Mathematical Models and Simulations of Complex Social Systems

by

María del Rocío Gómez Bardón

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
Faculty of Science
The University of Birmingham
September 2009

Abstract

In this thesis we present two different models of Complex Social Systems. The first model represents a vector-borne disease that takes place in a heterogeneous environment composed of areas of different types. Two populations take part in the epidemic process: humans and vectors. The population of humans moves around the heterogeneous environment. The idea of this model is to understand how the movement of people in the heterogeneous environment can affect the dynamics of the disease. The second model represents a Susceptible-Infected-Susceptible process on a social network. The population is represented as nodes, and the edges represent the possible transmissions between two people. We investigate how different topologies in the network affect the spread of the disease in the system. We simulate both models, and we perform a mathematical analysis of both of them. For the mathematical analysis we use an adapted version of the Random Heuristic Search framework, which was originally used for the understanding of Genetic Algorithms. In this thesis we investigate the predictability power of the mathematical approach.

To Ana Luisa

Acknowledgements

Firstly, I would like to thank my supervisor, Jon Rowe, for his support and guidance over the last seven years; the first two years when I worked with him as Research Associate, and the the last five years as his PhD student. I really appreciated his help, his patience, and his inspiring comments.

I would like to thank the School of Computer Science for funding this PhD for three years and making it possible for me to work on this thesis.

I also would like to thank the members of my thesis group, Russell Beale and Steve Vickers, who always made the time to read my reports and make insightful comments about this research.

I am most grateful to Ian, for his love and his patience, for his unconditional support during these years, and for his infinite faith in me.

And last, I would like to thank my parents who simply put me where I am.

Contents

List of Figures

LIST OF FIGURES

Part I

Background

Chapter 1

Introduction

1.1 Modelling Social Systems

For many years there has been a great interest in trying to understand what are the rules that govern traffic, market economies, communication networks, distribution of populations in cities, spread of diseases, and many other dynamic social systems. Social dynamics are, in general, of great interest but not easy to understand; social systems form part of the group of Complex Systems.

We can define a Complex System to be a system formed by a group of particles that are interdependent on each other. By interdependent we mean that, in order to describe the system as a whole, we need to describe all its parts and each part must be described in relation to the other parts (13). Formally a Complex System has been defined as *"a system where the collective behavior of their parts entails emergence of properties that can hardly, if not at all, be inferred from properties of the parts"*, (88).

This definition may look very twisted and complicated but, in fact, most of the systems that we find in the real world fall under the category of Complex

Systems. Examples of Complex Systems can be found in almost any studied field such as economics (18; 97), ecology (96; 97), biology (29; 45), anthropology (8; 12), sociology (78; 79), computer sciences (47; 102), physics (11; 108), etc. The fascinating thing about these systems is that, regardless of the complexity of the interactions between the parts, or the complexity of the parts themselves, the emergent effects of that local interaction at global scale are remarkably simple. A question that many researchers working in different fields that deal with Complex Systems, including social systems, are concerned about is how local interaction between individual parts gives rise to different emergent behaviours in the system as a whole.

Modelling and simulating these systems can provide us with a deeper understanding of how they behave in real life. The process of building and analysing models has also helped the scientific community to realise that there are many systems that have common emergent behaviours at a larger scale, even when they appear to be very different in detail (11; 13; 25). This is called universality of emergent behaviour. These universal properties allow us to find commonalities between diverse disciplines. We can use different modelling techniques depending on the characteristics of the systems we want to study. The first method used to model dynamical systems was differential equations; this method assumes uniformity in the system and smoothness in its motion and, therefore, it has limits in its application to the study of Complex Systems (54).

A different perspective in modelling systems is to describe the system as a group of particles that interact with each other. The process that defines the dynamics of the system under study is understood as a process due to the local interaction between the particles. Due to these local interactions, there are some properties that emerge in the system as a whole. This seems a better approach to capture the characteristics of a Complex System. Individual based modelling was successfully used by the physicist Johann Diderik van der Waals in the late 19th century for the study of kinetic theory of gases (11). However, it has not been popular in social sciences until recently. One of the first individual based models in social sciences are the well known Schelling's models (78; 79). He modelled a

neighborhood as a two dimensional lattice and people from the neighbourhood as coloured dots occupying spaces in the lattice. The dots could be of two different colours, each colour representing a different ethnicity. The dots moved around the lattice following certain rules based only on the principle that an individual does not want to live in a minority. Although Schelling did this individual based dynamical models by hand, after moving around the dots in the lattice a few times, he came to the conclusion that the dots of the model would be clearly segregated in groups of different colours. It was then clear that individual based models could give a very good insight of the basic principles of social behaviour and a better understanding of the relation between local interaction and global behavior in the system (38).

Individual based modelling could be very cumbersome if it should be done by hand, as was the case in Schelling's models; he could not take advantage of any other technology at the time. Nowadays the conceptual models and their dynamics can be very easily recreated with the help of computer technology. A good example of the use of computer technology for individual based modelling is **Agent Based Simulation**. Agent Based Simulations can be understood as a collection of computational elements that interact with each other and with some environment. These computational elements are called agents. When we create a multi-agent simulation we implement specific behaviours to specific individuals or agents. When the individuals interact with each other, a new general behaviour emerges in the system. This property of multi-agent systems makes them especially suitable for the simulation of social systems. Agent Based Social Simulation has been defined as the intersection of three fields: Agent Based Computing, Computer Simulation and Social Sciences (23). Yet, I would consider agent-based computing part of the Computer Simulation field and also part of the Individual Based modelling approach. In my opinion, Agent Based Social Simulation would be more accurately defined as the intersection of three fields: Social Sciences, Individual Based modelling and Computer Simulation.

1.2 Mathematics or Simulation?

Mathematical analysis has great advantages as, for example, being able to formulate very precisely the properties of the system and have rigourous proofs about these properties. A mathematician wants to understand formal structures by formal means. Mathematics can help to make an accurate prediction of the emergent behaviour of the dynamics by using a rigourous analysis of the system, and it has the capabilities to make a thorough analysis of relation between different parameters in the model without having to test one by one.

One of the major difficulties with this approach is the rapid growth of the mathematical complexity of the systems used to describe the various aspects of phenomena in sufficient detail. Even when the system has been successfully stated mathematically, we can encounter problems in solving them in an analytical form. Hence, their practical use in specific cases is very limited. In order to avoid the overload of mathematical complexity, the systems tend to be very simple and abstract. This seems to be a common reason for questioning the usefulness of these models, (9; 21). Another great difficulty is the necessity of knowing the language of mathematics itself. Often researchers that work in areas that can directly benefit from the use of mathematical models, as is the case in social sciences, do not have a broad knowledge of mathematical language and techniques.

There are also great advantages in the use of agent based simulations. Simulations can capture emergent phenomena while the modeler only focuses on the description of the individual and its relations with other individuals and its environment. It provides a natural environment for the study of social systems, and it is a flexible technique, specially when compared with mathematical modelling, particularly in relation to the development of spatial models. There is, of course, a need to create an abstraction and simplification of the system, but there is still plenty of room for adding a great number of parameters and heterogeneity in the actors of the model and their interaction. Even the individuals themselves

can be very complex agents (40). Simulations can then be used empirically as a case-scenario of situations that we could not mimic in the real system.

While computer simulations can be very useful, they also have some downfalls. The possibility to increase the complexity of the model may be very tempting for the modeler as reality is associated with complexity. In any case, adding complexity to a simulation does not imply making it more realistic. In fact, there is a great chance to be adding more artifacts in the model which are very difficult to identify if the conceptual model has a high number of parameters or if the process of the dynamics get too complex. Analysing the results of the system can be very difficult as it is necessary to test parameters one by one in order to have a clear understanding of their effects in the system, and, even if a thorough test is done, results obtained through simulations do not formally validate the observed behavior (50).

1.3 What Is This Thesis About?

In this thesis we present two different models of Complex Social Systems. They are both related to epidemiological processes or contact processes in a population. In the first model we keep the topology of the environment very simple and we focus more on the complexity of the contact process by modelling two populations that take part in the dynamics of the model. We also add movement of one of the populations around the environment. The second model keeps the contact process in the population very simple, but the dynamics take place on social structures that are medelled as networks, and that can be very complex.

The first model is a model of the spread of a vector-borne disease in a country (taking Malaria as an example), in which we investigate how the heterogeneity of the environment and the movement of people between different areas of the environment affects the dynamics of the disease. We adopt a meta-population approach by modelling the environment as a set of patches which represent different

areas of a country. The patches are connected to each other forming a very simple grid. We modelled two different populations, one representing human and one representing vectors. The disease spreads among the populations by transmitting the infection from people to vectors and from vectors to people. There are three different types of patches, each one representing one of the three geographical areas *Swamp*, *Rural* area or *City*. Each type of patch is characterised by its area size w, its number of mosquitoes V, and the probability for a person to leave the patch l. The mosquitoes in the model moved within each patch but do not move from patch to patch. Each time step the people in the model move from patch to patch. Once the population has moved, the dynamics of the disease take place in each patch independently of each other. This model has been implemented in five steps creating five different sub-models with increasing complexity by adding different factors to the original model in each step. Each sub-model has been simulated in a computer and analysed mathematically.

In the second part we study a model normally known as SIS model. SIS stands for susceptible-infected-susceptible and it is a well known epidemiology model (104). The model represents a contact process among individuals of a populations. Each individual of the population can be either susceptible or infected. Susceptible individuals that are in contact with infected individuals can get the infection transmitted with certain probability p. Infected individuals get spontaneously back to the susceptible state with certain probability b. We represent the population as a network in which the nodes represent the individuals of the population and the edges represent the contacts through which the process of the dynamics take place, so two nodes connected through an edges represent two individuals in contact with each other. Different network structures would represent different contact networks in the population. The idea of this model is to understand how the topology of the contact network of a population can affect the dynamics of the studied process. We present three models of this dynamical process taking place in different network structures. We also present a mathematical analysis for the dynamics of processes in these networks studying mainly the fixed points of the system.

The SIS model is commonly used to understand the spread of a disease in a population (such as a common cold), but it can also represent other contact processes that take place in a population. Other type of dynamic social systems that can be modelled by the SIS contact process are: rumor spread in a community, computer virus in a computer network, extinction and colonisation of species in islands, or voter behaviour in a population.

Although the work for this thesis was not specially focused in any particular social phenomenon, it seems all models presented in this work are highly related to epidemic models so, in many cases, we will explain these models, the outcome of the analysis, and the result of the simulations with terms used in epidemiology.

1.3.1 General Methodology and Approach

In order to understand the social systems mentioned above, we intend to create a model that captures a minimum number of parameters that may have an impact on the system dynamics. We are aware that oversimplification may lead to understanding a toy model but not the system under study. However, we still advocate for simplicity and clarity in our models as we want to understand general principles of social systems; describing the actors in too much detail may sometimes restrict the model to very particular cases. The second reason for simplicity and clarity comes from the desire to describe the model mathematically. In fact, our main objective is to create a mathematical description of the macro-behaviour of the system based on the local interactions of the actors. If we do so, we could formulate very precisely the properties of the system and have rigourous proofs about these properties. It could also help to make an accurate prediction of the emergent behaviour of the dynamics. One of the main questions of this thesis is to understand how accurate our mathematical analysis is and what information we can infer from the mathematical analysis about the dynamics of the model.

The general idea of the approach we use is, firstly, **identifying clearly** what

are the factors that we need, or want, to understand. We will always try to model them in a simple way to reflect only their more relevant characteristics. We then create several models of a system, starting from the most simple conceptual model and carry out an **incremental implementation** of these factors or parameters in order to analyse and understand them individually. For each conceptual model we create **a simulation and a mathematical model**. We carry out experiments in the simulation and a mathematical analysis.

The computer implementation of the multi-agent system will help us to contrast the validity of our system with real systems, to understand what this mathematical description should look like, and to validate the mathematical model and the conclusions drawn from its analysis. The mathematical analysis will help us to have a rigourous representation of the dynamics of the system allowing us to have a formal analysis of the relation between the parameters. The equations of motion of these systems have been calculated and contrasted against the simulation with several experiments. These mathematical models can also be used as a tool to validate the simulations.

The incremental implementation of the factors that will bring complexity to the simulation will help us to understand how each factor contributes in the behaviour of the population in the final simulation and to create different mathematical models with incremental complexity.

The models have not been compared with data of real epidemics. The idea in this thesis is not to present models that mimic or predict how real epidemics work. The focus of this thesis is more centered on comparing results between simulations and mathematical results at "macro" level so the conceptual models have been kept very simple. Both models are too simple to yield useful quantitative results from the simulation or from the mathematical models. We are aware that transmission of infectious diseases depends on more than a few parameters. Nevertheless, we still believe that models like ours are useful in getting a qualitative understanding of the basic transmission mechanisms. We believe that a lot

must be understood about these basic transmission mechanisms before one can confidently look at models with more parameters.

The simulations are time driven with discrete time steps. Populations are represented in a discrete manner (as it is individual based modelling), and the environment is also discrete. The simulations are built with RePast[1], (87). The reasons for this choice are, firstly, that it is written in Java language, which is the programming language I can program with best, and secondly, because it is one of the best available free libraries for support of social agent simulation in terms of user support, capabilities, usage of memory vs speed, restrictions in the agent modelling, documentation, graphical interface etc... (95).

Other systems that hold similar properties to the social systems we want to model are populations of genetic algorithms. Genetic algorithms can be understood as a kind of multi-agent system in the following way: there is a population of individuals, each represented by its "DNA", and we can consider this DNA as the state of the agent or individual. There is a finite number of possibilities in the set of states of the agent, Ω (usually called the *search space*). The individuals interact with each other by crossing over, and they interact with the environment when its fitness is evaluated (the ones with more fitness are more likely to survive). An important question that can arise when studying the population of genetic algorithms is: under what conditions can the dynamics of the population be inferred? This is the same question we are interested in when we analyze the behavior of multi-agent systems.

Michael Vose developed a mathematical framework for the study of the populations in Genetic Algorithms, (76; 98). This mathematical framework is known as the Random Heuristic Search. The idea of this theory is to consider a population distributed over a set of possible states (DNA) and define a map G that takes distributions of population to distributions of population, $G : \Lambda \to \Lambda$, where Λ

[1]REcursive Porous Agent Simulation Toolkit. Free and open-source, agent-based modelling and simulation toolkit.

is the set of distributions over Ω. G is called *The Heuristic Function*; the terminology comes from the development of GA theory in which genetic operators are used to stochastically search some problem space. The search proceeds by taking a population distribution p and applying the heuristic function to form $G(p)$. This distribution is sampled a number of times to form a new population. The process forms a Markov chain whose transition matrix can be written in terms of the function G.

In the work presented in this thesis I analyse multi-agent systems using this mathematical framework. In order to adapt this framework to multi-agent systems, we have to think of Ω (*search space*) as the set of possible states that the agents of the simulation can adopt. The problem is now to write down the heuristic function G of the system. This is equivalent to finding the equations of motion for the system, i.e. how the population distribution changes at each time step. In the case of multi agent systems we know the rules of the agents in advance, and the equations of motion represent a mathematical abstraction to be deduced from those rules. This will be the first stage of applying the RHS framework to the analysis of multi-agent system.

Once the map G has been determined we investigate analytically different properties of the system like, for example, its fixed points. The reason to study the fixed points of a system is the following: In an epidemic model of finite population (as it is in our simulations) there is always a probability for the infection to disappear completely, regardless of the parameters of the model. It may happen that all infected individuals become susceptible at some time step and then the epidemic disappears. We say that the probability for an epidemic to disappear is 1, but we don't know how long it would take, and it can be a very long time. In the meantime, while the epidemic has not disappeared, the epidemic seems to be in *equilibrium*, with the size of the epidemic oscillating around some value which we call *attractor*. These oscillations not only appear in simulations but also in real epidemiological systems. This apparent equilibrium in the simulations is only temporary and represents a case of "*metastability*", where a temporary equilibrium persists for a long time until a transition occurs to a final equilibrium,

which in these models would be extinction of the disease. In order to study analytically this temporary equilibrium observed in the simulations, we assume infinite population in the model. The *fixed point* of the mathematical model of the system, which represents an equilibrium state in an infinite population, would correspond to the attractors in a finite population system. The fixed points can also be referred to as *equilibrium point* or *steady-state* of the system.

1.4 Structure of This Thesis

In Chapter 2 we review different techniques used for the study of Social Systems. We divide our review in four sections. The first sections, 2.1, reviews models with a very simple environment where complex dynamics take place. Section 2.2 reviews Networks Theory, where there is a greater focus on the complexity of the topology of the network where the dynamics takes place, and the interaction dynamics tend to be very simple. The third section, 2.3, gives some examples of models built for experimental proposes, and the last section, 2.4, introduces a concept of Effective Reproductive Rate, which we will use in our second model.

In Chapter 3 we present a model of vector-borne disease in which two populations take part, mosquitoes and people, and the disease spreads by transmissions between the two populations. The idea of this model is to understand how the movement of a population in a heterogeneous environment can affect the dynamics of the disease. In Section 3.2 we present the motivation for this model. In Section 3.4 we present different sub-models implemented with increasing complexity and a mathematical analysis of each of them, focusing mainly in the study of the fixed points. During the mathematical analysis of some of the models we make use of *The Lemma of the Fixed Points* stated and demonstrated in Appendix A. This lemma is used for understanding the behavior of the system in terms of its fixed points. In Chapter 4 we present a set of experiments and run in the simulated sub-models. We aim to compare these results with the estimations made by the mathematical models.

In Chapter 5 we present an implementation of the SIS dynamics on different networks. Different network topologies represent different contact networks in the population. The idea of this model is to understand how the topology of the contact network of a population can affect the dynamics of the studied process. We present three models of the SIS dynamics taking place in three different network topologies. We also present a mathematical analysis for the dynamics of process in each of the networks, studying mainly the fixed points of the system. In section 5.7 we introduce the concept of Effective Reproductive Rate in a population and present a mathematical analysis of our models investigating this parameter. In Chapter 6 we present the relevant experiments testing the accuracy and predictability of the mathematical models.

In Chapter 7 we conclude the thesis summarising what we have learned about the different models and methodologies. In addition we summarise several future studies that are suggested by these investigations.

Chapter 2

Modelling Social Systems

In the previous chapter we introduced the topic of studying Social Systems using simulation and mathematical modelling. In this chapter we introduce in more depth different techniques used for understanding Complex Social Systems and give some examples. We divide this review in four sections. In the first section we review models in which the environment of the system is represented in a very simple way and the models focus on the complexity of the dynamical process of the system. In the second section we review models of Complex Networks in which the dynamical process is very simple and there is a greater focus on the topology of the network where the dynamics takes place. In the third section we give some examples of models used for experimental purposes, which tend to model the systems with high complexity in both the topology of the environment, and the dynamical process of the system. In the last section of this chapter we introduce the concept of Effective Reproductive Rate in epidemiological models which will be further discussed and analysed in the second part of this thesis.

2.1 Models with Landscape with Homogeneous Topology

There is a number of different techniques that are used to study the dynamics of systems which assume the homogeneity in the topology of their environment. This assumption in the environment makes the system easier to analyse, but it may also oversimplify the system giving results that do not reflect behaviour of real systems (21). Mathematical tools used by this approach are *Spatial Moment Equations*, *Partial Differential Equations*, or *Lattice-based* models. Spatial Moment Equations and Partial Differential Equations normally assume perfect mixing in the interaction of the population in the model, and there is not a spacial representation of the environment as such. Lattice models represent a continuous landscape as a regular square lattice. Each lattice may contain a single individual or a population. Lattices change their state deterministically or stochastically according to the rules. Examples of lattice models may be *Interacting Particle Systems*, *Spatial Stochastic Models*, *Cellular Automata*, or *Coupled Maps Lattices*. The models studied when using these tools are usually designed to investigate very general principles of the systems rather than tactical methods designed to answer specific questions. Below we give some examples of models and analysis using these techniques.

2.1.1 Cellular Automata

Hegselmann and Flache discuss various applications of Cellular Automata in the social sciences in paper (38). The paper seeks to demonstrate that Cellular Automata, although are models based on rigourous simplification, are also a good method for the understanding of social systems. The characteristic of Cellular Automata models are the following:

- Cells are arranged in a regular grid (normally 2-dimensional)

- Every cell adopts one state of a finite set of states
- Time is discrete
- Cells change their state according to the local rule
- The same transition rule applies to all cells
- In each period of time cells are updated simultaneously or sequentially.

In paper (38) Hegselmann and Flache present a check-board model (or 2-dimensional lattice model) in which cells are occupied by individuals with opinions. Each individual's opinion will be modified depending on its neighbours' opinions. It presents cases in which opinion has a discrete value (for example individuals can have one of 5 different opinions), or it can be a continuous value (individuals will have an opinion represented by a value between 0 and 1). The experiments suggest that, the more opinions there are in the set of states, the closer the stable state gets to general consensus and the smoother are the transitions between areas of different opinion. The results of the simulation also show that discretisation of the opinion value has a great effect in the result of the model, so this is an example of how discretisation of space, time, and state can matter on the qualitative behaviour of the model; in conclusion, the modelling of the system has to be carried out cautiously as it may produce artifacts. The author argues that the simplicity of these models allows us to 1.-focus on the dynamics of elementary social interaction when studying complex social systems, and 2.-make it very clear how certain macro effects are dynamic consequences of decisions and mechanisms operating only at micro level.

Another example of a Lattice Model is the work presented by Boccara and Cheong in their paper (19). They present a SIS model implemented in a Lattice network. In this Lattice not all nodes are occupied by agents, only a fraction of them, and the rest are empty. The agents move to a new empty node with some probability m. This empty node is chosen at random and it depends on a previously chosen range. To illustrate the importance of this range they considered

two extreme cases. In one case, the called "short range", the empty node will be one of the four neighbours. The "long range" will be any vertex in the lattice. The rules for the disease transmission between agents are exactly the same as the ones we use here. The size of the lattice is fixed, and the number of agents in the lattice is also fixed. Having fixed those values, the number of agents infected in the network will depend on the probability to transmit the infection p_i, the probability to recover p_r, and the average number of movements per agent m. Let us consider m a fixed value too. Then the number of infected agents depends on the pair of probabilities (p_i, p_r). If we now fix one of these probabilities, let's say p_i, the number of infected agents in the lattice will be 0 in some cases (if p_r is too high), or a positive value (if p_r is not high enough). If we find a value p_r for which the number of infected agents changes from 0 to a positive number we say that (p_i, p_r) is a transition point. For each p_i we will find a value for p_r such that (p_i, p_r) is a transition point. If we run different simulations of the model and plot the transition diagram of pairs (p_i, p_r), we can see that the diagrams change for different values of m.

