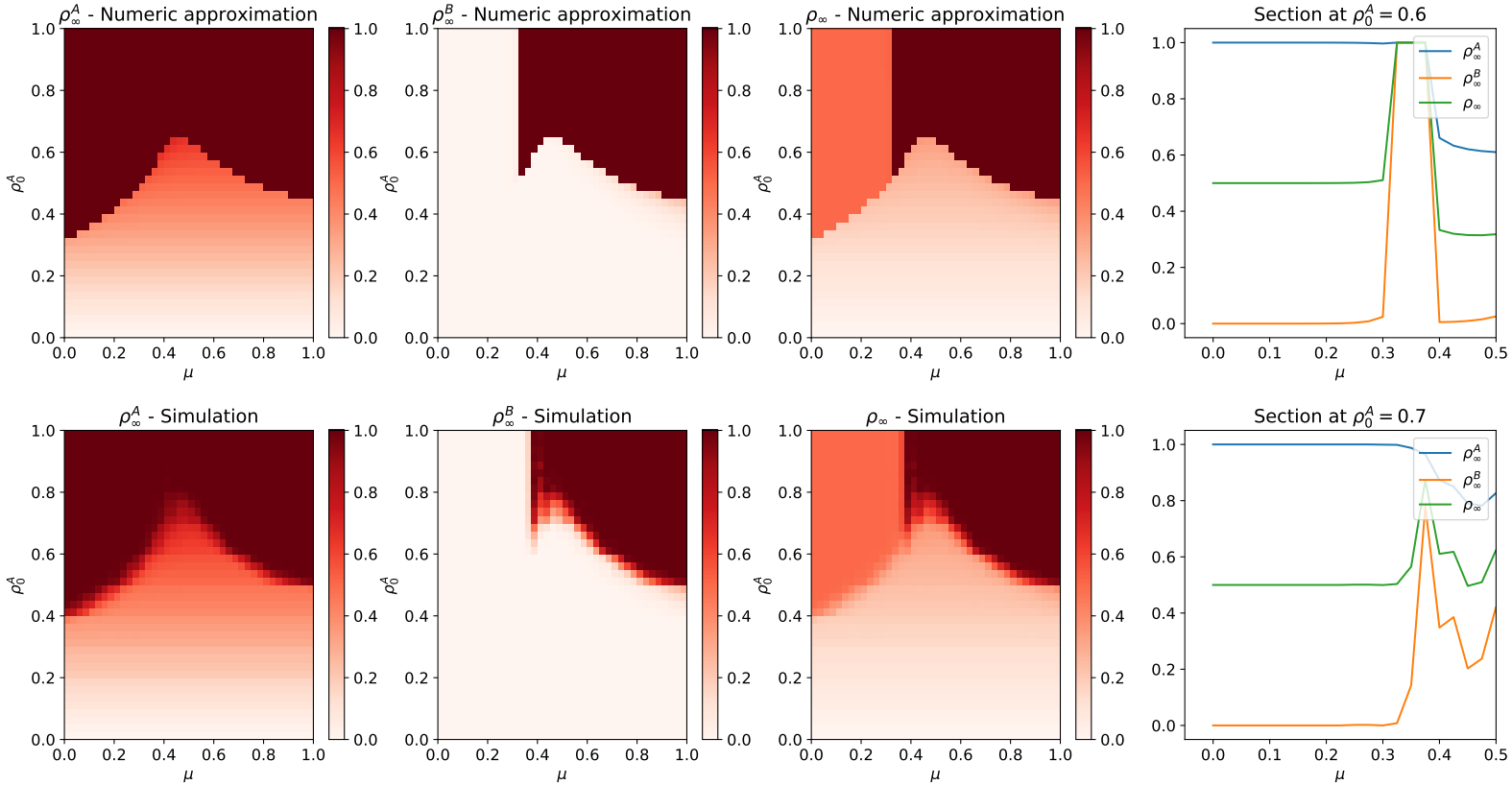

Figure A.4: This is the same plot represented in figure 5.4 but with $\theta = 0.5$.
For all plots we have fixed $p = 100$ and $q = 1500$. μ and ρ_0^A are both incremented by a step of 0.025.
The simulated results are obtained choosing 10 different initial seeds in A on each of 10 different Nematzadeh's graphs.
For further informations see section 5.3.

Bibliography

Threshold models of diffusion

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, *83*(6), 1420–1443. doi:10.1086/226707

Nematzadeh, A., Ferrara, E., Flammini, A., & Ahn, Y.-Y. (2014). Optimal network modularity for information diffusion. *Physical Review Letters 113, 088701*. doi:10.1103/PhysRevLett.113.088701. arXiv: http://arxiv.org/abs/1401.1257v3

Newman, M. E. J., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E 64, 026118*. doi:10.1103/PhysRevE.64.026118. arXiv: http://arxiv.org/abs/cond-mat/0007235v2

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, *99*(9), 5766–5771. doi:10.1073/pnas.082090499

Other resources

Buonocore, A., Di Crescenzo, A., & Ricciardi, L. (2011). *Appunti di probabilità*. I manuali. Liguori.

Ross, S. M. (2006). *Introduction to probability models*. Academic Press Inc.

van Steen, M. (2010). *Graph theory and complex networks: An introduction*. Maarten van Steen.

Contents