The mean-field approximation for this model says that, when time $t \to \infty$, the number of infected I in the lattice is given by equation:

$$I = (1 - p_r)I + (C - I)(1 - (1 - I)^z) \tag{2.1}$$

where C is the constant density of agents in the model, and z is the number of neighbouring vertices of a given vertex. From equation 2.1 they deduce that $I = 0$ is always a solution of the system that represents the *disease free* state of the model, and it is stable if, and only if, $zCp_i - p_r \leq 0$. Otherwise, if $zCp_i - p_r > 0$, the system is in an *endemic state* and the proportion of infected is given by the only positive solution of equation 2.1. The results of the mean-field analysis are compared to the transition diagrams obtained by running the simulations and it is concluded that, when they simulate a large range movement (i.e. for high values of m), the mean-field model is a good approximation but, in general, it cannot predict correctly the critical behaviour of short-range interaction systems. In this paper it is also analysed how the introduction of the parameter m, which reflects movement, has an impact on the speed of the spread of the disease when

they compare the effect of the simulation results for two different values of m. If these values are high enough, the difference between the two results is small. However, for small values of m, although the disease still spreads slowly, the difference between the results of the simulation says that the number of infected agents decreases dramatically as the value of m approaches 0. This reflects clearly the impact of movement in the model that cannot be captured by the mean-field approximation.

2.1.2 Interacting Particle Systems

Interacting Particle Systems is one example of stochastic lattice models that run in continuous time. In these systems lattices interact with neighbour lattices with some probability and change their states according to the results of those interactions (depending on the rules). Mathematical models used to analyse Interacting Particle Systems are *Pair Approximation Method*, which focuses on joint occupancy probability of neighbour pairs, or *Mean-Field Approximation Method*. In this section we give some examples of studies that have used some of the techniques mentioned above.

Schinazi introduces an Interacting Particle System to model the emergence of drug-resistant diseases in his paper (82). The main result of the analysis is that, if there is a mutation from a natural strain into a drug resistant strain, and if there is an effective drug against the natural strain, there will be an outbreak of the drug resistant strain and it will, eventually, replace the natural strain. This is observed by analysing the Interacting Particle System model. The mean-field model of the system, which is also presented in this paper, only tells us that under certain conditions a drug resistant strain and a natural strain cannot coexist.

2.1.3 Spatial Stochastic Models

Schinazi also published two different papers studying epidemiological problems of tuberculosis using Spatial Stochastic models. The first paper, (80), presents a model for the spread of tuberculosis and HIV focusing on the influence of social clusters in the spread of the epidemic. He divides the population into clusters and analyses when an epidemic is possible depending on two parameters: the infection rate within a social cluster and the size of the social cluster. One of the results of this paper is that, for low casual infection rate, the epidemic is possible only if the cluster size is large enough. If the cluster size is too large, the epidemic may occur only if the within cluster infection rate is large enough. The model suggests that, given a low casual infection rate, the cluster size and the within cluster infection rate are determinant. In his second publication, (81), he focuses on a different aspect of the tuberculosis disease. With tuberculosis, once you have been transmitted the infection, you become a *dormant* infected so you don't transmit the infection or become ill. You become actively infected later and it is not clear the process that makes you get to that state. It is still discussed whether it is by a re-infection of the disease (exogenously) or just an internal process of the disease (endogenously). In this paper Schinazi creates two different models of the system; a Spatial Stochastic model analysed mathematically with a technique called *coupling*, and a mean-field approximation model of the system. He compares both models getting to the conclusion that they give different results because they may behave in strictly different manners. He shows that the mean-field model does not allow an epidemic, even if the first infection rate is high enough, unless endogenous infections can occur. However, in the spatial stochastic model it tells us that, even when there is no endogenous infection, if the rate for a second re-infection is high enough, there is a chance of epidemic in the system. He argues that the different results given by analysing the two models may come from the fact that in the mean-field model, even if the second re-infection rate is high enough, the probability to get in contact with an infected person is low because you get mixed with the non-infected population, whereas in the spatial stochastic model, once you are infected, you are more likely to get

re-infected since your neighbourhood is likely to contain more infected people than the average.

Schinazi also uses spatial stochastic models for the understanding of the role of spatial aggregation in extinction of species (83; 84). He creates and compares two models of spatial aggregation of a species. They both are represent the environment as a two dimensional lattice. Each node of the lattice may host a flock of up to A individuals. Each individual in the flock may give birth at the same node at rate φ, until the maximum of A individuals has been reached at that node. Once the site reaches the maximum number of individuals that it can support, each individual in this node starts giving birth on each of the 4 neighbouring nodes at rate ψ. This rule mimics the fact that individuals like to stay in a flock, as long as they can. The two models have different rules for death of individuals in the model. The first model, (83), mimics a disaster (such as an encounter with greedy hunters, or a disease) in which the a whole flock disappears at rate 1. The analysis of this first model shows that there is a threshold in the flock size of the species so that, if the maximum flock size is above a certain threshold, the population is going to get extinct with probability 1 but, if the maximum flock size is below the threshold, there is a strictly positive probability that the population will survive. In the second model, (84), individuals die one by one at rate 1. This models the population in the absence of an external disaster. The analysis for the second model says that, if the maximum flock size A is large enough, the survival of the species is possible for any birth rates φ or ψ but, if $A = 1$ and the external birth rate ψ is too low, then the species dies out. Comparing these two models, Schinazi gets to the conclusion that animals living in large flocks are more susceptible to mass extinctions than animals living in small flocks. However, in the absence of disasters that may cause death of large flocks we have that, under some circumstances such as low birth rate, spatial aggregation makes the species survive.

2.1.4 Metapopulation Models

The word *metapopulation* is understood as a group of populations that live in a fragmented space. All these populations are located in different areas of the fragmented space and have their own independent dynamics but they are connected via migration (106). The idea of modelling populations in this way was introduced by R. Levins with the intention of understanding the processes of extinction and re-population of certain species in islands. His original model assumes that the space is discrete and it is formed by a set of patches which have all the same area and qualities. The dynamics of populations in different patches are independent from each other; the individuals will migrate from patch to patch, but the migration rate is low enough so the movement of the individuals does not affect the dynamics of the populations, yet, it creates a balance between extinction and colonisation (52).

In the metapopulation approach there is a more abstract representation of landscapes than in other approaches used in ecologist studies. In this approach the environment consists of discrete patches of suitable habitat ignoring the shape of the patches surrounded by the landscape matrix through which individuals may migrate. There is a stronger background in mathematical theory and, thus, there is a tendency to approach the problem analytically rather than numerically (34; 36). The success of metapopulation theory comes from the importance of the study of population in fragmented spaces and the simplicity of the theory. However, although it is constantly argued that real systems do not show a metapopulation structure and, therefore, their behaviour cannot be predicted using this methodology (10). There are still some authors, (28), that suggests that there is an issue in the scale of the population dynamics that can be studied by metapopulation approaches because, sometimes, local dynamics of species cannot be predicted by metapopulation models but regional dynamics can. Metapopulation theory was originally developed for the study of terrestrial ecosystems although it has subsequently been applied to other fields such us marine ecosystems, (89), or the spread of Pandemic Influenza around the world, (22).

As part of metapopulation theory we can mention the *Spatially Realistic Metapopulation Theory* (SMT) (35), which uses mathematical models for the analysis of *Stochastic Patch Occupancy Models* (SPOM) (37; 60). In this mathematical model the patches have only two states: occupied or empty. This gives the possibility to have 2^n different states in the system (system with n patches) which makes calculations difficult to perform. To simplify the mathematical analysis of the resulting equations it is necessary to assume that all patches are homogeneous (33). The SMT model is obtained by combining heterogeneous SPOM (different rates of extinction and colonisation in the patches) with assumptions in how the structure of the landscape affects these probabilities. In order to make a prediction of these models, it is necessary to create a SPOM model with the same dynamical function as in the SMT models and then fit the parameters of the SPOM model so that it "replicates" the dynamics of the SMT model. The results of analysing the homogeneous version of the model make a good approximation of the behaviour of the heterogeneous model (34).

2.2 Simple Dynamics in Complex Environments - Networks Theory

We define a network as a set of items called *nodes* that are connected to each other by *edges*. This is just a model to represent groups of individuals and links between them. In mathematic terminology they are known as *graphs*.

Although the definition is very simple, when we model a real-life system as a network, the outcome can get very complicated and difficult to understand. Networks can have a very complex wiring diagram, or they can change over time (like the World Wide Web), they can have different kind of nodes, or their links can have different weights. Some times the nodes themselves can be nonlinear dynamical systems.

The aim of the so called *Theory of Complex Networks* is to understand how complex dynamics take place in complex networks. But, in order to make some progress, different fields have overlooked certain complications and focused on the complexity of other properties. For example, there is a tendency to oversimplify the wiring diagram of the system by, for example, assuming they are geometrically regular and static in order to concentrate on the dynamics (93; 109). In other fields there is a stronger emphasis on the study of the network topology and its properties by using computer simulations and statistical mechanics theory (3; 66).

There are different kinds of real world systems that can be, and have been, modelled as networks. We can talk about **social networks**, which represent sets of people with some pattern of contacts or interaction between them, like friendship, business relationship, etc. **information networks** which reflect the information stored in its vertexes, like for example the citations between academic papers, (75), or links in pages of the World Wide Web, (4), **technological networks** like the electric power grid, computer networks, networks of roads, trains, or airline routes, (64), **biological networks** such as food webs, (61), neuronal networks, (51), network of insects, (27), or even genetic regulatory networks with the use of random Boolean net by Kauffman, (45).

2.2.1 ER-Random Graphs

One of the first important papers about networks (26) was written by Erdös and Rényi in 1959. In this paper they introduce a definition for **Random Graph**, also referred to as **ER-Random Graphs**, and the solution for some of its statistical properties when the number of edges tends to infinity. Modelling and studying random graphs was the first attempt to try to understand the complexity of networks. They are modelled as follows: *"We draw n nodes in a paper, and we choose a number m that will be the number of edges in the graph. We choose two nodes at random and we connect them with a link between them. Then you repeat*

this m times, until there are actually m edges in the graph. We have the model of a random network." (39).

One question that we can ask about this kind of network is: how many of the nodes are linked together? And the answer is: when the number of edges m is small we will find a lot of small clusters isolated from each other. As m gets bigger the clusters start joining together and, when m passes a threshold ($m > \frac{n}{2}$), all small clusters will join spontaneously to form one big cluster, called a **giant component**, and some other smaller clusters. It has been rigourously proved that the size of the giant component is of the order of n and there will be some other smaller components (3; 66).

2.2.2 Small World Effect

Another question we can ask about this network is: What is the average distance between the nodes of this big cluster? The formal definition for **average path length** in a graph is as follows: Consider an unweighed graph with the set of vertices X. Let $d(x_1, x_2)$ denote the shortest distance between x_1 and x_2. Assume that $d(x_1, x_2) = 0$ if $x_1 = x_2$ or x_2 cannot be reached from x_1. Then, the average path length l_G is:

$$l_G = \frac{1}{N(N-1)} \sum_{i,j} d(x_i, x_j)$$

where N is the number of vertices in X (definition taken from (103)). When the average path length between two vertices in a graph will grow in logarithmic manner, or slower, with the number of nodes in the network (which is the case with ER-Random Graphs), we say that the graph has a short path connectivity. The consequence of having short average path length, or short path connectivity, on a network is that most of the vertices of the network are connected by a short path no matter how big the network is. In social networks this effect is known as the **small-world effect**[1]. This property has been found in most of the real

[1]This is different from the definition of **Small World Graph** from Strogatz and Watts in (99) which is given later in this chapter (see Section 2.2.4).

world networks that have been studied (4; 64), and it has a great importance in the dynamics that are going to take place in the network because, if all vertex are connected with 'short paths', the information in the network will spread more efficiently. For example, it has been observed that the average diameter of the World Wide Web was 19 links in 1999, and it is expected that its number of links will increase 1,000 per cent, but the average diameter will be 21. This means that all the information that you can find in the internet is 'a few clicks away' (4). This property has also great impact in the spread of disease (68). The idea of studying the small-world effect as a property of a network comes originally from the conclusions drawn from the well known Milgram's experiment (48) (the "six degrees of separation" effect), although it has been argued that the validity of the conclusions are not as strong as it is claimed due to the restriction of the experiment (48; 86).

2.2.3 Degree Distribution

Other property that characterise a network is its degree distribution. By **degree distribution** of a network we mean the following: let us define p_k as the fraction of vertices of the network that is connected to other k vertices (we say k is the degree of the vertex). If we plot the histogram of p_k, we get the degree distribution of the network.

An advantage of working with ER-Random networks is that they have been intensively studied and there are very accurate results of their statistical properties (4; 64; 93). One of the well known properties of these networks is that the degree distribution follows a Poisson distribution. Unfortunately, most of the real-world networks that have been plotted in a diagram do not follow a Poisson degree distribution (4; 5; 63).

In 1999 it was already discussed in a publication that ER-Random networks may not be an accurate model for real networks (15). Barabási et al. discuss that

many real networks found in nature follow in fact a power-law degree distribution. Power law degree distribution is a common phenomenon in social networks, computer networks such as the World Wide Web (2; 4), citation networks (75), or even in electricity power networks as they show in their work. This issue is also discussed by other authors (4; 70). Barabási suggests that two specific characteristics are necessary for the formation of these networks: 1.- constant growth in the network, and 2.- preferential attachment. He discusses why power-law distribution is a very important property in a network that had not been predicted by the generated random networks at that time. Most real-word networks show power-law distributions with exponent between -2 and -3, (65; 74), or exponential distributions mixed with power-law distribution, (14; 64). Other observed degree distributions of real world networks are exponential, like in the railway networks or the power grid (5). Networks with power-law distribution are referred to as **scale free** networks because the average number of connections per node remains constant regardless the size of the network.

Newman has made substantial progress in the research on statistical properties of networks of arbitrary degree distributions by using **generating functions** (71) (which we will introduce later in this chapter), and he has used this tool to study statistical properties of scale free networks.

Although we are mainly interested in the dynamical process that take on complex networks, this type of properties are of our interest as they do make a difference on the overall results of the system as a whole. This is shown in (92), where they present a study of how the structure of a network can affect the stability or instability of opinions in individuals. The conclusions are that in flat hierarchies networks (such as lattices) the opinion of the people are more likely to be unstable, whereas in scale-free networks there is an inherit stability.

2.2.4 Clustering

In many real world networks, specially in social networks, it is very common that two nodes are connected to each other if they share a common neighbour. In social terms we could say that there is a higher probability that two people know each other if they have a common friend than if they do not. This relation between the individuals is represented as triangles in the network. This property is called **clustering** and it is measured by a **clustering coefficient** that determines the density of triangles in the network (99). The formal definition of cluster coefficient C was first introduced by Watts and Strogatz as follows: *Suppose that a vertex v has k neighbours; then at most $\frac{k(k-1)}{2}$ edges can exist between these neighbours (this occurs when every neighbour of v is connected to every other neighbour of v). Let C_v denote the fraction of these allowable edges that actually exist. Define C as the average of C_v over all v.*

As I said before, social networks show very high clustering coefficient compared with those in random networks (66; 69; 99). In fact, most of the social networks studied have the properties of clustering (they have a high clustering coefficient) and the small world effect. A simple network that has the form of regular lattice has a very high clustering coefficient, but they normally do not satisfy the small world effect (that the average path length between its nodes increases linearly with the number of nodes). On the other hand, ER-Random Networks have the short path connectivity property, but do not have as much clustering as real-world networks. Strogatz and Watts (99) propose a different way of modelling networks by creating what they called **small world networks**, and they are modelled as follows: *Starting from a ring lattice with n vertices and K edges per vertex, we rewire each edge at random with probability p. This construction allows us to 'tune' the graph between regularity (p=0) and disorder (p=1) and probe the intermediate region $0 < p < 1$, about which little is known.* This process adds some short cuts in the network that allow its average path length to be reduced notably but its clustering coefficient will remain almost the

same. This small world model is therefore a better approximation than Regular Lattices or ER-Random Networks to what the social networks may really be. Unfortunately, Small World Networks' degree distributions do not follow a power-law distribution.

Newman also proposes an algorithm to create networks with the desired number of triangles in a network (67). We define a sequence of numbers $s_0, s_1, \ldots, s_N$ that determines the number s_i of "stubs" or single edges that emerge for vertex i, and another sequence of numbers $t_0, t_1, \ldots, t_N$ that determines the number t_i of corners of triangles that emerge from vertex i. We create the network by choosing pairs of stubs uniformly at random and joining them to make complete edges, and also choosing trios of corners at random and joining them to form complete triangles. The only constraint is that the total number of stubs must be a multiple of 2 and the total number of corners a multiple of 3. The end result is a network drawn uniformly at random from the set of all possible matchings of stubs and corners.

We can still mention other properties of complex networks. In most kind of modelled networks there are different types of vertices, and sometimes vertices of the same type have more probability to be connected between them. This property is called **assortative mixing** or **correlation**, and it is also a common property of social networks (74). For example, it is known that people tend to preferentially mix with others of a similar age. This age-assortative mixing may have great influence in the dynamics of a disease in a community such as measles, where the mixing between school children drives the epidemic process, or the recent H1N1 pandemic, which has shown a great age-dependent susceptibility in its records (46; 62).

If we can classify types of vertices according to their vertex degree this correlation is called **degree correlation** and, again, high degree correlation is another common property of social networks (66). Although there is extensive literature on how to calculate statistical properties of networks of arbitrary degree distributions it is has not been so thoroughly done for the calculation of correlation or

clustering coefficients (65). Another property of social networks is that they are divided into communities as well as showing positive correlation. Newman suggests that the division into different groups may affect this positive correlation. He has done a rigourous analysis predicting the correlation coefficient in terms of the variation of the sizes of the different groups and it compares well with what has been observed in real-world networks (69).

2.2.5 Analysing Networks - Generating Functions

In 2001 Newman et al., (70), present and discuss the statistical properties of random graphs with arbitrary degree distribution (not just for ER-Random graphs). The analysis of these statistical properties is based on generating functions (107). They build the generating function $G_0(x)$ for the degree distribution in the graph. Let us suppose that we have a graph of N vertices, with N large. They define the generation function for the degree distribution in the graph as: $G_0(x) = \sum_{k=0}^{N} p_k x^k$ where p_k is the probability that a randomly chosen vertex on the graph has degree k. The probability p_k is given by the k^{th} derivative of G_0 as follows:

$$p_k = \frac{1}{k!} \frac{d^k G_0}{dx^k} \bigg|_{x=0}$$

So the one function $G_0(x)$ encapsulates all the information contained in the probability distribution p_k. They say that function G_0 "generates" the probability distribution p_k. For example, if we want to find the generation function of the Erdős-Rényi Random Graph, (26), we would get:

$$G_0(x) = e^{-z} \sum_{k=0}^{\infty} \frac{z^k}{k!} x^k = e^{z(x-1)}$$

By analysing this function they get to deduce some statistical properties of the graph. We are only going to mention a few properties that relevant for the

analysis of a SIR[1] system in a network. Among other properties they get to deduce: 1.- the mean component size of the graph when there is not a giant component in the graph, 2.- the threshold at which the giant component of the graph is formed in terms of the average size of the components of the graph, and 3.- the size of the giant component S as a solution of the equations system:

$$S = 1 - G_0(u), u = G_1(u)$$

where G_0 is the generating function of the degree distribution of the graph and $G_1(x) = \frac{G_0'(x)}{\sum_k k p_k}$. The calculation of the component size distribution as a solution of these equations is not normally solvable in closed form but a solution can be found using iteration starting from a suitable initial value of u.

This method of studying the properties of a graph using generating functions is also extended to directed graphs[2] and to bipartite graphs [3] (70; 71). The mathematical results are contrasted with real data and it is found that for bipartite graphs the similarity of the results with real data is very good, but it is not so in the case of directed graphs. There are substantial quantitative differences between the predictions of the results and the characteristics of these networks

[1]SIR stands for susceptible-infected-removed and it is a version of the susceptible-infected-susceptible process in which infected individuals cannot go back to a susceptible state. Instead, infected individuals will go to a "R" (removed or recovered) state at a certain rate and they will remain in such state for ever. In some models the "R" state represents a state of recovery from a disease in which reinfection is not possible. In other models it may represent death. Once an individual reaches the removed or recovered state, it will not play a role in the dynamics of the model any longer.

[2]A directed graph is a graph whose edges have a direction from one vertex to another, but not viceversa. These type of graphs are used, for example, to represent a network of web-sites with hyper-links between them. A web-site A may have a link to another web-site B but there may not be a link from B to A.

[3]A bipartite graph is a graph that has two different types of vertices representing elements with different properties. These type of graphs are used, for example, to model affiliation networks. An affiliation network is a network in which actors are joined together by common membership of groups or clubs of some kind. They are represented using two kinds of vertices, one representing the actors, and the other representing the groups. Edges then run only between vertices of unlike kinds connecting actors to the groups to which they be long.

shown in real data. This may indicate that there are interesting social structures in these networks that are not captured by a simple ER-Random Graph model.

In (64) Newman et al. present an analysis of what they call "network resilience" based on percolation theory which is then going to be used for solving the SIR problem in a network. They present the "network resilience" problem as follows: Consider a model defined on a network on which each vertex is either "present" or "absent". They call it "absent" because for some reason they may be non-functional on the graph. They define probability b of being present in the graph (let's assume this probability does not depend on the degree of the vertex). They define the generation function of the probability for a vertex of degree k to be present in the graph as: $F_0(x) = \sum_{k=0}^{\infty} p_k b x_k$, so $F_0(x) = bG_0(x)$. Having defined this generating function and following the same steps used to find the characteristics of the graph mentioned in the previous paragraph they can find out: 1.- the mean size of a cluster of connected vertices, 2.- the critical value, in terms of the probability b for which a giant cluster is formed, and 3.- the size of the giant cluster of "present" vertex on the network once the giant cluster has been formed.

In (64) the SIR model is solved by reducing the problem to a percolation problem. They formulate the SIR model in a network as follows: Consider an outbreak on a network that starts with a single infected individual and spreads on the network until a subset of the network is infected. The vertices of the network represent the potential hosts and the edges represent contact between the possible hosts. If we imagine occupying the network in all edges that have resulted in transmission of the disease, the set of vertices representing the infected hosts form a connected percolation cluster in the network. They denote by τ the time that a host remains infected and by r the probability that an infected host transmits its infection to a neighbour, the total probability for infection is:

$$T = 1 - e^{-r\tau}$$

They call this quantity *transmissibility*. By using the transmissibility T as the probability of being "present" in the graph (i.e. by substitution b by T in the

analysis above) the SIR problem is reduced to a percolation problem. Using the "network resilience" analysis, based on percolation theory, we can find: 1.- the average size of the distribution of outbreaks of the disease (given by the distribution of percolation clusters on the network), 2.- *epidemic threshold* for the disease (given by the percolation threshold at which the giant cluster is formed in the network), and 3.- the total size of the epidemic (given by the size of giant cluster of the percolation problem).

In (68) Newman et al. presents a two dimensional small-world model to study the spread of plant disease. A mathematical analysis of the model using generating functions predicts accurately the percolation threshold and the epidemic size in function of the transmission probability of the disease and the density of the shortcuts added to the initial lattice.

In paper (67) Newman extends the use of generating functions to study the statistical properties of networks created with a given number of triangles in the network. As we explained before, he constructs a network from a sequence that determines the number s_i of stubs and the number of corners of triangles t_i that emerge from each vertex i of the network. He then creates two generating functions, based on the two sequences, and by combining them he makes a similar analysis of properties such as the size of the giant component if there is one, the sizes of the small components of the network, average path lengths, and other. He is also able to calculate percolation properties such as percolation threshold, mean size of small clusters, and complete distribution of small clusters. As we mention above, this analysis of percolation properties is then used for the understanding of SIR processes in clustered networks.

2.2.6 Networks with Hierarchical Structures

The spread of diseases in social networks has also been studied by a network model with hierarchical structure (31). This hierarchical structure represents in-

teractions of different intensities among individuals depending on their relations, for example, families, co-workers, or other more sporadic relations. In this work the spread of a SIS epidemic model is implemented in a scale-free network model with hierarchical structure. The implementation of the hierarchy structure affects the formation on the network (there is more probabilities to find an edge between nodes of the same group), and the spread of the disease (probability of infection is higher among nodes in the same group). The experiments focus on studying the speed at which the epidemic spreads in the network, the peak number of infections, how the epidemic spreads among different hierarchies, the clustering coefficients, and how the maximum degree value of the network affects the spread of the epidemic. They found that the epidemic spreads faster among nodes in the same group, that high clustering coefficient values slowed the spread of the epidemic, and that an increase in the maximal degree value of the network accelerates the spreading process. They also compare the results with a differential equation called *Master Equation*, which is some sort of mean-field approximation that does not take into consideration hierarchy in the network. The value at which the epidemic stabilises is very similar in the Master Equation and the model, but the speed at which the epidemic spreads at the beginning and the maximal number of illnesses in the network are higher in the model than in the Master Equation, so it is concluded that the Master Equation should not be used as a predictor.

As an extension of the work presented above, the same authors also study the effect of random and targeted vaccination on the spread of the disease and the effect of taking sick leave from work (32). The results show that randomly targeted vaccination is not very effective since the number of people that need to be vaccinated is very high, but targeted vaccination and taking sick leave are good ways of preventing and controlling the epidemic.

2.3 Models for Experimental Purposes

Below I give some examples of systems that simulate the movement of population using agent-based techniques, and which have been developed for experimental purposes. These models are designed to answer very specific questions about very specific populations and specific landscapes, and they implement the environment, the agents, and the interaction between them in very detail and with high complexity.

I previously introduced the Metapopulation approach used for the study of persistence of ecosystems in fragmented landscapes. Other field that is interested in this matter is *landscape ecology* (105). Landscape ecology and metapopulation ecology share the common goal of predicting the persistence of spatially structured populations in fragmented landscape, but differ in the different approaches employed by ecologists from the two disciplines. Landscape ecology uses models designed to answer specific questions about specific populations in specific landscapes whereas metapopulation theory is more general and focuses on the construction of "toy" models, which are more abstract and can explore more general principles of the system's dynamics. Landscape ecology methodology uses a very detailed description of landscape structure and emphasis of the individual movement. Their most important tool is individual-based simulation. They tend to use real data based on Geographical Information Systems and spatial statistics to build simulations that model dynamics in very complex landscapes (7; 53; 58).

SimPed (41) is a system that is concerned with the evolution of urban structure and with the social activities of humans within urban environments. This system is modelled with StarLogo (90) (agent based simulation tool developed from Logo as a programming language for children) and the urban environment is represented using an approach called space syntax, used for spatial analysis. The results of experiments using this tool showed a significant correlation between pedestrian movement and the morphological structure of the space in which these

agents move. This system was used as an experimental tool, to test some hypothesis made about urban behaviour, but no theoretical proof of the results has been done.

Another system developed for the study of urban movement is TRANSIMS (TRansportation ANalysis SIMulation System) (1). This system is being developed in Los Alamos National Laboratory. It consists of a set of mutually supporting simulations, models, and databases that are used to create an integrated regional transportation system analysis environment. Los Alamos has done studies of Albuquerque and Dallas transportation and completed a microsimulation of auto traffic patterns in 25 square miles of Dallas that represented about 200,000 vehicles over a five-hour period. The goal of TRANSIMS is to develop technologies that can be used by transportation planners in any urban environment.

Other systems have been developed for the understanding of crowd behaviour in emergency situations and used as support in developing efficient rescue plans (91; 110). These systems are also used as tools for predicting crowd behaviour in different situations, but no deeper analysis of the dynamics of these systems has been done.

2.4 Other Matters of Epidemiology - Effective Reproductive Rate

Anderson and May have done a great work introducing and analysing thoroughly a few simple mathematical models of the transmission of infectious agents in humans (6; 56; 57). The aim for these models is to help to interpret observed epidemiological trends, to understand collections of data in more depth, and to design programs for control of these infections, (6). They use mathematical studies as a tool for understanding epidemiological processes, therefore the models

are very simple. They regard these models as something that provides insight into essential aspects of epidemics and as a point of departure for adding realistic complications step by step in an understandable way. In their book they introduce the concept of Reproductive Rate, R_0, (the number of new infections that a host would produce in a totally susceptible population), and the effective reproductive rate, R, (the average number of actual new infections produced by a host among his contacts during the epidemic process). These quantities satisfy the inequality $R \leq R_0$.

As it is explained in paper (94), for an epidemic (with no recovery) that spreads polynomially over a period of time, the value of the effective reproductive rate must be very close to 1, i.e., the average number of people that one individual transmits its virus to is just 1. If $R > 1+c$ with c positive, the number of infections after time t will exceed $(1 + c)^t$, and the disease spreads exponentially. On the other hand, if $R < 1 - c$, the epidemic dies out. The initial reproductive rate R_0 parameter also plays an important role on the extinction or persistence of the disease in the population. It is known that disease dynamics present random fluctuations in their size (44). This fluctuations cause the disease-free state to be reached, and this extinction process occurs even when R_0 is greater than unity, but not large enough (85).

Anderson and May (56) estimate in their calculations that the basic reproductive rate R_0 of HIV is well above unity. However, if R was also above unity the spread of the disease would be exponential and it is shown otherwise (94). Therefore there must be a significant reduction in the effective reproduction rate of the epidemics R.

Different people explain this phenomena using different arguments, but in paper (94) Balázs proposes that the main reason for viruses without recovery to be able to spread at a rate lower than exponential is clustering in social networks. He argues that clustering among the social networks can have sufficient local effect on the epidemic to modify the basic reproduction rate form a high R_0 to an effective reproductive rate $R = 1$.

Part II

Vector-Borne Disease Model
A Metapopulation Approach

Chapter 3

Vector-Borne Disease Model A Metapopulation Approach

3.1 Overview

In this chapter we present a model of spread of a vector-borne disease in a country taking malaria as an example. The idea of this model is to investigate how a population moving around a heterogeneous environment affects the dynamics of the disease. This model has been implemented in five steps, creating 5 different sub-models with increasing complexity by adding different factors to the original model in each step. Each sub-model has been simulated RePast and analysed mathematically. The mathematical analysis seeks to understand what is the steady-state of the disease for any parameter values. For this we create a mathematical abstraction of each system to be deduced from the rules the individuals follow in each sub-model. We then try to find the fixed point of this mathematical abstraction and study their stability.

3.2 Background

Malaria is one of the best known parasitic diseases. More than 200 million people die of malaria every year and it is endemic in more than 90 countries. Over 40% of the world population are exposed to varying degrees of malaria risk in about 100 countries (24). Malaria is caused by a parasite that hosts in the human body and transmits from human to human via mosquitoes. There are four different types of malaria, but, for the purposes of our research, we will only consider *Plasmodium falciparum* as it is the one that causes the most serious infections.

The parasite is transmitted via the female mosquitoes *Anopheles*. Female mosquitoes need blood for the development of the eggs in their reproduction cycle. If this blood is infected, the parasites can multiply and develop inside the mosquito and can be transmitted to another susceptible human in the next blood meal. The female mosquito takes a blood meal approximately every two or three days and the parasite needs to be hosted in the mosquito for at least 10 days before it is ready to be transmitted to another human. A female anopheline mosquito normally lives between two or three weeks (43).

The number and type of mosquitoes differ depending on the seasons and the environmental conditions of the area. In urban areas the mosquitoes are more rare than in rural areas, but they normally reproduce in swamps, a fact that makes those areas very highly populated by mosquitoes and, therefore, very likely for transmission. Incubation period takes between 8 and 25 days before the symptoms start. If malaria is detected at an early stage (within the first 2 or 3 days of symptoms) it is very easy to treat and it can take only three days to combat it. If it is not treated, complications can develop. In fact, the delay in diagnosis and treatment is the main reason for the high level of mortality for the disease (24; 42). It has been documented that in endemic areas immunity to malaria is acquired after repeated infections, but the immunity process at the molecular level is not well understood (20; 59).

Human mobility between different countries and areas influences strongly the re-emergence of the disease in different countries (55). Most of the models of malaria assume a constant population and constant rates of infection. We have modelled the spread of malaria in a country considering different areas with different probabilities of transmitting the infection. Our major priority is to understand how the movement of the population around different areas affects the spread of the parasite.

The model presented in this chapter is an abstract representation of the malaria mechanisms in a moving population. However, it can be used to understand the dynamics of any vector-borne disease in which there is movement in the population. There are two main factors that determine this model, the participation of two different populations in the spread of the disease (mosquitos and human), and the movement of people around different areas (which may represent countries, villages or other type of areas of smaller or larger scale). This model can also be used for the understanding of any other vector-borne disease in which movement of the population may take a role in the disease dynamics. Vector-borne diseases are defined as the diseases which are transmitted through vectors. By vectors we understand any animal that transmits human diseases such as mosquitoes, rodents, fleas etc. Vectors are able to transmit the disease to humans by biting, burrowing, or contaminating living spaces.

Given the characteristics of the system we want to model it seems natural to take a metapopulation approach (introduced in the Background Chapter), in which there is an emphasis in two aspects: the heterogeneity of the environment where the system dynamics take place, and the movement of the population in the modelled environment. It has been already discussed in the background section how the introduction of space and movement in a model can make a difference in the system dynamics as, for example, (19) shows that there is a difference between a perfect mixing mathematical model and the model spread in a lattice (when it is only partially occupied). They also show a difference between a lattice models only partially occupied without movement and the introduction of movement in the models.

3.3 The Model

We consider two different populations, human and vectors (mosquitoes), both modelled as a set of individuals that are located in the country. It is the interaction between these two populations that gives rise to the epidemic dynamics. We do not model birth or deaths in either population, so the number of vectors and people are going to be constant in the model.

We know that, although most people live in cities, and in cities there are not many mosquitoes, people do move around the country by car or by plane, and spend certain time in a given area before they move to the next area. We are interested in how this movement affects the dynamics of the disease, so we will model people moving between different areas of the environment by following certain rules stated in the next section. We know that mosquitoes fly and move too, but we believe that, at the spacial scale that we are focusing on, the movement of the mosquitos should not have an effect in the system.

We know that in different parts of the country there may be different numbers of mosquitoes and different numbers of people. We model the country as a group of different patches connected to each other by roads. The people would move from patch to patch with no pre-determined route. The movement of the population around the country is simulated by giving each individual, with certain probability, the intention to leave and go to some random patch that is connected to its current location by a road. We model three different types of patches representing three different types of area: **Swamps**, **Rural** areas and **Cities**, each of which have different characteristics. What is going to characterize each type of patch is a set of parameters: 1.-the number of mosquitoes in the patch V_i, 2.-the probability for a person to leave this patch l_i, and 3.-the dimension of the patch w_i.

Given these three different types of patch we model them so that in a rural patch there are more mosquitoes than in a city, but less than in a swamp patch.

The probability to leave a patch depends on the location of the individual, so the individual has higher probability to leave a rural patch than a city patch, and the highest probability to leave is given by being in a swamp. Also the dimension of each patch modelled will depend on their type, with swamps having the smallest dimension, then city patches being larger, and rural patches being the largest patch type.

We model two different states in the human population, susceptible and infected, and the same for the vector population. We know that humans can get immune to malaria after having been infected for some time, but we do not really know how it works so we have decided to ignore immunity in this model. We assume that all infected people get treatment and recover. A human gets infected when he/she is bitten by an infected vector. We assume that this human remains infected for some time and then he/she recovers, becoming again susceptible to the disease. The infection time length in people is the same for all, R_p time steps. A vector gets infected when it bites a human that carries malaria. A vector also stays infected for some time until it becomes susceptible. This is not the case in the real system, however, this would represent the death of infected vector and the replacement of those by new susceptible vector. This is a more simple way to represent its life dynamics; by keeping the population of mosquitoes constant. In the model the infection time length in mosquitoes is also the same for all, R_v time steps.

3.3.1 The Environment

The representation of the environment and the movement of this vector-borne disease model is based on the Botellón simulation (16; 30; 77). The different areas are represented as vertices of a graph. Areas that are directly connected by a road have edges between them (see Figure 3.1). Each area has a type (City, Rural or Swamp) and each area type has a given number of vectors (V_i) and a probability to leave that area (l_i) and an area size (w_i).

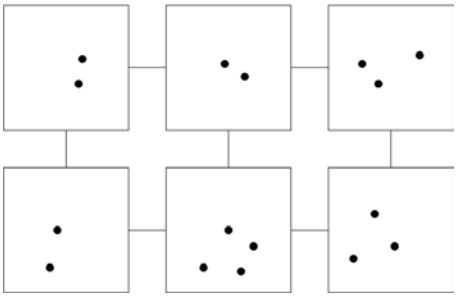


Figure 3.1: Graphical representation of the environment used in the vector-borne disease models.

Each area type is modelled as a 2 dimensional grid formed by w_i cells. At each time-step every individual (vectors and people) takes a random position within the patch (see Figure 3.2).

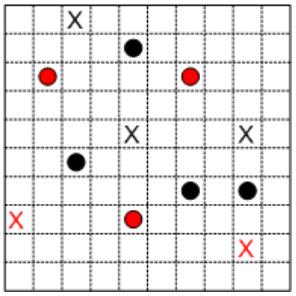


Figure 3.2: Graphical representation of a single patch in the vector-borne disease model. The patch is subdivided into a number of independent cells. Each cell can be occupied by none, one, or more than one individual. Crosses represent vectors, and dots represent people. Red colour represents infected state, and black colour represents susceptible states.

3.3.2 Movement Rules

People are modelled as individuals moving around these areas following simple probabilistic rules. The rules determine how the individuals interact with the environment:

- **Deciding** to stay or to go: This is calculated probabilistically. Different patch types have different probabilities for each person to stay or to go.

- **Choosing** what neighbour patch to move to: If the person has decided to leave its current location, it will choose a random neighbouring patch to move to. Neighbouring patches have been predetermined by the formation of the grid 3.1. All neighbouring patches have the same probability to be chosen.

- **Placing** the person in its chosen location: If the person decides to stay we randomly give the person a new random location within the set of cells in its current patch. If the person decides to leave we randomly give the person a new random location within the set of cells of the new patch.

3.3.3 The Disease Dynamics

The disease dynamics take place in each patch independently. In each patch we have a number of **susceptible** people and a number of **infected** people. We also have a number of **susceptible** vectors and a number of **infected** vectors.

In each patch of the environment the following happens: At each time step each person and each vector in the patch is positioned into a randomly chosen cell within the patch. The number of cells to chose from for each individual are exactly w. The process of choosing the cell is independent for each vector and person, so the same cell may be chosen by more than one individual. We say that two individuals get in contact with each other if they get placed on the same cell.

If a susceptible person gets in contact with an infected vector, the person will change its health state to infected, and it will remain in this group for R_p time-steps. If an infected person gets in contact with an infected vector before this time has elapsed, the infection is sustained by another R_p time steps before

the person goes back to the susceptible state. This process is an assumption in our model as it is not known how the re-infection process works in malaria.

In a symmetric manner, if a susceptible vector gets in contact with an an infected person, the vector will change its health state to infected, and it will remain in this group for R_v time-steps. If an infected vector gets in contact with an infected person before this time has elapsed, then the infection is sustained by another R_v time steps before the vector goes back to the susceptible state.

3.4 Step by Step Implementation and Mathematical Analysis

In order to understand the complexity of the model described above, we use a "step by step" approach. We first implement a simple model with not many parameters, analysing it mathematically and understanding the dynamics of it. We then implement new parameters in the model increasing its complexity. We introduce all the parameters in an incremental manner making a mathematical analysis of those models individually.

The incremental implementation and analysis will be as follows:

Model 1 We model the movement of people in a set of patches and try to find the average number of people in each patch in a steady-state. The disease process is not modelled at this point.

Model 2 We model a simple disease model in a single patch. The probability of transmission in the patch will be a constant number.

Model 3 We model a set of patches with movement of people between the patches. Within each patch we implement the simple disease model studied in the previous model. Each patch will have a constant probability of transmission

and we are interested in understanding how the movement of people between those patches affects the general dynamics of the model.

Model 4 We implement and analyse a vector-borne disease model in one single patch.

Model 5 We implement and analyse the model above in a set of patches considering different values for some parameters to give a heterogeneous structure in the space. At this point we also introduce movement of people between the patches.

3.4.1 Model 1 - Implementing and Analysing the Movement

In this first step we will only model a set of patches and movement of people between the patches. The environment is always modelled as a set of patches forming a 2 dimensional grid that are connected to each other as follows: Each patch is connected with the patches that are located next to it on the four directions right, left, above and below, as long as there are patches positioned on those locations (i.e. the patches that are on the top row will not have a neighbour above them, and so on...).

In order to get a sensible mathematical description of the model we number the vertices of the graph $\{0, 1, \ldots, N-1\}$, (see figure 3.3). We also name the probability for a person to leave a patch n at time t as $l_n(t)$. We name $p_n(t)$ the number of people in patch n at time t. The state of the system at any discrete time step is given by the vector $\vec{p}(t) = (p_0(t), \ldots, p_{N-1}(t))$. The total number of people in the model is $P = \sum_{n=0}^{N-1} p_n(t)$.

The procedure is very simple. First, let us try to find the number of people in one single area i. Supposing that we know the number of people at time *t-1*, let us try to predict the number of people at time t:

$$p_i(t) = p_i(t-1) - peopleLeaving_i(t) + peopleComing_i(t)$$

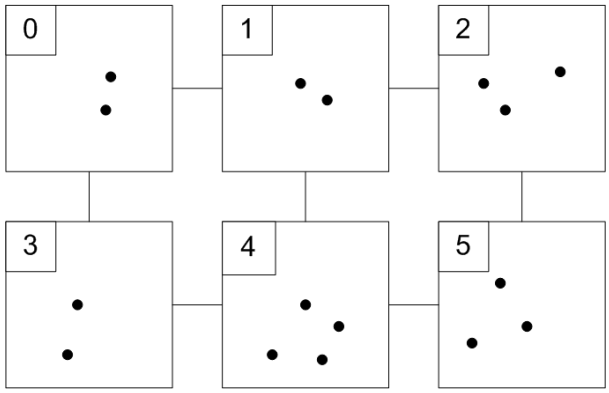


Figure 3.3: Numbering the patches of the network.

- The number of people leaving the area i at time-step t can be estimated as the number of people in i at time t, times the probability l_i of leaving the patch.

$$peopleLeaving_i(t) = l_i p_i(t-1)$$

- The number of people coming to patch i at time-step t can be estimated as the sum of the people leaving each patch, times the probability of those people to choose this patch i as its new location. Let us name $e_{i,j}$ the probability for a person located in j to choose i:

$$peopleComing_i(t) = \sum_j e_{i,j} peopleLeaving_j(t) = \sum_j e_{i,j} l_j p_j(t-1)$$

All neighbouring patches have the same probability to be chosen:

$$e_{i,j} = \frac{[i \text{ is connected to } j]1}{c_j} \tag{3.1}$$

where $[i \text{ is connected to } j] = 1$ if i is connected to j and 0 otherwise, and c_j is the number of neighbouring patches that a patch j has.

We define the matrix A with entries

$$A_{i,j} := e_{i,j} \tag{3.2}$$

(probabilities of choosing patch i when the person is located in patch j) and the diagonal matrix L with entries

$$L_{i,i} := l_i \tag{3.3}$$

Substituting all this terms in our original equation we have:

$$\vec{p}(t) = \vec{p}(t-1) - L\vec{p}(t-1) + AL\vec{p}(t-1)$$

We define a new matrix, $\boldsymbol{M}$, called ***Matrix of Movement*** that calculates the change of number of people at each time step:

$$M := I - L + AL \tag{3.4}$$

and therefore the equality above is transformed into:

$$\vec{p}(t) = M\vec{p}(t-1)$$

M is going to be the matrix that describes the flow of people between patches at each time step, and a steady-state of the simulation is the vector $\vec{p}$ that satisfies:

$$\vec{p} = M\vec{p} \tag{3.5}$$

and $\sum_{i=0}^{n-1} p_i = P$ where P is the total number of population in the simulation.

So we have a continuous approximation of the steady-state of the number of people in each patch in the model (assuming there is a steady-state) given by 3.5, and a matrix M that describes the flow of people in the model at each time step.

The model presented here is a simplification of previous work presented on (16; 30; 77).

Example 1. *Let's consider a system of population of 350 individuals living in an environment consisting of 6 patches connected to each other.*

Let us consider the topology of this environment as shown in 3.4.

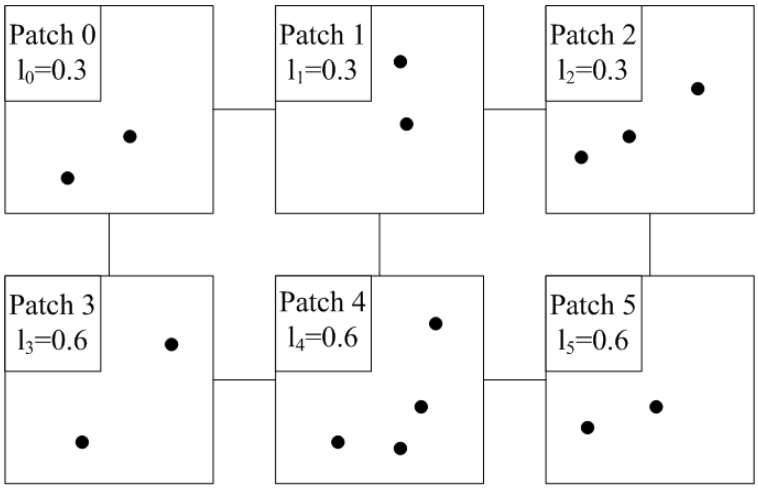


Figure 3.4: Model 1 - Graphical representation of the environment of the model used in Example 1. The dots do not reflect the actual number of individuals used in the example.

We name the patches as $\{0, 1, 2, 3, 4, 5\}$. *Let us assume that the probability for a person to leave the patch* n *is constant over time and it will be 0.3 for those patches in the first row (0, 1 and 2) and 0.6 for the patches in the second row (3, 4 and 5). So we have:* $l_0 = l_1 = l_2 = 0.3$ *and* $l_3 = l_4 = l_5 = 0.6$.

Looking at the topology of the environment we have that the number of connections of the patches are: $c_0 = 2, c_1 = 3, c_2 = 2, c_3 = 2, c_4 = 3, c_5 = 2$. *That makes our matrix* M *as:* $M = \begin{pmatrix} 0.7 & 0.1 & 0 & 0.3 & 0 & 0 \\ 0.15 & 0.7 & 0.15 & 0 & 0.2 & 0 \\ 0 & 0.1 & 0.7 & 0 & 0 & 0.3 \\ 0.15 & 0 & 0 & 0.4 & 0.2 & 0 \\ 0 & 0.1 & 0 & 0.3 & 0.4 & 0.3 \\ 0 & 0 & 0.15 & 0 & 0.2 & 0.4 \end{pmatrix}$

Solving our equation $\vec{p} = M\vec{p}$ *we have that* $\vec{p} = (\lambda\frac{2}{0.3}, \lambda\frac{3}{0.3}, \lambda\frac{2}{0.3}, \lambda\frac{2}{0.6}, \lambda\frac{3}{0.6}, \lambda\frac{2}{0.6})$. *Since the total number of people in our model is* $P = \sum_{n=0}^{5} p_n = 350$, *we get* $\lambda = 10$.

Therefore, a continuous approximation of the expected number of people in each patch, when the model is in its steady-state, should be:

$$\vec{p} = (66.6667, 100, 66.6667, 33.3333, 50, 33.3333)$$

and the movement of people in the model at each time step is given by the matrix M above.

We can see that the vector that defines the number of people in each patch when the simulation is in its steady-state $\vec{p} = (\frac{200}{3}, 1, \frac{200}{3}, \frac{100}{3}, \frac{100}{2}, \frac{100}{3})$ is an eigenvector of M with eigenvalue 1, i.e. the movement of the population does not affect the overall number of people in each patch.

3.4.2 Model 2 - Simple Disease Dynamics in a Single Patch

At this point we model the spread of a disease in a single patch. We introduce a new parameter d in the patch that indicates the probability for a person to get infected. This parameter will be constant over time.

The people will have one of two possible states: infected or susceptible. At each time step a susceptible person can change its state to infected with a probability d. The person will remain infected for a number of steps R. The number of people in this model is constant and it will be named as P.

In this model we also implement re-infection so, if an infected person gets transmitted the infection, it will remain infected for the next R following time steps.

If we name the number of susceptible people at time t as $s(t)$ and the number of infected people at time t as $i(t)$ we can develop the mathematical model for the described system as follows:

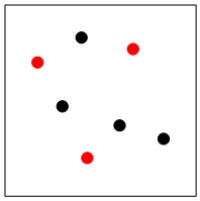


Figure 3.5: Model 2 -Graphical representation of the environment in Model 2. There is only one single patch in which people don't have a specific location. The disease is acquired by a susceptible individual with probability d, which is constant over time. Infected people are represented in red and susceptible in black.

We know that a person is susceptible if and only if it has not been infected in the last R time steps. The probability for not being infected in R time steps is $(1-d)^R$ so, the probability for a person to be susceptible is $(1-d)^R$. Therefore the probability for a person to be infected is: $1-(1-d)^R$.

$$\begin{aligned} s(t) &= (1-d)^R P \\ i(t) &= p-(1-d)^R P \end{aligned} \tag{3.6}$$

So it seems this model has a steady-state as the values of $s(t)$ and $i(t)$ don't depend on t, so we can say:

$$\begin{aligned} s &= (1-d)^R P \\ i &= p-(1-d)^R P \end{aligned} \tag{3.7}$$

Example 2. *This time we will consider a system of 350 individuals all living in the same patch. Let us assume that each person has a probability 0.01 to get infected at each time step, and it will remain infected for 40 time steps.*

In this case the mathematical result says that an estimate of the number of susceptible and infected people in the model in its steady state is:

$$\begin{aligned} s &\approx 234.14 \\ i &\approx 115.83 \end{aligned}$$

3.4.3 Model 3 - Movement Between Patches with Simple Disease Dynamics

We now create a model consisting of a set of "perfect mixing" patches $\{0, \ldots, N-1\}$. Each patch will have a value for the parameter that determines the probability of infection transmission $\{d_0, \ldots, d_{N-1}\}$. The recovery rate R is constant over the model. The number of people in each patch at each time step t is given by $\vec{p}(t) = (p_0(t), \ldots, p_{N-1}(t))$. We also represent the number of susceptible and infected people in each patch with vectors $\vec{s}(t)$ and $\vec{i}(t)$ where the n^{th} component is the number of susceptible $s_n(t)$ or infected people $i_n(t)$ in the patch n at time t respectively.

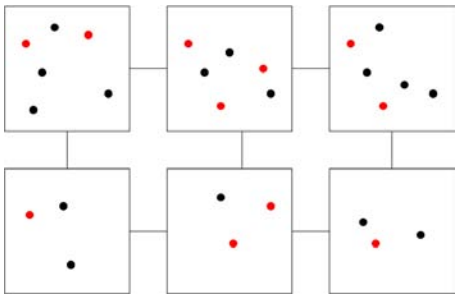


Figure 3.6: Model 3 - Graphical representation of the environment used in Model 3. A set of patches with simple disease dynamics forming a grid structure.

The equations that describe the spread of the disease in the model at each time step are based on the previous model. Assuming that the people in the model do not move from patch to patch, the dynamics of the disease can be described as:

$$\begin{aligned} s_n(t) &= (1-d_n)^R p_n(t) \\ i_n(t) &= p_n(t) - (1-d_n)^R p_n(t) \end{aligned} \tag{3.8}$$

If we define I_N as the identity matrix of size N, and D as a diagonal matrix with entries $D_{n,n} = d_n$ in its diagonal we can also represent the model above as follows:

$$\begin{aligned} \vec{s}(t) &= (I_N - D)^R \vec{p}(t) \\ \vec{i}(t) &= \vec{p}(t) - (I_N - D)^R \vec{p}(t) \end{aligned} \tag{3.9}$$

The equations above describe the dynamics of the disease in the patches when there is no movement in the model. If we want to include the movement that takes place in the model at each time step we have:

$$\begin{aligned} \vec{s}(t) &= [(I_N - D)M]^R \vec{p}(t - R) \\ \vec{i}(t) &= \vec{p}(t) - [(I_N - D)M]^R \vec{p}(t - R) \end{aligned} \tag{3.10}$$

In the equations above we consider the people at time $\vec{p}(t - R)$ rather than $\vec{p}(t)$ because we are looking for those people that have not been infected in the last R steps, therefore we want to trace this location of the population from time step $t - R$. The number of people in each patch will be updated at each time step with matrix M, and the infections in the population will be updated with matrix D. In the equation systems 3.8 and 3.9 we know that $\vec{p}(t) = \vec{p}(t - R)$ as there is no movement between the patches.

Going back to the equation system 3.10, and assuming there is a steady-state in the number of people in each patch, we would have as a result that there is also a steady-state in the number of infected and susceptible people in each patch given by:

$$\begin{aligned} \vec{s} &= [(I_N - D)M]^R \vec{p} \\ \vec{i} &= \vec{p} - [(I_N - D)M]^R \vec{p} \end{aligned} \tag{3.11}$$

Example 3. *Let's consider again a system with a population of 350 individuals living in an environment consisting on 6 patches connected to each other. The topology will be the same as in the first example, shown in 3.7. In this case we will*

have the patches classified in different types of patches: Big city, Small city, Rural or Swamp. The parameters for a person to get infected and the parameter for a person to leave the patch depend on the type of patch they are classified under. In this example the people will be classified following the graph in the Figure below:

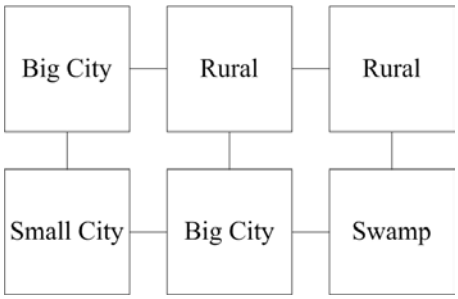


Figure 3.7: Classification of the patches of the environment used in Example 3

Let us assume that the parameters for each of the are as follows:

Patch type	*Value of* l_n	*Value of* d_n
Big city	0.2	0.01
Small city	0.4	0.02
Rural	0.7	0.05
Swamp	1	0.5

Knowing the topology of the environment and the values of l_n*, we construct the matrix M as we did in Example 1. We can also construct matrix D, knowing the values of* d_n*, and we also know the value of the number of people in each patch* $\vec{p}$ *as we did in Example 1. In order to find the fixed points of the system we run a program that computes the equations given by 3.11. This tells us that a continuous approximation of the number of infected s in each patch is:*

$$\vec{i} = (89.4161, 38.3212, 25.5474, 44.7080, 134.1241, 17.8832)$$

3.4.4 Model 4 - Model of Vector-Borne Disease in a Single Patch

The next step in our simulation is to model a vector-borne disease in a single patch. We now implement the two types of agents, people and vectors. They both have two states susceptible and infected. The patch will be a 2 dimensional grid formed by w cells. At each time-step all agents (vectors and people) take a random position within the patch.

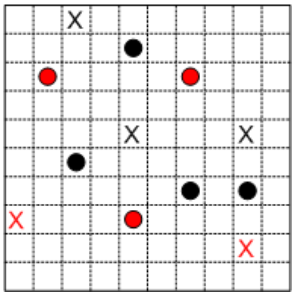


Figure 3.8: A single patch with Vector-borne disease.

If a person and a vector are in the same cell, we consider there is contact between them. If a person gets in contact with an infected vector it will remain infected for the following R_p time steps. If a vector gets in contact with an infected person it will remain infected for the following R_v time steps.

Given these rules, the probability for a person to remain in susceptible state at a given time t is equal to the probability for this person not to be in the same cell as any infected vector. This probability is: $(1 - 1/w)^{v(t)}$ [1] where $v(t)$ is the number of infected vectors in the system. Conversely, the probability for a vector to remain in susceptible state is equal to $(1 - 1/w)^{i(t)}$, where $i(t)$ is the number of infected people in the system.

[1] Note that the probability to be in the same cell as a given agent is $1/w$ as there are exactly w different cells in the system.

We are going to approximate $(1 - 1/w)^{v(t)}$ with $\exp\left(\frac{-v(t)}{w}\right)$. Looking at the Taylor series of these two functions [1] we can see that this approximation is really good for values of w sufficiently large compared to $v(t)$, i.e., when the area of the space represented in the model is large compared to the number of vectors in the system. If we look at the difference between these two functions when the number of vectors in the system is 10,000 and the total area w varies from 0 to 100,000, we can see that, in the worse case, the difference is still less than 0.00003.

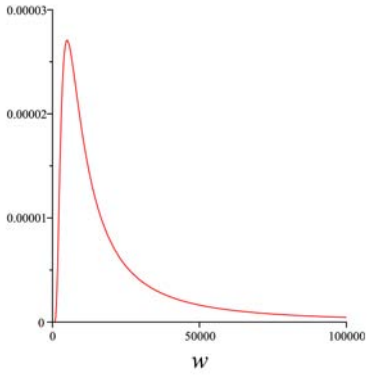


Figure 3.9: Comparing expressions $(1-1/w)^v$ and $\exp\left(\frac{-v}{w}\right)$ with $v = 10,000$ and w ranging from 0 to 100,000

In fact, this second mathematical expression represents the probability for a person to remain in susceptible state at a given time t in a model with continuous space of size w. Let us consider a continuous two dimensional space or area size w. Let us consider a number h of particles located in this space. Each particle will have a circle of radius r centered around it. If we randomly put another

[1] Let us define $g(x) = (1 - x)^v$ and $h(x) = \exp(-vx)$. The Taylor series of these two functions are: $g(x) = (1 - x)^v = 1 - vx + (v^2x^2)/2 - (vx^2)/2 + ...$ and $h(x) = \exp(-vx) = 1 - vx + (v^2x^2)/2 + ...$ As we can see these two functions approximate each other when vx are close to 0

particle in this space, and considering the circle of radius r centered around this new particle, the probability that this new circle does not overlap any other circle in the space would be $\exp\left(\frac{-hr}{w}\right)$.

If we consider each vector in our model as a particle in a 2 dimensional continuous space and we give the value $r = 1$, as the people in our model are located randomly in the space, for each person in the space we could approximate the probability for this person not to be in contact with any infected vector in the model to:

$$\exp\left(\frac{-v(t)}{w}\right)$$

in the same way, the probability for a vector not to get in contact with an infected person at time t can be approximately:

$$\exp\left(\frac{-i(t)}{w}\right)$$

The probability for a person to be susceptible at a given time step is the probability for that person not having been in contact with an infected vector for the last R_p time steps. That translates to:

$$s(t) = P\exp\left(-\frac{v(t-R_p)+v(t-R_p+1)+\cdots+v(t-1)}{w}\right)$$

So the number of infected people and infected vectors is and the infected population would be:

$$\begin{aligned} i(t) &= P - P\exp\left(-\frac{v(t-R_p)+v(t-R_p+1)+\cdots+v(t-1)}{w}\right) \\ v(t) &= V - V\exp\left(-\frac{i(t-R_v)+i(t-R_v+1)+\cdots+i(t-1)}{w}\right) \end{aligned}$$

Assuming there is a steady-state in the dynamical system we define v and i as:

$$i = \lim_{t\to\infty} i(t) = P\left(1 - \exp\left(\frac{-vR_p}{w}\right)\right) \tag{3.12}$$

$$v = \lim_{t\to\infty} v(t) = V\left(1 - \exp\left(\frac{-iR_v}{w}\right)\right) \tag{3.13}$$

This model coincides with the *Transmissibility* model by Newman (64) which is:

$$T = 1 - e^{-r\tau}$$

where τ is the time that an infected host remains infective,and by r the probability per unit time that the host will infect one of its neighbours in the network. However, Newman uses only one type of agent in its model while we use two different ones. That is why we get two coupled equations instead of one.

From the equations 3.12 and 3.13 we get one single equation:

$$i = P\left(1 - \exp\left(\frac{-V\left(1 - \exp\left(\frac{-iR_v}{w}\right)\right)R_p}{w}\right)\right)$$

We can see that $i = 0$ is always a fixed point ($i = 0 \Rightarrow v = 0$), but the mathematics get too complicated to find any other fixed point. An alternative way to find the fixed point would be to run the coupled equations 3.12 and 3.13 in a mathematical software and observe if they get to a fixed point. We could think that this approach does not give us the information we are interested in for two reason: First, the fact that we find some fixed points does not mean that we have all fixed points of the system. Second, since there may be more fixed points in the system the solution found by running the mathematical software may depend on the initial parameters. And last, finding these points does not tell us anything about their stability.

However, we are going to use Lemma of The Fixed Points (Lemma 4), stated and demonstrated in Appendix 1, to show that, defining the operator G as below 3.14, we can see that when the fixed point $i = 0$ is stable, there are not more fixed points in the system, but, when $i = 0$ is unstable, G has one more fixed point and this fixed point will be stable. In this case, if we set the right conditions in the system so that our fixed point $i = 0$ is unstable, the option running a mathematical software to find a second fixed point would be enough as we know that 1.- it would be the only fixed point in the system, and 2.- it would be stable.

In order to analyse the stability of any fixed point (i_0, v_0) we study the function $G(i)$ defined as below:

$$G(i) = P\left(1 - \exp\left(\frac{-v(i)R_p}{w}\right)\right) \tag{3.14}$$

We know that a point $i = i_0$ is stable if $|\partial G/di| < 1$ when it is evaluated in i_0.

$$\frac{\partial G}{di} = P\exp\left(\frac{-vR_p}{w}\right)\frac{R_p}{w}\frac{\partial v}{di}$$

and

$$\frac{\partial v}{di} = V\exp\left(\frac{-iR_v}{w}\right)\frac{R_v}{w}$$

This gives us:

$$\frac{\partial G}{di} = \frac{VPR_pR_v}{w^2}\exp\left(-\frac{vR_p + iR_v}{w}\right)$$

Assuming that there is a stable state in the dynamical system, a point $(i_0,\ v_0)$ is a fixed point of the system if it satisfies equations 3.12 and 3.13, and it will be a stable point if

$$v_0R_p + i_0R_v > w\ln\left(\frac{VPR_pR_v}{w^2}\right) \tag{3.15}$$

and it will be unstable if

$$v_0R_p + i_0R_v < w\ln\left(\frac{VPR_pR_v}{w^2}\right) \tag{3.16}$$

In the case of $i_0 = 0$, $v_0 = 0$ we have that it is stable if

$$w^2 > PVR_vR_p \tag{3.17}$$

and unstable if

$$w^2 < PVR_vR_p$$

However, we can still get more information from G. Because the number of vectors V, number of people P, infection intensities R_p and R_v, and the size of the area w, are all constant and always positive in our system, we can say that G is continuous in $\mathbb{R}$ and $G(i) \leq P$ in $\mathbb{R}$. We also know that $G(0) = 0$ and always positive.

Looking at G' we can see that it is always positive in $\mathbb{R}$. Also that G' is continuous in $\mathbb{R}$.

$$\frac{\partial G}{di} = \underbrace{\frac{VPR_pR_v}{w^2}}_{+} \underbrace{\exp\left(-\frac{vR_p + iR_v}{w}\right)}_{+}$$

and that makes G strictly increasing in $\mathbb{R}$.

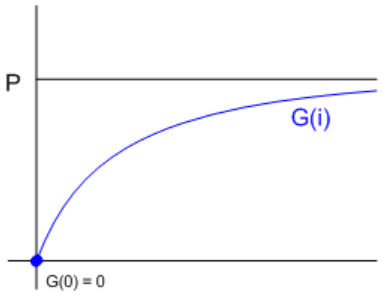


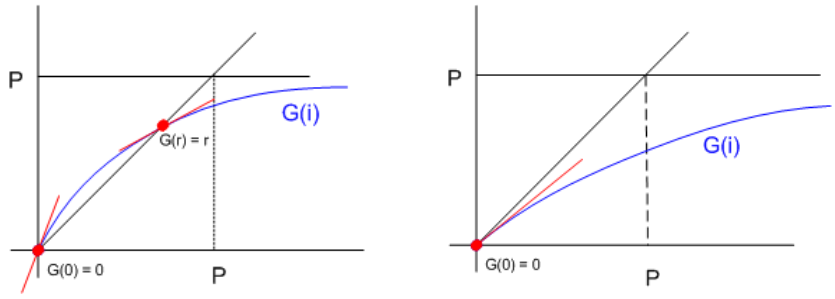
Figure 3.10: Graphical representation of a function G with $G(0) = 0$ and G is smaller than P, always positive, strictly increasing and concave downward.

Looking at the second derivative of G:

$$\frac{\partial^2 G}{di^2} = \underbrace{\frac{VPR_pR_v}{w^2}}_{+} \underbrace{\left(-\frac{R_v}{w} - \frac{R_pR_vV\exp\left(-\frac{iR_v}{w}\right)}{w^2}\right)}_{-} \underbrace{\exp\left(-\frac{vR_p + iR_v}{w}\right)}_{+}$$

we can see that it is also continuous in $\mathbb{R}$ and always negative. That makes G concave downward (see Figure 3.10).

Given the properties of G we can we can use Lemma 4 to prove that if $G'(0) < 1$ (i.e. $1 < \frac{P}{w}\frac{V}{w}R_vR_p$), then $i = 0$ is stable and there are no other fixed points in the system. On the other hand, if $G'(0) > 1$ (i.e. $1 > \frac{P}{w}\frac{V}{w}R_vR_p$), then $i = 0$ is not stable and there is one and only one more fixed point in the system, and that fixed point is stable. See Figure 3.11.



(a) $G'(0) > 1$ (b) $G'(0) < 1$

Figure 3.11: Graphical representation of a function G with $G(0) = 0$, $0 < G < P$ in $(0, P]$. If $G'(0) > 1$ then G has one more fixed point r, with $0 < r < P$ and $|G'(r)| < 1$. If $G'(0) < 1$ then G does not have any fixed point in $(0, P)$.

Having proved this, we know that the fixed points given by running the equations in the mathematical software are the only points in the system and they are stable. Because of these two reasons, the initial values given to i should not affect the result of running the equations, as long as $i \neq 0$.

3.4.5 Model 5 - Movement Between Patches with Vector-Borne Disease Dynamics

We again implement the environment as a set of homogeneous patches $\{0, \ldots, N-1\}$ with different values for the parameters w_n (area of the patch), V_n (total

number of vectors), and l_n (probability to leave the patch). In each patch the dynamics of the disease is the same as for a single patch. However, we will again implement movement of people between the patches as we did in previous cases (vectors will remain always in the same patch).

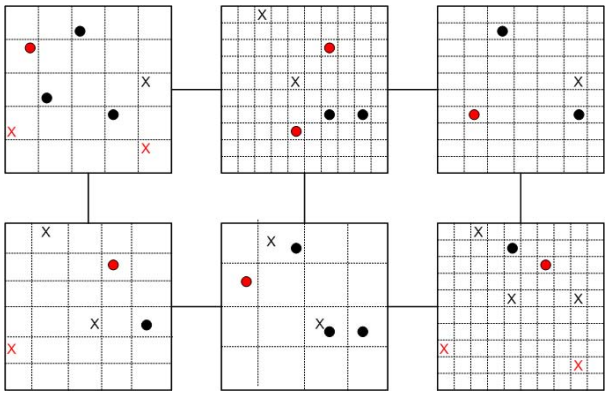


Figure 3.12: A set of patches with vector-borne disease

In order to do the mathematical analysis of this model we define two diagonal matrices $E_{v(t)}$ with entries $E_{n,n} = \exp\left(\frac{-v_n(t)}{w_n}\right)$ (probabilities for a person not to get in contact with any infected vector in patch n at time t) and $E_{i(t)}$ with entries $E_{n,n} = \exp\left(\frac{-i_n(t)}{w_n}\right)$ (probabilities for a person not to get in contact with any infected person in patch n at time t)

If there was no movement in the system the number of infected vectors in each patch of the model the model would be:

$$\vec{v}(t) = \vec{V} - E_{i(t-1)}E_{i(t-2)}\dots E_{i(t-R_v)}\vec{V}$$

At each time step each person will move with certain probability from patch to patch. Then the disease transmission dynamics take place. Having the matrix M that determines the movement of people in the model at each time step we have the number of expected infected people at time t is:

$$\vec{i}(t) = \vec{p}(t) - E_{v(t-1)}ME_{v(t-2)}M\dots E_{v(t-R_p)}M\vec{p}(t)$$

Assuming that the model has a steady-state we have:

$$\vec{v} = \vec{V} - E_i^{R_v}\vec{V} \tag{3.18}$$

$$\vec{i} = \vec{p} - [E_v M]^{R_p}\vec{p} \tag{3.19}$$

3.5 Chapter Summary

In this chapter we have presented a model of the spread of a vector-borne disease in a country taking malaria as an example. The idea of this model is to investigate how a population moving around a heterogeneous environment affects the dynamics of the disease. This model has been implemented in five steps, creating five different sub-models with increasing complexity by adding different factors to the original model in each step. Each sub-model has been simulated using RePast and analysed mathematically.

The first model focuses only in the movement of a population around a set of patches, and the disease has not been implemented in the model yet. The environment is modelled as a set of patches connected to each other forming a grid. There is only one population living in the model. Each patch i has a probability l_i for each individual to leave the patch. The people that leave each patch will move to one of its neighbouring patches with equal probability. The rules that define the movement dynamics are explained in more detail in Wection 3.3.2. The mathematical analysis says that movement of the people is determined by the topology of the grid which is given by matrix A (3.2) and the probability of leaving each patch given by matrix L (3.3) so when the model is its steady-state, the number of the people in each patch is given as an eigenvector of the matrix $M = I_N - L + AL$ where I is the identity matrix of size $N = \{$ number of patches $\}$ (see 3.4).

The second model implements a simple disease in a single patch. There is still only one population and the probability for each individual to get infected is

given by a constant d. An infected person will remain infected for R time steps unless it gets re-infected, in which case it will remain for another R time steps since the last time it got infected. The mathematical analysis for this model says that, when the model is in its steady-state, the number of infected people in the model is given by $i = P - (1-d)^R P$ where P is the total number of people in the system.

The third model implements the environment as a set of patches of different types connected to each other forming a grid. The population moves around these patches following the same rules as in the first model. The disease dynamics, which take place independently in each patch, are defined as in the second model. The mathematical analysis for this system says that the number of infected people in each patch in its steady-state is given by equation $\vec{i} = \vec{p} - [(I_N - D)M]^R \vec{p}$, where M is the matrix movement defined as above and D is a diagonal matrix which holds the probabilities d_i for a person to get infected in each patch i.

The dynamics of the fourth model are implemented in one patch only, and consist of a vector-borne disease that is transmitted between two populations. The patch is subdivided in a grid of w number of cells. At each time step each person and each mosquito gets randomly e independently located to one of the w cells. If two individuals are in the same cell we say they are in contact with each other. The disease gets transmitted from infected people to susceptible mosquitoes or from infected mosquitoes to susceptible people when they are in contact. The mathematical system that defines the dynamics of this model is given by two coupled equations 3.12 and 3.13. We have not been able to solve these equations analytically. However, we have found that there are two fixed points in the system, one will be a disease free state $i = 0,\ v = 0$ and one will imply an endemic state in the system. To find the second fixed point is the system, in which the disease remains endemic we can run these coupled equations with an initial number of infected mosquitoes different from 0. The mathematical analysis also says that, although there are two fixed points in the system, only one of them will be stable and the other one will be unstable. In fact, the fixed point $i = 0$ will be stable when the product of the density of vectors and people are low

enough (3.15) and the second fixed point (whose value is obtained by running the coupled equations) will be stable otherwise. This means that the disease in an infinite population would die out whenever the density of the populations is low enough, regardless of the initial number of infected in the system. It also means that the disease will be endemic otherwise.

The fifth model implements the same environment as the first model formed by a set of patches of different types. We now include a population of vectors V_i in each patch i. Each type of patch will have a different number of vectors. Vectors move within the patch but do not move from patch to patch. The population of people moves around these patches following the same rules as in the first model. In each patch there is a vector-borne disease defined as in the fourth model and which takes place each time step after people have moved around the patches. The mathematical analysis of this model says that the steady-state of the system is given by the coupled equations 3.18 and 3.19. One fixed point of these equations is the trivial solution $\vec{i} = 0, \vec{v} = 0$. The second fixed point of the system can be found by running the coupled equations 3.18 and 3.19 since we have not been able to solve the system analytically. We have not been able to make an analysis regarding the stability of these fixed points.

One of the first questions of this thesis is to understand how accurate our mathematical analysis is and what information we can infer from the analysis about the dynamics of the model. In the next Chapter 4.1 we present several experiments and contrast the results of the simulations with the outcomes of the mathematical analysis of each model. The experiments are normally set with a large number of individuals in them, as we know that the mathematical model reflects the behavior of an infinite population. However, sometimes we will take it to an extreme and run the experiment with only one individual and see that we can still infer information from the mathematical analysis.

Chapter 4

Experiments and Results

4.1 Model 1 - Experiments and Results

In this section we present the experiments made for the study of the mathematical result given by the first mathematical model given by $\vec{p} = M\vec{p}$ (3.5). This mathematical result says that the expected distribution of people in the patches in its steady-state is:

$$\vec{p} = (p_0, \ldots, p_{N-1})$$

$$\text{where } p_n = \lambda \frac{c_n}{l_n} \text{ and } \sum_{n=0}^{N-1} p_n = P$$

I remind the reader about the parameters used for this model:

P	total number of people in the model
p_n	number of people in patch n
$\vec{p}$	vector distribution of people in the model
l_n	probability for a person to leave patch n
c_n	number of connections of patch n

As explained in the previous chapter, the parameter l_n depends on the type of patch n is classified under, and the parameter c_n depends on the topology of the environment. For these experiments I create a 3x3 grid structure environment and give the patches the classification type shown in Figure 4.1:

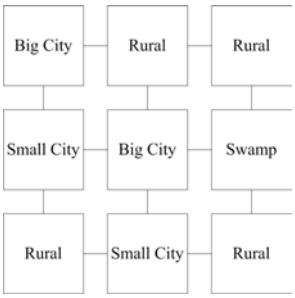


Figure 4.1: Structure of the environment in the vector-borne disease models.

In this set of experiments the parameters l_n will be:

Patch *type*	Value of l_n
Big city	0.2
Small city	0.7
Rural	0.4
Swamp	1

Looking at Figure 4.1 we see that the set of probabilities of leaving each patch is:

$$\{l_0, \ldots, l_8\} = \{0.2, 0.7, 0.7, 0.4, 0.2, 1, 0.7, 0.4, 0.7\}$$

and the set of numbers of connections of each patch is:

$$\{c_0, \ldots, c_8\} = \{2, 3, 2, 3, 4, 3, 2, 3, 2\}$$

Following the same steps as in Example 1 from previous chapter we know that the distribution of people in the simulation should be:

$$\vec{p} \approx P * (0.1643, 0.0704, 0.0469, 0.1232, 0.3286, 0.0493, 0.0469, 0.1232, 0.0469)$$

We run different sets of experiments that take different number of people in the simulation.

In the first set of experiments we set P to be $10,000$ people. Following the mathematical result we should expect to have the following distribution in the steady-state:

$$\vec{p} \approx (1643, 704, 469, 1232, 3286, 493, 469, 1232, 469)$$

For this experiment we run the simulation 100 times for 50 time-steps each time.

In the Figure (4.2) we show a graph with the number of people in each patch as we run the simulation. In the x-axis we have the 9 different patches in different colours and the way they change over time. To create this graph we have calculated the average population in each patch at each time step over 100 runs. We can see in the graph that the average distribution of people in each patch finds its steady-state very quickly. In the graph we can also compare the average number of people given by the experiments and the expected number of people given by the mathematical result 3.5. As we can see, the mathematical result is a very accurate approximation of what happens in the simulation.

If we run the simulation only once, for 50 time steps, and we compare the results with the mathematical approximation, we get the graph shown in Figure 4.3. We can see that the result is still very accurate.

If we change the number of people to only 1, as we expect the mathematical approximation is not accurate any more. The mathematical result says that the distribution of people will have a steady-state approximated to:

$$\vec{p} \approx (0.1643, 0.0704, 0.0469, 0.1232, 0.3286, 0.0493, 0.0469, 0.1232, 0.0469)$$

which seems very unfeasible since we can only have either 1 or 0 people in each patch at any time step. However, we can interpret the solution of the fixed point given by 3.5 in a different way. If we run the simulation for 10,000 time steps

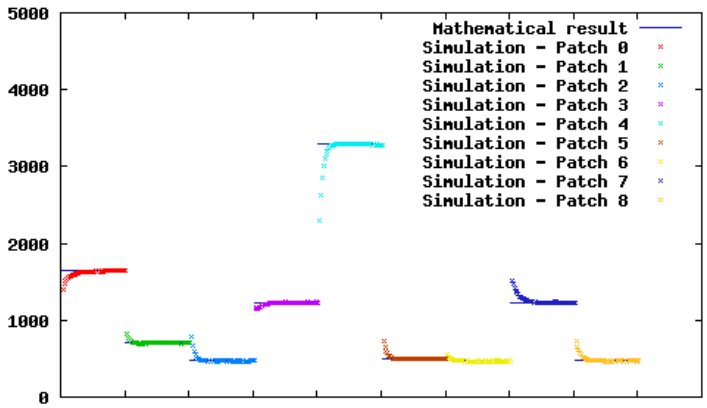


Figure 4.2: Model 1 - Studying the steady state of the distribution of people with 10,000 people in the system. View of the average of 100 runs.

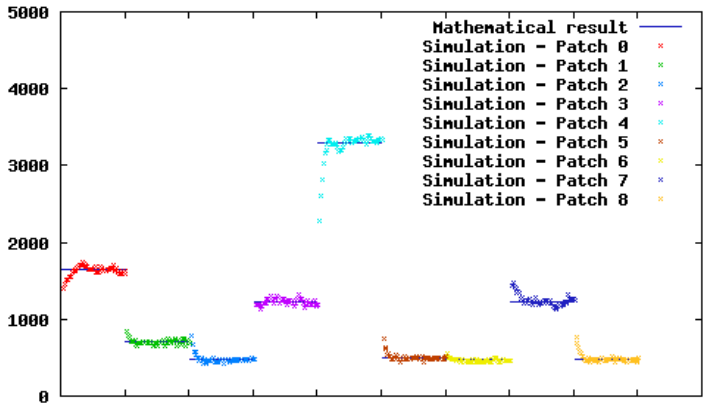


Figure 4.3: Model 1 - Studying the steady state of the distribution of people with 10,000 people in the system. View of 1 single run.

and we estimate the probability to find the person in each patch, we can see that the mathematical result 3.5 is a very good approximation of this estimated probability. This result is also shown in Figure 4.4.

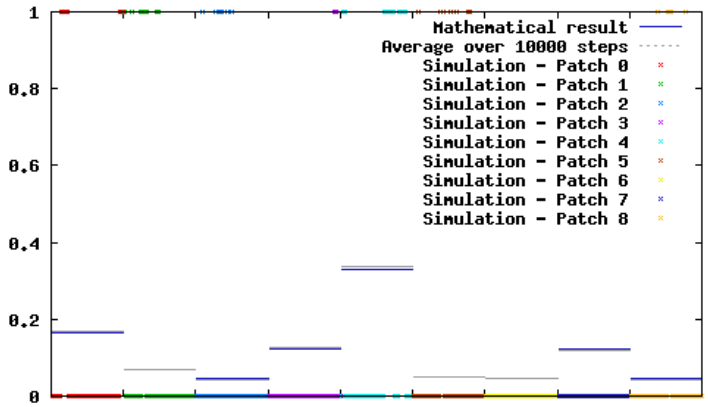


Figure 4.4: Model 1 - Studying the average time that a person spends in each patch when there is only 1 person in the system. View of the average in time of 1 single run.

4.2 Model 2 - Experiments and Results

For the second model we have the result 3.7, which gives us the number of infected people in the steady-state. Note that in this model we are working with only one patch. The parameters used for this model are:

d	probability for a person to get infected
R	number of time-steps a person remains infected
s	number of susceptible people at the steady-state
i	number of infected people at the steady-state

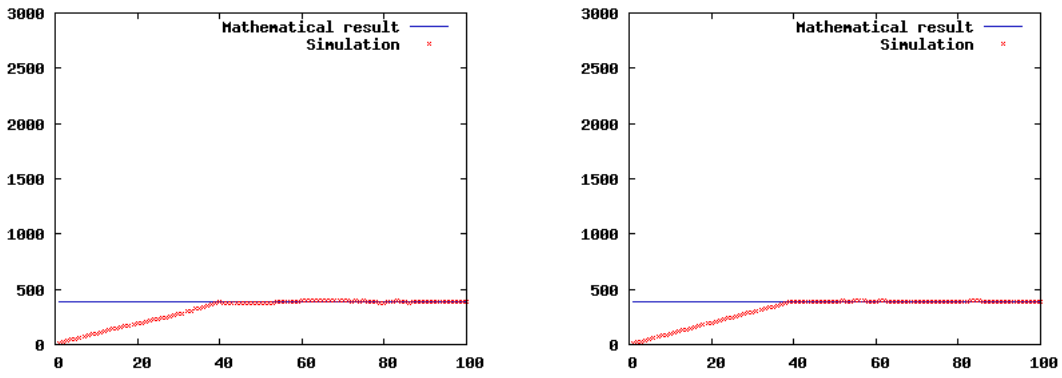
The parameter values chosen for the experiments are $d = 0.001$ and $R = 40$. The mathematical result 3.7 says that the expected number of infected people

should be:

$$i = (1 - d)^R * P \approx 0.0392 * P$$

We again run different sets of experiments that take different numbers of people. For all different values of P we run 100 simulations for 100 time-steps each.

When we have 10,000 people the mathematical result says we should expect $3,920$ people to be infected. In Figure 4.5 we show two graphs. The first one shows the number of infected people in one single run over 100 time-steps and compare it with the mathematical result. In the second graph (4.5 b) we have the average number of infected people in 100 time-steps. The average is calculated over 100 different runs. We can see that in both cases the mathematical result is very accurate.



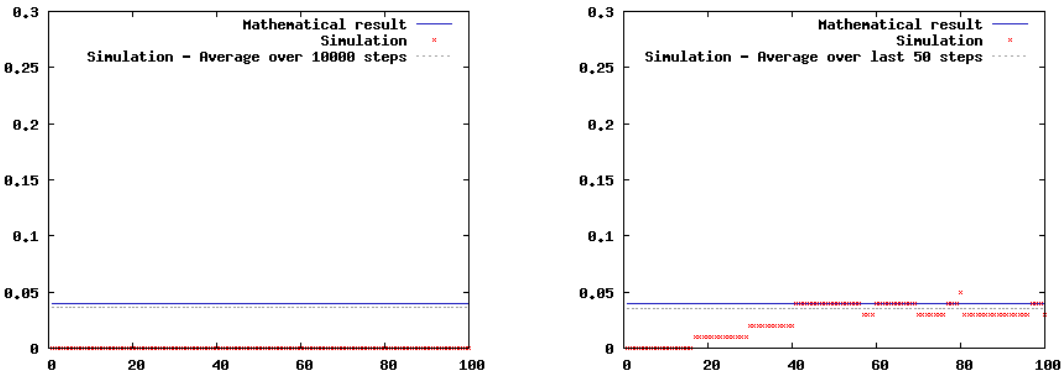
(a) 1 Run

(b) Average of 100 runs

Figure 4.5: Model 2 - Studying the steady state of the number of infected people in the system. Parameter values are transmission probability $d = 0.001$, infection time $R = 40$, and number of people $P = 10,000$.

We can observe in Figure 4.6(a) that, when we run the simulation with only one person in the system, the mathematical result is not telling us anything about the steady-state of the simulation. In fact, in most of the individual runs the number of infected people is 0 in the first 100 steps of the simulation. We also

know there will not be such a steady-state since the possible number of infected people is 0 or 1. However, the mathematical result seems to be a very accurate prediction of the probability for the person to be infected. In Figure 4.6(a) we can also compare the mathematical result (in blue) with the average number of infected people in a run over 10,000 time-steps (in grey). In Figure 4.6(b) we can also see the average number of infected people over 100 runs (in red) and the average of those values over the last 50 time-steps (in grey). We can compare how the average over time is very close to the mathematical result (in blue).



(a) 1 Run

(b) Average of 100runs

Figure 4.6: Model 2 - Studying the average time a person spends infected when there is only one person in the system. Average over time for 1 run and 100 runs. Parameter values are transmission probability $d = 0.001$, infection time $R = 40$, and number of people $P = 10,000$.

4.3 Model 3 - Experiments and Results

In this model the environment is formed as a set of patches and the people move between them. The structure of the environment is the same as we used for Model 1 (4.1). The mathematical result of this model is:

$$\begin{aligned} \vec{s} &= [(I_N - D)M]^R \vec{p} \\ \vec{i} &= \vec{p} - [(I_N - D)M]^R \vec{p} \end{aligned} \tag{4.1}$$

The new parameters used in this model are:

d_n	probability for a person to get infected in patch n.
D	diagonal matrix with entries d_n

The probability for a person to get infected in each patch n depends on the classification of that patch. In this set of experiments the parameters d_n are:

Patch *type*	Value of d_n
Big city	0.0001
Small city	0.001
Rural	0.01
Swamp	0.1

Matrix M is calculated as in Example 1, using the values l_n and c_n. The values for these parameters are the same as used in the experiments for Model 1.

In order to calculate the fixed point of the system 3.11, we run a program with these equations with the parameter values being the same as the ones given in the simulations. As a result we find that a continuous approximation of the steady-state of our simulation should be:

$$P * (0.1643, 0.0704, 0.0469, 0.1232, 0.3286, 0.0493, 0.0469, 0.1232, 0.0469)$$

As in the previous cases we run two experiments, one with 1 person in the system, and another one with 10,000 people. In both cases we run 100 simulations for 100 time-steps each.

For the simulation using 10,000 people, as we can observe in Figures 4.7 and 4.8, the mathematical model is again very accurate for both the results of averaging 100 runs and the results of one single run.

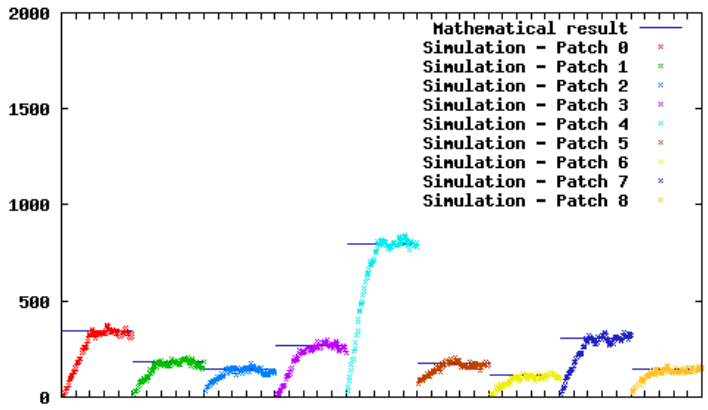


Figure 4.7: Model 3 - Studying the steady state of the infections with 10,000 people in the system. View of 1 single run.

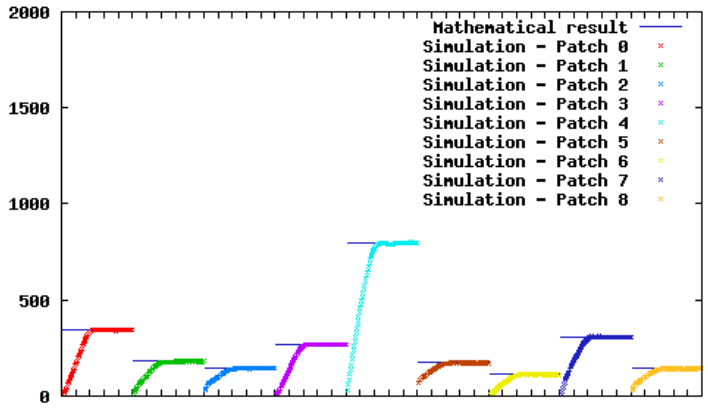


Figure 4.8: Model 3 - Studying the steady state of the infections with 10,000 people in the system. View of the average of 100 runs.

When we run the simulation with 1 person we again can use the mathematical result as the probability to find an infected person at a given patch, but it does not tell us anything about the steady-state of the simulation, as shown in Figure 4.9.

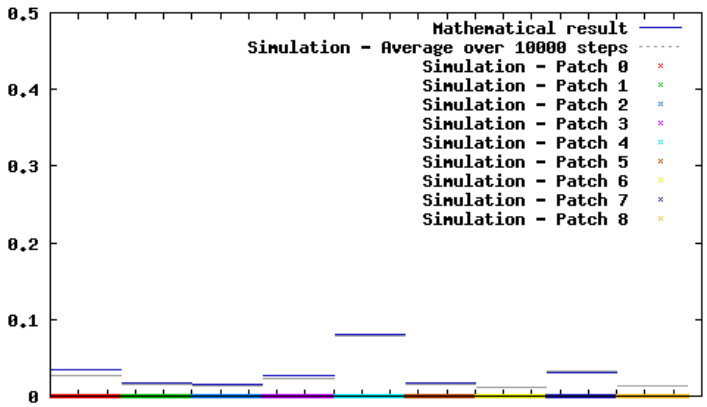


Figure 4.9: Model 1 - Studying the average time that a person spends infected in each patch when there is only 1 person in the system. View of the average in time of the average of 100 runs.

4.4 Model 4 - Experiments and Results

In this model we introduced the vectors population and the dynamics in the spread of the disease changed. We again work with only one patch.

The first mathematical result for this model is given by equations:

$$\begin{aligned} i &= P\left(1 - exp\left(\frac{-vR_p}{w}\right)\right) \\ v &= V\left(1 - exp\left(\frac{-iR_v}{w}\right)\right) \end{aligned}$$

I remind the reader of the new parameters used in this model:

P	number of people in the patch
p	number of infected people in the patch
V	number of vectors in the patch
v	number of infected vectors in the patch
R_v	number of time-steps a vector remains infected
R_p	number of time-steps a person remains infected
w	area size of the patch (number of cells)

We run several experiments exploring the different aspects of the behaviour of the model and the mathematical result. In the first experiment we will run the simulation for 2,000 time-steps with the following parameter values:

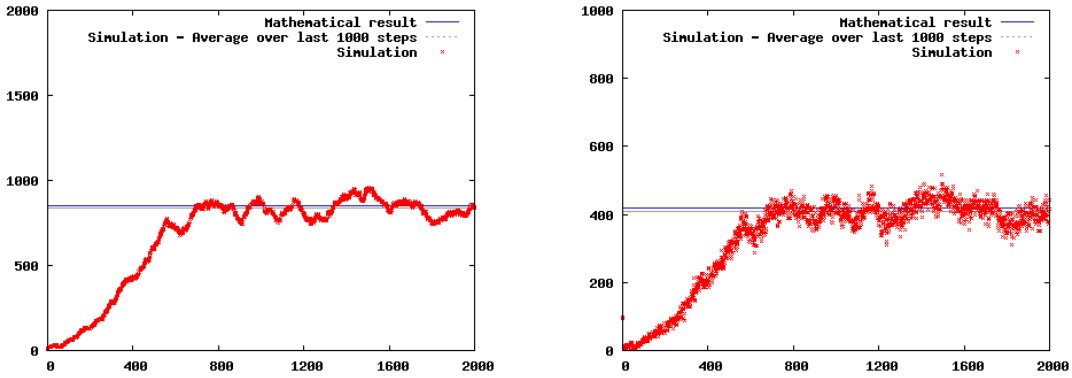
$$\begin{aligned} P &= 2{,}500 \\ V &= 10{,}000 \\ R_p &= 40 \\ R_v &= 2 \\ w &= 40{,}000 \text{ units} \end{aligned}$$

For the disease dynamics to take place in the simulation, we need to start with a number of infected agents. We have chosen to leave all people susceptible and 1% of the vector population infected. In order to get the values of the mathematical approximation given by 3.12 and 3.13 we run the two equations in a program until it settles to a steady-state. For the parameter values above we find that the mathematical approximation is:

$$\begin{aligned} \text{i} &= 854.5392 \\ \text{v} &= 418.2702 \end{aligned}$$

In Figure 4.10 we can see how the number of infected people (a) and vectors (b) vary as the simulation runs. Although they are close to the value given by the mathematical result the are large oscillations as time moves. The grey lines in the graphs are the average number of infections in the last 1,000 time-steps. We can see that the approximation is very accurate.

In the next experiment we will vary the parameter w (area size of the patch) leaving the rest of the parameters fixed. So the parameter values will be:



(a) Infected people

(b) Infected vectors

Figure 4.10: Model 4 - Number of infected people and vectors as we run the simulation for 2,000 time-steps. The parameter values of the simulation are: P=2,500, V=10,000, R_p=40, R_v=2 and w = 40,000. The initial number of infected vectors is 100.

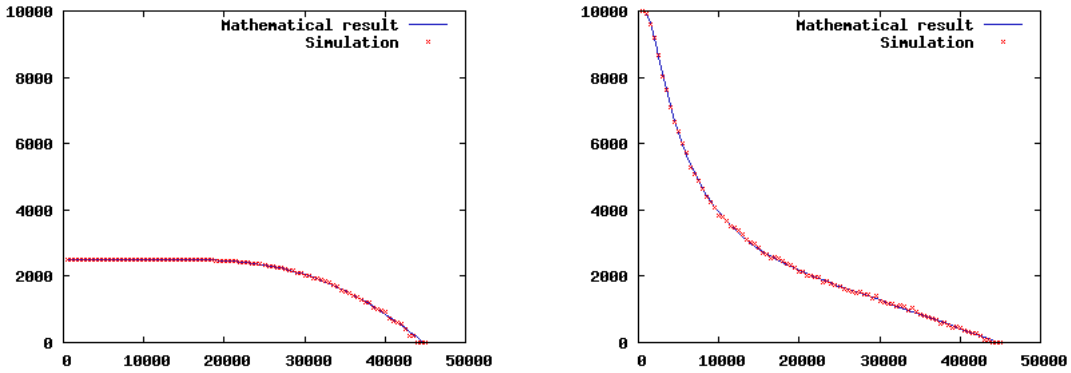
$$\begin{aligned} P &= 2{,}500 \\ V &= 10{,}000 \\ w &= \text{will vary from 0 to 40,000 units (with 500 unit intervals)} \end{aligned}$$

For each value of w we will run the simulation once for 2,000 time-steps and record the value of infected people and vectors at the last time-step. In Figure 4.11 we can compare the values given in this experiment with the mathematical approximation. As we can see in Figure 4.11 the values given by the simulation are very close to our mathematical estimation.

If for each value of w we run the simulation 100 times and calculate the average, the results we get are even closer to the mathematical result, as expected.

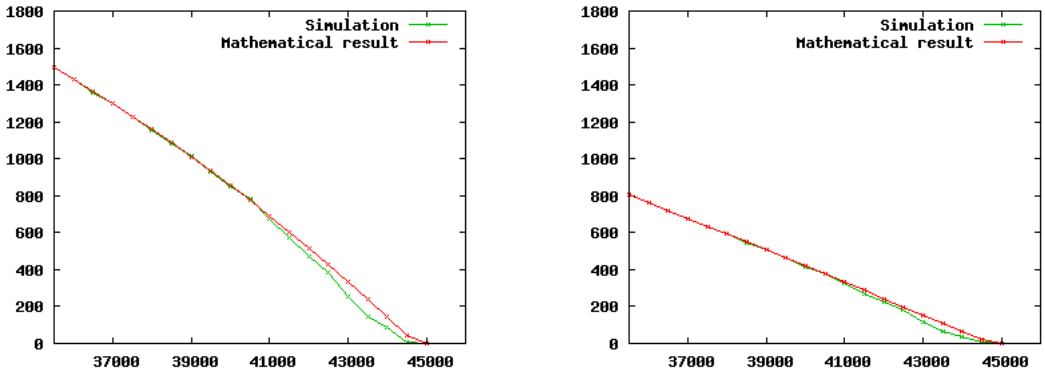
However, if we zoom in on the area where the number of infections are getting close to 0 we can see that our results differ from the mathematical result, as the simulation gives lower values than the mathematical result. (Figure 4.12).

The mathematical result 3.17 says that when $w^2 < PVR_vR_p$ the fixed point



(a) Infected people (b) Infected vectors

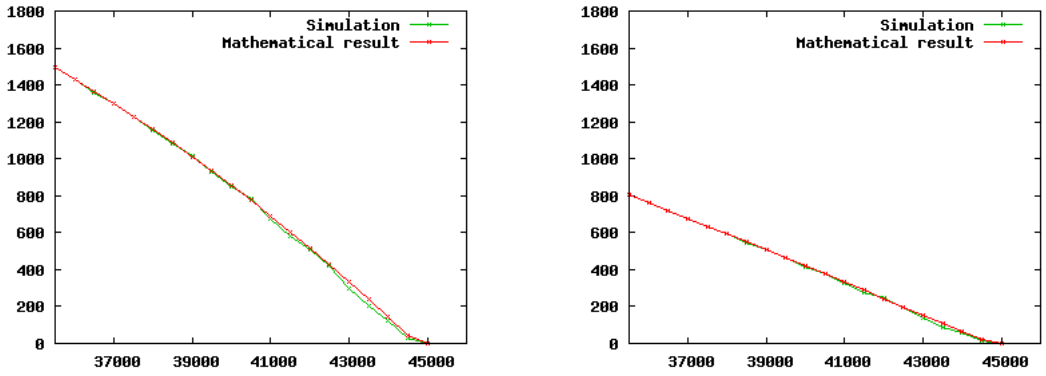
Figure 4.11: Model 4 - Steady-state of the infected people and vectors as the area size of the patch changes value w. The graph represents the average steady-state of 100 different runs, including those who get into the absorbing state $i = 0, v = 0$.



(a) Infected people (b) Infected vectors

Figure 4.12: Zoom of Figure 4.11 for the largest values of the area size w.

$v_0 = 0, i_0 = 0$ is unstable, and the other one is stable. However, that works well with a continuous number of agents. The simulation has a discrete number of agents and they change their state from susceptible to infected randomly (with certain probability). These factors make the simulation so there is always certain probability that none of the agents get infected and therefore get to the point where $v_0 = 0, i_0 = 0$. This is an absorbing state in the system, i.e. once they get in that state, they cannot get in any other state. It is obvious that the less number of infected people there is in the system, the higher the probability to get in that state.



(a) Infected people

(b) Infected vectors

Figure 4.13: Model 4 - Steady-state of infections in the system as w changes its value. This graph represents the average of only those runs that have not fallen in the absorbing state $v = 0, i = 0$.

In Figure 4.13 we have drawn the same graph as before but considering only those runs which did not fall into the absorbing state $v_0 = 0, i_0 = 0$. We can see that the average value of those runs and the value given by the mathematical result are considerable closer than the previous case.

Figure 4.14 shows the number of steps that a simulation has to run before it reaches the absorbing state $i = 0, v = 0$. In the experiment we left fixed the number of mosquitoes and the parameters R_v and R_p. We incremented the number of people in the model and the area size w, so that the density of people

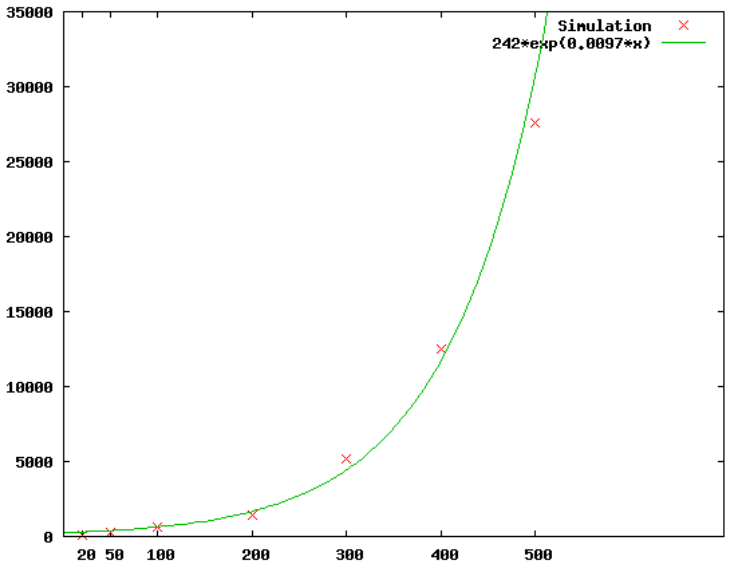


Figure 4.14: Number of steps before simulation reaches the absorbing state for different number of people and area sizes, leaving the density fixed

in the patch remains fixed. We can see that the time for the disease to disappear increases exponentially with the number of people in the model.

4.5 Chapter Summary

The first model focuses only in the movement of a population around a set of patches. The mathematical analysis says that, when the model is in its steady-state, the distribution of the people around the patches is given as an eigenvector of a matrix M (see 3.4). The results of the experiments presented in section 4.1 show that this mathematical estimation is really good when there is a large number of people in the model. However, when the number of people in the model is not large enough, the mathematical model is not a good estimator of the distribution but it can be understood as the average time that a person remains in each patch.

The second model implements a simple disease in a single patch. There is still only one population and the probability to get infected is given by a constant d. The mathematical analysis for this model says that when the model is in its steady-state the number of infected people in the model is given by equation 3.7. The results of the experiments presented in section 4.2 show that this mathematical model is also very accurate in estimating the size of the disease when there is a large number of people in the system. When there are only few people in the system the mathematical model estimates the average time that each person remains in its infected state.

The third model implements the environment as a set of patches of different types connected to each other forming a grid. The population moves around these patches following the same rules as in the first model. The disease dynamics, which take place independently in each patch, are defined as in the second model. The mathematical analysis for this system says that the distribution of infected people in the model in its steady-state is given by equation 3.11. The results of the experiment show that this mathematical model is also very accurate when there are a large number of people and, when there are only a few people, it estimates the probability to find an infected person in a patch at any time step.

The fourth model represents a vector-borne disease that is transmitted between two populations in one single patch. The mathematical system that defines the dynamics of the disease is given by two coupled equations (3.12 and 3.13) that we have not been able to solve analytically. We have found that there are two fixed points in the system, one will be a disease free state $i = 0,\ v = 0$ and the second fixed point implies an endemic state in the system. To find the second fixed point of the system, in which the disease remains endemic in both populations, we run the coupled equations with an initial number of infected mosquitoes of 1%. The simulation results show that the results of the mathematical model are a good estimate of the steady-state of the system when the population is large enough. However, we have not run simulations with low number of vectors or people in the model. The mathematical analysis also says that, although there are two fixed points in the system, only one of them will be stable and the

other one will be unstable. In fact, the fixed point $i = 0$ will be stable when the density of both vectors and people are low enough (3.15) and the second fixed point (whose value is obtained by running the coupled equations) will be stable otherwise. This means that the disease in an infinite population would die out whenever the density of the populations is low enough, regardless of the initial state of the system. It also means that the disease will be endemic and persist (for an indeterminate amount of time) otherwise. In the results presented in Section 4.4 we can see that the mathematical estimation of the stability of these points is again very accurate. We start having problems only when the expected size of the infection is quite small and the simulation gets into the absorbing state $i = 0, v = 0$. But even taking into consideration those runs, the prediction of the mathematical model seems very accurate.

4.6 Conclusions

In this chapter we have successfully implemented a model of vector-borne disease in a country and we are able to understand to certain extent the dynamics of the system under study. We have successfully undertaken a mathematical analysis of the different models adapting the Heuristic Random Search framework to our models. This approach has helped us to create mathematical abstractions of each of the sub-models, formulating very precisely the properties of the system and having rigourous proofs about these properties. We have, in all cases, been able to make an accurate prediction of the emergent behaviour of the disease dynamics. However, we have also observed that the complexity of the mathematical models increases rapidly with the complexity of the model. In the later models 3.4.4 and 3.4.5 we have been unable to solve the motion equations analytically. In the last model 3.4.5 we have been unable to carry out a mathematical analysis of the stability of the fixed points.

During the process of creating the simulations and the mathematical models in parallel we have found that, having two different representations of the system,

one as an agent based simulation and another one as a mathematical model, has been crucial for the understanding of the behavior of the model. For example, the mathematical equation that defines the movement of the population around the patches was based on the results of experiments run in the simulation. Also, the realisation of being able to interpret some mathematical models as the average amount of time that an individual spends in a certain state came from observation of the simulation.

It is also clear how the strategy of carrying out an incremental implementation with increasing complexity of the models has great advantages for the understanding of the more complex models. For example, the fact of analysing the movement of people around the patches before introducing any sort of disease in the model was crucial for the definition of the Matrix of Movement in the first mathematical model, which was used for later models.

Part III

SIS Dynamics on a Network

Chapter 5

SIS Dynamics on a Network

5.1 Overview

Many of the social dynamical systems studied have been modelled as homogeneous systems where all actors interact with each other with the same probability. This homogeneity in the interactions is referred to as "perfect mixing". As we discussed in the background section, there are different ways of introducing heterogeneity in the interactions between the actors of our models. Among others there is the metapopulation approach introduced in the Background Section and used in the models presented in the first part of this thesis. Another approach would be by representing the interaction between individuals through a social network.

During the last few years it has been frequently discussed that many social and natural systems can be represented as a graph where nodes represent individuals, and links between these nodes represent an interaction between these individuals. In this way we represent our population as a network and we can study a given phenomena on the population as processes running on top of this network.

5. SIS DYNAMICS ON A NETWORK

There has been a large amount of research studying many properties of social networks such as degree distribution, clustering, size of the components, small world effect, etc. All these different properties of networks, which have been presented in the background section, can have a great impact in the dynamics of the phenomena studied in our population (21). From all these properties we are going to pay most attention to the degree distribution of our networks.

The process we are going to study in our networks is called SIS model. SIS stands for susceptible-infected-susceptible and it is a well known epidemiological model. The nodes of the network represent people and can be in either of these two states, susceptible or infected. The edges of the network represent some sort of relation between the individuals of the population that make possible the transmission of the disease between two people, so the infected nodes may transmit the infection to other nodes that are in contact with them through an edge.

Although the most common name for the process we are studying is called the SIS model, and it is used for epidemiological studies, we are not going to state the process in terms of infections but in terms of colours. The process is stated as follows:

Let us suppose that we have a set of N nodes connected to each other. These nodes may be in red or white state. At each time step two things happen:

1 Red nodes try to transmit their state to their white neighbours and convert them into red nodes with probability p of success.

2 Red nodes spontaneously turn white with probability b.

The reason to state the problem in terms of colours is because this behavioural model can be used, and has been used, for the study of different phenomena such as epidemiology, Extinction and Colonisation dynamics of species in islands or spread of computer viruses in computer networks.

For the study of Extinction and Colonisation of species in islands the idea was to model a population of a species in a fragmented space, to model a population of populations. This model is somehow the beginning of Metapopulation Theory and it was introduced by R. Levins with the intention of understanding the processes of extinction and re-population of certain species in islands. His original model assumes that the space is discrete formed by a set of patches which have all the same area and qualities. The dynamics of populations in different patches are independent from each other. The individuals will migrate from patch to patch, but the migration rate is low enough so the movement of the individuals does not affect the dynamics of the populations, but yet it creates a balance between extinction and colonisation. Again, the Levin's model is also a case of perfect mixing model in which each patch is connected with every other patch, while in the model we present the dynamics will take place in different structured networks. In our model the state red would correspond to *populated* and white would correspond to *extinct*. This was the most simple model of studying populations in a fragmented space and it was the beginning of the field of meta-population theory (33).

This model can also be used for the study of the spread of computer viruses in a computer network (47; 102). A computer virus is a computer program that can copy itself and infect a computer without the permission or knowledge of the owner. A computer virus can only spread from one computer to another (in some form of executable code) when its host is taken to the target computer via the Internet, or via a removable medium such as a floppy disk, CD, DVD, or USB drive. Viruses can increase their chances of spreading to other computers by infecting files on a network file system or a file system that is accessed by another computer (104).

5.2 General Methodology and Approach

The question we try to answer in this model is: what is the steady-state distribution? How many red or white nodes there are after a long period?

We have studied different graph topologies taking the approach of adding more complexity to the model as we have understood the more basic models. The models that we have studied are named 1.-Symmetric Graph, 2.-Erdös-Rényi Random Graph and 3.-Scale-Free network. These models are introduced below.

The rigourous definition of *Symmetric Graph* is as follows: A symmetric graph is a set of vertex V and a set of edges $E \subseteq (V, V)$ in which for any pair of vertices x and y there exists a bijection $f : V \to V$ such that 1.- $f(x) = y$, and 2.- if (u, v) is an edge, $(f(u), f(v))$ is an edge. We can get a more intuitive understanding of this definition by looking as some examples in Figure 5.1.

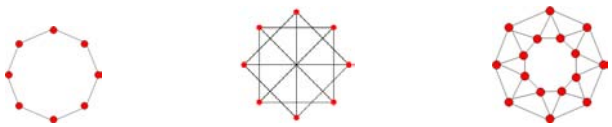


Figure 5.1: Symmetric Graphs with 2, 3 and 4 edges per node respectively.

The reason to start with this type of topology is no more than the simplicity of it. We are aware of the lack of realism in this network; it still has some ingredient of homogeneity as it is assumed that all agents of the model interact to exactly the same number of other agents. But we have, somehow, restricted the number of interactions to a local neighbourhood.

The next two models we work with are random graphs models. By random graphs we mean graphs that are generated by some random process. This random process does not have, in principle, any constraint in any properties of the

networks other than the way we generate them. However, we will see that the random process we use has a direct effect in the degree distribution of the networks built.

The first model of random graph we used is called *Erdös-Rényi Random Graph* model (also referred to as ER-Random Graph model), and it is built as follows: We first create the N nodes that will belong to our graph. On the side we consider the set $\mathcal{E}$ of all possible edges between those N nodes, as if the graph was fully connected. We are now going to choose only some edges from the set $\mathcal{E}$ to include in our graph. We choose them by creating a random process that includes each possible edge of $\mathcal{E}$ independently with probability d. All the edges from $\mathcal{E}$ chosen by this process will be the edges of my graph. The probability d is called *density* of the graph, and it measures the number of edges that we expect to have in the graph compared with the total number of possible edges so, if $d = 1$, we obtain a fully connected network and, if $d = 0$, we would get N independent nodes with no connections between them. This random process produces a graph with a Poisson distribution that is well approximated by the a normal degree distribution with mean $\mu = (N-1)d$ and variance $\sigma^2 = (N-1)d(1-d)$. For example, let's say that we want to recreate an ER-Random Graph of 10,001 nodes with density $d = 0.05$, we then would have a random graph whose degree distribution can be approximated by a normal distribution with mean 500 (see Figure 5.2).

The second model of random graph we use is called *Scale-Free Network*. The network models presented so far are very interesting from the theoretical point of view and they bring some complexity and heterogeneity in the model. Yet, they are not a good representation of social networks, in fact, they are not a good representation of most of the networks observed in nature. Most networks seem to have more sophisticated properties than the previous models, and one of them is the degree distribution. The observed degree distribution in most networks found in nature is what it is called power-law distribution. That a network has the property of power-law distribution means the following: If we define the probability that a given node of the network has k number of edges, this probability would be of the form: $d_k = ck^{-\gamma}$ with c and γ constants .

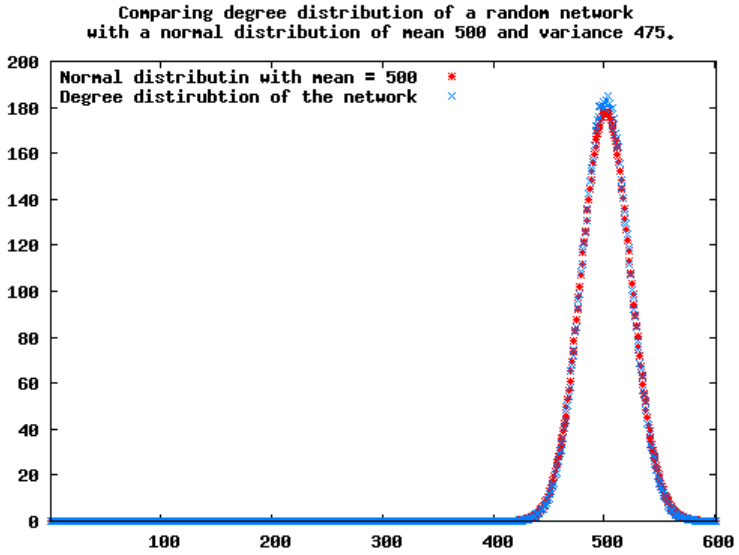


Figure 5.2: Degree distribution of a randomly generated graph of 10,001 nodes and density 0.05

Networks with power-law degree distribution are called scale-free networks. To randomly generate a scale-free network, we use the mostly widely accepted generative model created by Barabási and Albert, (15). The random process they use in the creation of the graph is called *rich get richer* or *preferential attachment* rule. It works as follows: Let's say that we want to generate a network of N nodes. First we create two nodes with one edge between them. Then we create one more node which will be attached to one of the existing nodes. The existing node it will be attached to is chosen at random, however, the probability for an existing node to be chosen is proportional to its degree, so the more edges an existing node has, the more probability to be chosen. This process will be repeated until we have all N nodes in the graph. In this first step the two existing nodes have both degree one, so they have equal probability to be chosen, but, as we add more nodes and edges to the model, the degree distribution of the nodes will follow a power-law distribution. When we generate the network we could also choose to add more than one edge per new node, for example, each new node may have initially 7 different neighbouring nodes. As long as the initial number of edges

per new node is determined at the beginning, generating a network following this process will always create a scale-free network.

For the mathematical analysis of the dynamic system in the different graphs we use an approach similar to what we did in our previous chapters, only that, instead of grouping the population depending on its location, we group the population according to the number of connections a person (represented as a node) has on the model. Having the local interaction rules of the system we try to define the general rules of the model. When we have the equations that describe the general behaviour, we look for the fixed points and, if we find them, we study the stability of this fixed points. Finally, we run the appropriate experiments to check if the fixed points given by our mathematical analysis are shown in the experiments.

5.3 Notation

$\mathbf{r(t)}$	probability for a node in the network to be red at time t,
$\mathbf{r}$	probability for a node in the network to be red when the system has reached its asymptotic value (assuming it exists),
$\mathbf{b}$	probability for a red node to spontaneously get white,
$\mathbf{p}$	probability to transmit the colour red to a white neighbour,
$\mathbf{m}$	$= \frac{b}{p}$,

5.4 Symmetric Graph

5.4.1 Simulation

The simulation creates a 2-dimensional lattice with form of a torus. The main parameters that determine the size of this lattice are the width and hight of this lattice and the depth of the connections. The nodes are connected to other nodes

that are next above, below, left, and right of our node. The depth parameter will tell us how many nodes will be connected to my original node in the four respective directions. For example, in Figure 5.3 we have a node in blue that is connected to 2 other nodes in each direction. This would be the case in a lattice of depth 2. To form a torus, the nodes of the top (the first row) are connected to the nodes of the bottom (the last row), and the nodes of the left column are connected to the nodes of the right column, with the same depth as the other nodes, (see Figure 5.4). In a torus all nodes have exactly the same number of neighbours and there are no borders in the lattice.

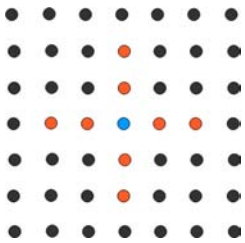


Figure 5.3: Graphical representation of depth in a 2-dimensional lattice. In this case the depth is 2.

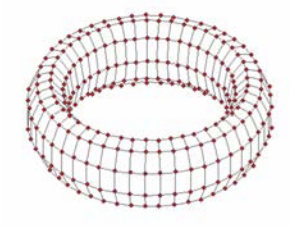


Figure 5.4: 2-dimensional lattice with form of a torus.

We set part of the nodes to be red and part of them to be white. We decide the initial number of red nodes at the start. The dynamics of the simulation work as follows. For each node in the lattice we check its state; if it is white we get its neighbouring nodes and, one by one, we allow the transmission of the colour with probability p. If one of these neighbours transmits its colour to our node, then our node will be red the next time step. If the node is red, it goes back to white with probability b. The states of the nodes get updated in a synchronous

manner, so they all get updated at the same time once the dynamics have taken place in parallel in all nodes. This forces the nodes to remain at least 1 time step in the same state, either red or white.

The simulation does not have a visual representation of this process. Each time we run the simulation we write in a file the state of the simulation in terms of the total number of red and white nodes in the network.

5.4.2 Mathematical Model

Let us suppose that each node on our symmetric graph has exactly D edges. In this topology we could mathematically describe the number of red nodes at time $t+1$ as:

$$r(t+1) = r(t) - br(t) + (1 - r(t))(1 - (1-p)^{r(t)\mathrm{D}}) \tag{5.1}$$

where $r(t)$ is the proportion of red nodes in the system at time t and D is the degree or number of edges of each node. The proportion of white nodes at time t would be $w(t) = 1 - r(t)$.

Let us find the equality 5.1 step by step. The proportion of red nodes at time $t+1$ will be $r(t+1) = \mathrm{A} - \mathrm{B} + \mathrm{C}$ where:

- A := Total proportion of red nodes at time t.
- B := Proportion of nodes that turn from red to white at time t.
- C := Proportion of nodes that turn from white to red at time t.

We know that $\mathrm{A} = r(t)$ and $\mathrm{B} = b{*}r(t)$. We estimate C as the total proportion of white nodes at time t, times the probability for each of them to turn from white to red. So let $\mathrm{C} = \mathrm{E} * \mathrm{F}$ where:

- E := Total proportion of white nodes at time t, so $\mathrm{E} = 1 - r(t)$.
- F := Probability for a white node to turn to red, i.e. the probability that a white node gets the red colour transmitted by at least one of its red neighbours.

We know that the probability for a white node to get the red colour transmitted depends on the number of red neighbours it has. We know that each node has D neighbours, so using a mean-field like approximation we can estimate that every node has $\mathrm{D} * r(t)$ red neighbours. We can also say that $\mathrm{F} = 1 - \mathrm{G}$ where G is the probability that a node does not get transmitted the red colour by any of its neighbours. The probability that a node does not get transmitted the red colour by a given red neighbour is $1 - p$, so the probability that a node does not get transmitted the red colour by any of its $\mathrm{D} * r(t)$ red neighbours is $\mathrm{G} = (1-p)^{\mathrm{D}*r(t)}$. So we get $\mathrm{F} = 1 - (1-p)^{\mathrm{D}*r(t)}$. Therefore

$$r(t+1) = \mathrm{A} - \mathrm{B} + \mathrm{C}$$

substituting $\mathrm{A} = r(t)$, $\mathrm{B} = b * r(t)$ and $\mathrm{C} = \mathrm{D} * \mathrm{E}$:

$$r(t+1) = r(t) - br(t) + \mathrm{D} * \mathrm{F}$$

and finally substituting $D = 1 - r(t)$ and $\mathrm{F} = 1 - (1-p)^{\mathrm{D}*r(t)}$:

$$r(t+1) = r(t) - br(t) + (1 - r(t))(1 - (1-p)^{\mathrm{D}*r(t)})$$

We now proceed to study the fixed points of the system and their stability. Let r be the asymptotic value of $r(t)$ as t goes to the infinity.

$$r = r - br + (1-r)(1 - (1-p)^{rD}$$

The fixed points of the system described are given by equation 5.1:

$$br = (1-r)(1 - (1-p)^{rD}) \tag{5.2}$$

Looking at this equation, we can see that $r_1 = 0$ is a fixed point in the system. In order to find any other fixed point in the system, we use an approximation to simplify the analysis. The approximation we will use is:

$$1 - (1-p)^{rD} \approx prD \tag{5.3}$$

which would give us one more fixed point:

$$r_2 = 1 - \frac{b}{Dp} \tag{5.4}$$

The use of this approximation is good only when p is small compared to rD. If p was not small enough we can simply run the equation 5.1 with initial parameter $r = 1/2$ [1] until we reach a fixed point. Below we prove that if the point $r = 0$ is stable in the system, there are no other fixed points in the system and, if $r = 0$ is unstable, there is one and only one more fixed point and it must be stable.

For the analysis of the stability of the fixed points we are going to use Lemma 4 stated and proved in Appendix A. In our system we restrict the values for b, p and D to be positive, and $b, p < 1$. To analyse the stability of these fixed points we define the function G:

$$G(r) = r - br + (1-r)(1-(1-p)^{rD})$$

Then we study its first and second derivatives evaluated at the fixed point:

$$G'(r) = (1 - (1-r)D\ln(1-p))(1-p)^{rD} - b$$

$$G''(r) = [2 - (1-r)D\ln(1-p)]\, D\ln(1-p)(1-p)^{rD}$$

It happens that $G(0) = 0$ and $G(1) = 1 - b < 1$. Analysing the first derivative G' we can see that $G' + 1$ is positive in $[0, 1]$, i.e. $G' > -1$ in [0,1].

$$G'(r) + 1 = \underbrace{(1 - (1-r)D\underbrace{\ln(1-p)}_{-})}_{+}\underbrace{(1-p)^{rD}}_{+}\underbrace{-b+1}_{+}$$

[1] The initial parameter value for r is $1/2$ so that it is far enough from the absorbing state $r = 0$.

Analysing the second derivative we can see that $G'' < 0$ in $[0, 1]$:

$$G''(r) + 1 = \underbrace{\left[2 - (1-r)D\underbrace{ln(1-p)}_{-}\right]}_{+} \underbrace{D\underbrace{\ln(1-p)}_{-}(1-p)^{rD}}_{-}$$

Therefore we can use Lemma 4 (in Appendix A) to prove that, if the point $r = 0$ is stable, that makes $G'(0) < 1$ and there are no other fixed points in the system. However, if the point $r = 0$ is unstable, that means that there is one and only one more fixed point in the system and this point must be stable. Because this point is stable we can find it by running the equation 5.1 with initial parameter $r = 1/2$ or any other $r \neq 0$. So we have:

- When $\ln \frac{1}{(1-p)^D} > b$ the fixed point $(0, 1)$ is unstable and the second fixed point, given by running the equation 5.1 (which can be approximated with $(1 - \frac{b}{Dp}, \frac{b}{Dp})$), is stable.
- When $\ln \frac{1}{(1-p)^D} < b$ the fixed point $(0, 1)$ is stable and there are no other fixed points in the system.

In (19) they study a very similar model of SIS in a lattice. We presented this work in the background section. There are two differences between our model and theirs. One is that, while each of the nodes on our model represents an agent, in their model the lattice is not fully occupied by agents but only some vertex are. The second difference is that they include movement in the model. For each time step each agent has a probability m to move to an empty node. The empty node is chosen at random from a set of nodes. This set depends on the range of the movement in the model. The two extreme cases of movement range they model are 1.- only to the neighbours next to the node or 2.- any node in the lattice. They are called "short range" and "long range" movement respectively. They do a mean-field analysis of their model and then compare it with their simulations for different values of the movement range, the probability to transmit the disease,

and the time that an agent remains infected. The mean-field equation for their model is:

$$I = (1 - p_r)I + (C - I)(1 - (1 - p_i I)^z) \tag{5.5}$$

where I is the number of infected in the lattice, p_i is the probability to pass the infection, p_r is the probability to become susceptible, C is the proportion of occupied nodes in the model, and z is the number of neighbours each node has in the lattice. This equation 5.5 gives as a result a solution for the *disease-free* state $I = 0$ and the *endemic* state would be the positive solution of the equality

$$zCp_i - p_r = 0 \tag{5.6}$$

We can see that the equations 5.5 and 5.6 which describe their system and its behaviour is a generalised case of our equations 5.1 and 5.4 where in our case the proportion of occupied nodes is $C = 1$.

In their work they simulate the model and compare the results to the mean-field equation 5.5. They conclude that this is a good approximation when the movement range m is large, but not for small values of it. However, they don't compare their analysis for simulations in which the density in the lattice C is 1, which would correspond to our model. Note that if $C = 1$ there cannot be movement in the lattice as there would not be any empty node for the agents to move to.

5.5 Erdös-Rényi Random Graph

5.5.1 Simulation

The Erdös-Rényi Random Graph has two main parameters. One is the number of nodes in the network and the second one is the density of connections. We create this graph by first creating all the nodes and then choosing with probability d each of the possible edges that join any two nodes in the model (we explained

this process in more detailed in the introduction of this chapter). However, this process does not make sure that our network has only one large component.

We know for previous work that there exists a threshold value in the density parameter d for which above that value a giant component is formed. We have chosen a value of d higher than this threshold so it is likely we get the desired effect. Once we have created our network we make sure that this network has only one large component. If it does not have one large component but different subcomponents we choose two random nodes of two separate components and create an edge between them. We follow the same procedure as many times as we need until all subcomponents are joined together.

The nodes are set to one of the two possible states by random, according to a parameter that determines the proportion number of red nodes there is in the network at the initial state. The nodes will transmit the colour to each other or become white spontaneously in a parallel manner so, once the dynamics has taken place, we update the state of the nodes synchronously.

5.5.2 Mathematical Model

In the Symmetric Graph model all nodes had the same number of edges D. In this model the nodes have different number of edges, but for our mathematical description and analysis we will calculate the average of number of edges per node $A = d(N-1)$. So in this model we will use the same mathematical description as in the previous model 5.1 but substituting the fixed number of edges D by the average number of edges per node A. So we have:

$$r(t+1) = r(t) - br(t) + (1 - r(t))(1 - (1-p)^{r(t)A})$$

$$r(t+1) = r(t) - br(t) + (1 - r(t))(1 - (1-p)^{r(t)d(N-1)})$$

In its steady-state this give us:

$$r = r - br + (1 - r)(1 - (1-p)^{rd(N-1)})$$

$$br = (1 - r)(1 - (1 - p)^{rd(N-1)}) \tag{5.7}$$

We would use the same approximation as in 5.3 we would have that the fixed points that can be estimated as:

$$r_1 = 0$$
$$r_2 = 1 - \frac{b}{pd(N-1)}$$

This is exactly the same mathematical description of the network as we used in the Symmetric Graph analysis so we would get the same fixed points and the same stability. We compare and discuss the results in the next chapter.

5.6 Scale-Free network

5.6.1 Simulation

In this model the population in the system will be represented as a scale-free network. The network is built using the Barabási Albert network model (or BA model) (15) using a *preferential attachment* rule. The parameters that determine the shape of this network are the total number of nodes in the network and the minimum number of edges we introduce in the network for each new node c.

We first create an initial number of nodes c. We connected all nodes to each other. We then create one node at a time and place it in the network by connecting it through edges to c different nodes already in the network. Each of these nodes to connect to is going to be chosen using the preferential attachment rule (the probability to chose a node is directly proportional to its number of edges). We need to make sure that all nodes are different so two different nodes cannot have two edges between them. We create and connect nodes until we have the desired number on the network.

Again, as in previous simulation the nodes are set to one of the two possible states by random according to a parameter that determines the proportion number of red nodes there is in the network at the initial state. The nodes will transmit the colour to each other or become white spontaneously in a parallel manner, so once the dynamics has taken place we update the state of the nodes synchronously.

We are going to present different mathematical models and their analysis with increasing complexity. We will use additional notation in each model so it will be clarified at the beginning of each subsection.

5.6.2 Mean-Field Mathematical Model

If we name $r(t)$ the proportion of infected nodes in the network at time t, and A the average degree in the network. $Ar(t)$ is the average number of infected neighbours a node in the network has. We can write the mean-field equation for the SIS model in a network as follows:

$$r(t) = r(t)(1-b) + (1-r(t))\left[1-(1-p)^{Ar(t)}\right] \tag{5.8}$$

This is the same model that we have used for the Symmetric Graph and the Erdös-Rényi Random Graph models. This model neglects the density correlations among the nodes in the network considering that they all have the same number of neighbours. The mathematical analysis has been explained in previous sections.

5.6.3 First Mathematical Heterogeneous Model

In all remaining models we are going to divide the set of nodes according to their degree. We will then calculate independently for each degree n the number of red nodes at time t, which we name $r_n(t)$. The rest of the notation we will use

in this model remains the same. I will remind the reader of the notation used in this model.

5.6.3.1 Notation

$\mathbf{r_n(t)}$	probability for a node of degree n to be red at time t,
$\mathbf{r_n}$	probability for a node of degree n to be red once the system has reached its asymptotic value,
$\mathbf{r(t)}$	probability for a random node of the network to be red at time t,
$\mathbf{r}$	probability for a random node of the network to be red once the system has reached its asymptotic value.
$\mathbf{b}$	probability for a red node to spontaneously turn white,
$\mathbf{p}$	probability to transmit the red colour to a white neighbour,
$\mathbf{m}$	$:=\frac{b}{p}$,
d_n	degree distribution of the nodes in the network,
d	$:=d_1$; proportion of nodes of degree 1 in the network,
$\langle n \rangle$	average degree in the network.

We modify the mean-field equation so that nodes with a different number of neighbours have different probabilities to get the colour transmitted. Let $r_n(t)$ be the probability that a node of degree n in the network is red. The new model is:

$$r_n(t+1) = r_n(t)(1-b) + (1-r_n(t))\left[1-(1-p)^{nr(t)}\right] \tag{5.9}$$

$$r(t) = \sum_n r_n(t)d_n \tag{5.10}$$

In this model we have replaced the exponent $ar(t)$ by $nr(t)$ where a is an average and n is the actual degree of the node we are studying in the first equation. This equation captures the fact that nodes with higher degrees have a higher probability to have the disease transmitted.

For the mathematical analysis of the system we will again use the approximation $1-(1-x)^a \approx xa$ so equation 5.9 becomes:

$$r_n(t+1) = r_n(t)(1-b) + (1-r_n(t))pnr(t) \tag{5.11}$$

The equivalents of the two equations above describing the asymptotic behaviour of the system would be:

$$r_n = r_n(1-b) + (1-r_n)pnr \tag{5.12}$$
$$r = \Sigma_n r_n d_n \tag{5.13}$$

We can see that the equation describing the dynamics of the process in nodes of degree n (5.12) depends on the total probability to find a red node r (calculated in 5.13) and viceversa.

We now proceed to study the fixed points of the system and their stability. Following the same steps as we did in previous models we find that equation 5.12 conveys to

$$r_n = \frac{nr}{m+nr} \tag{5.14}$$

and our system becomes:

$$r_n = \frac{nr}{m+nr}$$
$$r = \Sigma_n r_n d_n$$

By combining these two equations we get one equation with only one unknown r:

$$r = \sum_n \frac{nr}{m+nr} d_n$$

One solution of this equation is $r_1 = 0$, the other solution, r_2 would be given by:

$$1 = \sum_n \frac{n}{m+nr} d_n \tag{5.15}$$

This system cannot be solved analytically. It seems that numerically can be solved only sometimes. When we could find a solution for this last equality 5.15 numerically, we then substituted r in equation 5.14 and found each r_n independently.

If we define the operator $G : \mathbb{R} \to \mathbb{R}$ as $G(r) = \sum_n \frac{nrd_n}{m+nr}$ and analyse it, we see that $G(0) = 0$, $G(1) < 1$ (given that $m = b/p$ is not 0) $0 < G(r) < 1$ for $r \in (0, 1)$. Looking at its first and second derivative we see that:

$$G'(r) = \sum_n \frac{mn}{(m+rn)^2} d_n \quad > 0$$

with $G'(0) = \sum_n \frac{n}{m} d_n = \frac{\langle n \rangle}{m}$ and:

$$G''(r) = \sum_n -\frac{2nmr}{(m+nr)^3} d_n \quad < 0$$

therefore, using the Lemma of the Fixed Points (4) we know that when $\frac{\langle n \rangle}{m} < 1$ there would be a second solution r_2 given by (5.15) which would be stable, and the solution $r_1 = 0$ would be unstable. If $\frac{\langle n \rangle}{m} \geq 1$ we would have only one solution $r_1 = 0$ and it would be stable.

5.6.4 Extension of the First Mathematical Heterogeneous Model - Considering Nodes of Degree One

In this new model we are going to modify again the factor that determines the probability that each of the neighbours of a given node is infected. In the previous model we estimated this probability as $r(t)n$ for a node of degree n. However, if we look at a node of degree one in our network we see that it cannot be connected to another neighbour of degree one, given that the networks that we are building for this models have only one large component. If it did, then both nodes would form a single component of the network and would be isolated from the main component. That makes the degree distribution of neighbours of a node of degree 1 different from the normal degree distribution d_k. The node of degree 1 will not have neighbours of degree 1, but the proportion of neighbours of each other degree k will remain proportionally the same with respect of all other degrees (except for degree 1). In fact, the degree distribution of neighbours of a node of degree 1 should be $\sum_{k>1} \frac{d_k}{1-d_1}$, which makes the average number of red neighbours $a_1(t) = \sum_{k>1} \frac{d_k}{1-d_1} r_k(t) \neq r(t)$. In the case of having a network with

minimum degree 1 we will define $d := d_1$. So we get two equation systems for the two networks. If the network has been generated by adding only one edge per new node the system is:

$$\begin{aligned} r_1(t+1) &= r_1(t) - br_1(t) + (1 - r_1(t)) \left(1 - (1-p)^{\sum_{k=2} \frac{d_k}{1-d} r_k(t)}\right) \\ r_n(t+1) &= r_n(t) - br_n(t) + (1 - r_n(t)) \left(1 - (1-p)^{nr(t)}\right) \quad \forall\ n > 1 \\ r(t) &= \sum_{k=1} d_k r_k(t) \end{aligned}$$

with steady-state:

$$\begin{aligned} r_1 &= r_1 - br_1 + (1 - r_1) \left(1 - (1-p)^{\sum_{k=2} \frac{d_k}{1-d} r_k}\right) \\ r_n &= r_n - br_n + (1 - r_n) \left(1 - (1-p)^{nr}\right) \quad \forall\ n > 1 \qquad (5.16) \\ r &= \sum_{k=1} d_k r_k \qquad (5.17) \end{aligned}$$

$$(5.18)$$

Using the approximation as in the previous model and looking at the steady-state of the system we find that the asymptotic value of our system, assuming it exists, is given by:

$$r_1 = r_1 - br_1 + (1 - r_1)p\frac{r - dr_1}{1 - d} r_n = r_n - br_n + (1 - r_n)pnr \quad \forall\ n > 1$$

Let us see what would happen in the asymptotic value, assuming it exists:

$$r_1 = r_1 - br_1 + (1 - r_1)p\frac{r - dr_1}{1 - d} r_n = r_n - br_n + (1 - r_n)pnr \quad \forall\ n > 1$$

therefore we get

$$br_1 = (1 - r_1)p\frac{r - dr_1}{1 - d} br_n = (1 - r_n)pnr \quad \forall\ n > 1$$

using abbreviation $m = \frac{b}{p}$ we get:

$$\begin{aligned} r &= \frac{mr_1(1-d) + (1 - r_1)dr_1}{1 - r_1} \qquad (5.19a) \\ r_n &= \frac{rn}{m + rn} \quad \forall\ n > 1 \qquad (5.19b) \end{aligned}$$

By definition we have:

$$r = \sum_{k=1} d_k r_k = dr_1 + \sum_{k>1} d_k r_k \iff$$
$$1 = \frac{cr_1}{r} + \sum_{k>1} \frac{d_k r_k}{r} \tag{5.20}$$

So having 5.19a, 5.19b and 5.20 we get the system:

$$\begin{aligned} r &= \frac{mr_1(1-d) + (1-r_1)dr_1}{1-r_1} \\ r_n &= \frac{rn}{m+rn} \text{ for } n \neq 1 \\ 1 &= \frac{dr_1}{r} + \sum_{k>1} \frac{d_k r_k}{r} \end{aligned} \tag{5.21}$$

Substituting the values for r and r_k, that we got from the two first equations 5.19a and 5.19b, in the third equation 5.20 we get an equality with only one unknown r_1.

$$1 = dr_1 \frac{1-r_1}{m(1-d) + (1-r_1)d} + \sum_{k>1}^{M} d_k \frac{k}{m + k\frac{m(1-d)+(1-r_1)d}{1-r_1}} \tag{5.22}$$

Knowing m, d, and d_k we only need to find r_1 such that this equality is satisfied. This equality 5.22 could be solved numerically using Maple 12 and we found the value of r_1 for different values of m. Then we will calculate r by substituting r_1 in 5.19a, and the rest of the values r_n using the second equality 5.19b.

5.6.5 Second Mathematical Heterogeneous Model

Our previous model brought some heterogeneity in the model as it comprises the fact that nodes with different degrees have a different probability to have the disease transmitted (the higher the degree, the higher the probability). However, it does not yet capture the fact that nodes with different degrees may have different

probabilities to transmit the disease to other nodes, as they are connected to a different number of neighbours. This factor is included in the next model.

Let's take a look at the previous model. It says that if we have a susceptible node of degree n the probability for this node to get the infection transmitted is $nr(t)$. The factor $nr(t)$ is an average of infected nodes over the whole network multiplied by the number of neighbours of our node. This factor does not reflect the fact that nodes with different degrees may have different probabilities to transmit the disease. Let's think of our susceptible node again. Each of its n edges may or may not be connected to an infected node at the other end. A random edge in our network is more likely to be connected to a node of high degree. The probability for a random edge to be connected to a node of degree k is proportional to this degree k and to the density d_k. Then, the probability for a random edge on the network to be connected to an infected node of degree k is proportional to kd_kr_k. We define the probability that a given neighbour of our node is infected, at time t, as $M(t)$:

$$M(t) = \sum_k \frac{kd_kr_k(t)}{\langle k \rangle}$$

where the denominator $\langle k \rangle$ is just a normalising factor in $M(t)$. The definition of $M(t)$ encapsulates the fact that nodes of different degrees may have different probabilities to transmit the infection to a given node. The model becomes:

$$\begin{aligned} r_n(t+1) &= r_n(t)(1-b) + (1-r_n(t))\left[1-(1-p)^{nM(t)}\right] \\ M(t) &= \sum_k \frac{kd_kr_k(t)}{\langle k \rangle} \\ r(t) &= \sum_n r_n(t)d_n \end{aligned} \tag{5.23}$$

This implies that an asymptotic behaviour of the model can be described independently for different degrees as:

$$r_n = \frac{nM}{m+nM} \tag{5.24}$$

$$M = \sum_k \frac{k d_k r_k}{\langle k \rangle} \tag{5.25}$$

$$r = \sum_n r_n d_n \tag{5.26}$$

This means that the higher the degree of the node, the higher the probability to be infected. By using these two equation systems we can write M in terms of itself as:

$$M = \sum_k \frac{1}{\langle k \rangle} \frac{kM}{m + kM} k d_k \tag{5.27}$$

The solution for the equation system given by equations 5.24, 5.26 and 5.27 and we have not been able to calculate it analytically or numerically. If it was possible to solve the system numerically knowing the values of $\langle k \rangle, m$ and d_k, once calculated M we can calculate each value of r_n for each n and we have the total size of the epidemics given by r.

This model has been previously defined and studied by Pastor-Satorras and Vespignani in (73) and (72), although it was defined using a reaction rate equation which is a different approach from ours. This model was studied in terms of epidemiology. Since the system cannot be solved analytically or numerically they use an analytical approximation created as follows. The first use an approximation of the degree distribution the network as $d_k = \frac{2c^2}{k^3}$ taken from the predictions calculated in (15) (c is the number of connections that each new node brings into the network). This approximation of the distribution is claimed to predict the distribution for a network build using the Barabasi and Albert model. However, this has not been the case when we have compared the distribution of our networks as shown in Figure 5.5.

Using f_k as the approximation for the degree distribution, they solve equation 5.27 by substituting the infinite sum by an infinite integral. As a result they get:

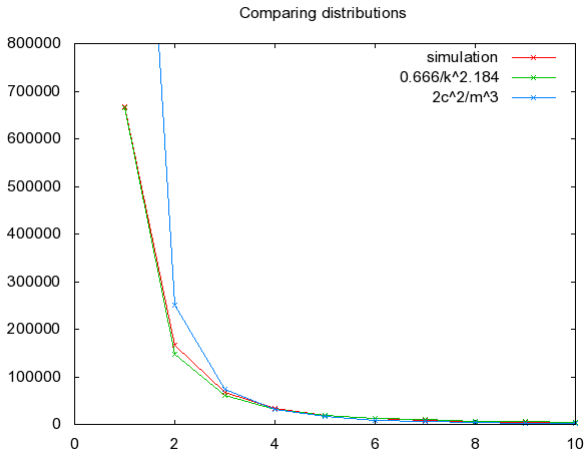


Figure 5.5: In red we have the degree distribution of a scale free network of 1,000,000 nodes created by adding 1 edge per node. In blue we have the approximated degree distribution given by $f_k = \frac{2c^2}{k^3}$ when c=1. In green we see the degree distribution given by our approximation $0.666/k^{2.184}$.

$$\Phi = \frac{e^{\frac{-1}{c\lambda}}}{c\lambda} \quad \text{which is equivalent to our } M \tag{5.28}$$

$$\rho = 2e^{\frac{-1}{c\lambda}} \quad \text{equivalent to our } r. \tag{5.29}$$

where $\lambda = \frac{p}{b}$ so it would be $\lambda = \frac{1}{m}$.

In their analysis they also conclude that the epidemic threshold of the network is given by $\lambda_c = 0$ when the exponent of the degree distribution is $2 < \gamma \leq 3$. This would mean that once a disease has started in the network it will always remain in an endemic stated regardless of the probability to transmit the disease p or the probability to recover b.

5.6.6 Extension of the Second Mathematical Heterogeneous Model - Considering Nodes of Degree One

In this new model we are going to modify again the factor $M(t)$ that determines the probability that each of the neighbours of a given node is infected. We are again going to consider the case in which, during the process of creating the network, we only add one node with one edge at a time. If we look at a node of degree one in our network we see that it cannot be connected to another neighbour of degree one, so the probability that an edge that comes from a node of degree 1 is connected to a node of degree k at the other end is 0 if $k = 1$ and proportional to kd_k if $k > 1$. We define the probability that the only neighbour of a node of degree 1 is infected as:

$$M_1 = \frac{1}{\langle k \rangle - d} \sum_{k>1} k d_k r_k$$

where the dividend $\langle k \rangle - d$ is a normalising factor in M_1. We define the probability that a given neighbour of a node of degree $n \neq 1$ is infected as:

$$M_n = \frac{1}{\langle k \rangle} \sum_{k} k d_k r_k$$

Our mathematical model becomes:

$$\begin{aligned} r_n(t+1) &= r_n(t)(1-b) + (1 - r_n(t)) \left[1 - (1-p)^{nM_n(t)}\right] \\ r(t) &= \sum_n r_n(t) d_n \qquad (5.30) \\ M_n = \frac{1}{\langle k \rangle} \sum_k k d_k r_k \;\text{ if } n \neq 1, \;\text{ and } \quad M_1 &= \frac{1}{\langle k \rangle - d} \sum_{k>1} k d_k r_k \end{aligned}$$

We believe that the absence of neighbours of degree one on nodes of degree one can have an effect in the epidemics of the model considering that nodes of degree one are the most number of nodes in the model.

5.6.6.1 Notation

$\mathbf{r_n(t)}$	probability for a node of degree n to be red at time t,
b	probability for a red node to spontaneously recover,
p	probability to transmit the colour red to a white neighbour,
m	$= \frac{b}{p}$,
$\langle k \rangle$	average degree in the network,
d_k	degree distribution of the nodes in the network,
d	$= d_1$; proportion of nodes of degree 1 in the network,
M_n	probability that a randomly chosen neighbour of a randomly chosen node (of degree n) is red

$$M_n = \frac{1}{\langle k \rangle} \sum_k k d_k r_k \quad \text{if} \quad n \neq 1 \qquad M_1 = \frac{1}{\langle k \rangle - d} \sum_{k>1} k d_k r_k$$

note that if $n \neq 1$

$$M_n = M_1 \frac{\langle k \rangle - d}{\langle k \rangle} + d r_1 \frac{1}{\langle k \rangle} \tag{5.31}$$

5.6.6.2 The Analysis

We have the same system as before

$$r_n(t+1) = r_n(t)(1-b) + (1 - r_n(t)) \left[1 - (1-p)^{n M_n(t)}\right]$$

with asymptotic value of r_n being:

$$r_n = r_n(1-b) + (1 - r_n) \left[1 - (1-p)^{n M_n}\right]$$

Using an approximation $1 - (1-p)^{nM_n} \approx pnM_n$ we have:

$$r_n = \frac{n M_n}{m + n M_n} \tag{5.32}$$

Substituting these values of r_n (5.32) in the definition of M_1 we get:

$$M_1 = \frac{1}{\langle k \rangle - d} \sum_{k>1} k d_k \frac{kM}{m + kM} \tag{5.33}$$

we have re-named M_k (when $k > 1$) as $M = \frac{1}{\langle k \rangle} \sum_k k d_k r_k$ because M_k does not depend on the degree k.

Using the relation 5.31 and the value of r_1 given by 5.32 we can also have M in terms of M_1 as

$$M = M_1 \frac{\langle k \rangle - d}{\langle k \rangle} + d \frac{M_1}{m + M_1} \frac{1}{\langle k \rangle} \tag{5.34}$$

so we can consider M as a function of M_1. To make calculation more clear and simple we are going to define a function $g(M_1)$ where $M = g(M_1)$.

We can therefore re-write M_1 in terms of itself as:

$$M_1 = \frac{1}{\langle k \rangle - d} \sum_{k>1} k d_k \frac{k(M_1 \frac{\langle k \rangle - d}{\langle k \rangle} + d \frac{M_1}{m+M_1} \frac{1}{\langle k \rangle})}{m + k(M_1 \frac{\langle k \rangle - d}{\langle k \rangle} + d \frac{M_1}{m+M_1} \frac{1}{\langle k \rangle})} \tag{5.35}$$

If we know the values d, $\langle k \rangle$ and m in our system we would be able to solve numerically the value of M_1 and also the values of M and r.

In our calculations we will use M_1 as:

$$M_1 = \frac{1}{\langle k \rangle - d} \sum_{k>1} k d_k \frac{k g(M_1)}{m + k g(M_1)} \tag{5.36}$$

It is easy to understand that the probability to find a red node in the network is positive if and only if the probability that a randomly chosen neighbour of a node of degree 1 is red is also positive (i.e. $r > 0$ if and only if $M_1 > 0$). Therefore, studying the stability of our system, given by equation 5.30, is equivalent to studying the stability of the system given by equation 5.36.

Let's define the operator $G(M_1)$ as:

$$G(M_1) = \frac{1}{\langle k \rangle - d} \sum_{k>1} k d_k \frac{k g(M_1)}{m + k g(M_1)} \tag{5.37}$$

Let's study G' and G'':

$$G'(M_1) = \frac{1}{\langle k \rangle - d} \sum_{k>1} k d_k \frac{k m g'(M_1)}{(m + k g(M_1))^2} \tag{5.38}$$

with

$$g'(M_1) = \frac{\langle k \rangle - d}{\langle k \rangle} + \frac{dm}{(m + M_1)^2} \frac{1}{\langle k \rangle} > 0 \quad \text{in } [0, 1]$$

that makes $G(M_1) > 0$ in $[0, 1]$.

$$G''(M_1) = \frac{1}{\langle k \rangle - d} \sum_{k>1} k^2 d_k \frac{kmg''(M_1)m(m + kg(M_1)) - 2mk(g'(M_1))^2}{(m + kg(M_1))^3} \tag{5.39}$$

where

$$g''(M_1) = \frac{-2dm}{(m + M_1)^3} < 0 \quad \text{in } [0, 1]$$

Given that g'' is negative in $[0, 1]$, all terms of G are negative for $M_1 \in [0, 1]$ so $G'' < 0$ in $[0, 1]$. G also satisfies that $G(0) = 0$ and $G(1) < 1$. Therefore we can use the "Theorem of the fixed points" stated in Appendix 1.

Let's now go back to G'. The fixed points (and their stability) of our system depends on the value of G' at 0. We have:

$$G'(0) \overset{1}{=} \left(\frac{d}{m^2} \frac{1}{\langle k \rangle} + \frac{\langle k \rangle - d}{\langle k \rangle} \frac{1}{m} \right) \frac{\langle k^2 \rangle - d}{\langle k \rangle - d} \tag{5.40}$$

The degree distribution of the scale-free network is of the form $d_k = ck^{-\gamma}$. In (15) it is shown that a network built using the *preferential attachment* rule has an exponent γ between 2 and 3. If the exponent γ has value $\gamma < 3$ then $\langle k^2 \rangle - d$ goes to the infinity as the k as the network grows. Therefore, in our network $G'(0)$ will never be less than 1 and the *disease free* state of the system would be unstable. There will be a constant endemic state in the system regardless of any value given to the probability to transmit the disease p or the probability to get recovered b. This is in agreement with models presented in (72) and (73).

In order to study the stability of the fixed point in the case where $\gamma > 3$ we look at $G'(0)$. Doing some manipulations in equation (5.40) we get:

$$G'(0) < 1 \Leftrightarrow m^2 - \frac{\langle k^2 \rangle - d}{\langle k \rangle} m - d \frac{\langle k^2 \rangle - d}{\langle k \rangle - d} \frac{1}{\langle k \rangle} > 0$$

We can think of the expression $m^2 - \frac{\langle k^2\rangle - c}{\langle k\rangle} m - d\frac{\langle k^2\rangle - d}{\langle k\rangle - d}\frac{1}{\langle k\rangle}$ as a polynomial ($p(m) = am^2 - bm - c$) in m with coefficients:

$$\begin{aligned} a &= 1 \\ b &= \frac{\langle k^2\rangle - d}{\langle k\rangle} \\ c &= d\frac{\langle k^2\rangle - d}{\langle k\rangle - d}\frac{1}{\langle k\rangle} \end{aligned}$$

Note that in our system the values for b and c will be always positive. To study the sign intervals of this polynomial we look for the roots of this polynomial which would be:

$$m_1 = \frac{b - \sqrt{b^2 + 4c}}{2} \quad \text{and} \quad m_2 = \frac{b + \sqrt{b^2 + 4c}}{2}$$

because of b and c this polynomial has two real roots with one of them m_1 being negative. In our system m can never adopt a negative value therefore, we are only interested in the sign of the polynomial when m is positive. When $m = 0$ the polynomial is negative and it will be negative until its first positive root m_2.

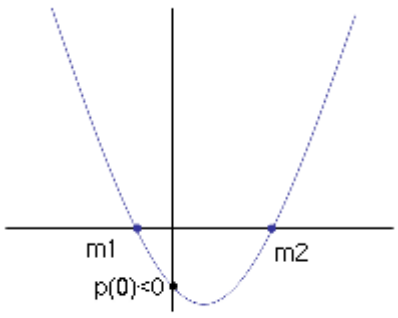


Figure 5.6: Form of the polynomial $p(m)$

We can say that $G'(0) < 1$ for every $m \in [0, m_2)$ and $G(0) \geq 1$ for any $m \geq \frac{b+\sqrt{b^2+4c}}{2}$. We can now substitute b and c for their real values and calculate

the epidemic threshold m_c in terms of $\langle k \rangle$, $\langle k^2 \rangle$ and d:

$$m_c = \frac{1}{2}\frac{\langle k^2 \rangle - d}{\langle k \rangle} + \frac{1}{2}\sqrt{\left(\frac{\langle k^2 \rangle - d}{\langle k \rangle}\right)^2 + 4d\frac{\langle k^2 \rangle - d}{\langle k \rangle - d}\frac{1}{\langle k \rangle}}$$

This last model again could not be solved analytically, and we failed trying to solve it numerically using Maple 12.

5.7 Effective Reproductive Rate R

In our network we have red and white nodes. At each time step a red node will either go back to white with probability b, or transmit its colour to each of its white neighbours with probability p. So, at some point a red node will go back to white state. During the time the node has been in red state, it has transmitted its colour to a number of neighbours. That number is what we call the number of transmissions in a lifetime. We want to know what is the average number of transmissions that a red node does during its lifetime.

In the field that studies the spread of diseases in a population there are two fundamental parameters studied. One is the called *basic reproductive rate* R_0, which measures the number of infections a host would produce in a totally susceptible population, and the *effective reproductive rate* R, which measures the number of actual new infections produced by a host among his contacts during the epidemic process (6).

It is argued that for an epidemic that spreads polynomially for a long period of time the value of the *effective reproductive rate* R is highly constrained; if this rate R is higher than 1 for a period of time t, the disease spreads exponentially (R^t). If this rate R is less than one after a period of time, it will die out. So to have a polynomial growth the rate must be 1, i.e. the average number of transmissions per host in its lifetime must be 1 (94).

This question is also touched in the Metapopulation field by the well known Fermi Paradox (49). The Fermi Paradox presents the following question:

> The galaxy contains roughly a hundred billion stars. If even a very small fraction of these have planets which develop technological civilisations, there must be a very large number of such civilisations. If any of these civilisations produce cultures which colonise over interstellar distances, even at a small fraction of the speed of light, the galaxy should have been completely colonised in no more than a few million years.

What this is really saying is that if each technological civilisation colonises at least one other planet (i.e. the colonisation rate R of a technological civilisation is greater than one), then the galaxy should have been completely colonised. Following the logic used in (94) we should conclude that, if there are planets that develop technological civilisations, the number of planets these civilisations colonise should not be more than one.

In this section we study mathematically what are the effective reproductive rates R for some of our networks according to our mathematical models. We then contrast the results with the simulations of the models.

5.7.1 Calculating R for our models of Symmetric Graphs and ER Graphs

To estimate the **average lifetime** of a red node we do as follows: The probability for a node to spend exactly x time steps in red state, defined in terms of the probability to go back to white state, follows a Geometric distribution: $b(1-b)^{x-1}$ (101) So the average lifetime of a red node is: $\sum_{x=0}^{\infty} xb(1-b)^{x-1} = \frac{1}{b}$.

If our node has exactly s white neighbours the probability to transmit the red colour to exactly k of its white neighbours follows a Binomial distribution (100); $p[k] := \left(\begin{array}{c} s \\ k \end{array} \right) p^k (1-p)^{s-k}$. The average number of transmissions in one time step is: $\sum_{k=0}^{s} kp[k]$. Therefore, we find that, if our node has exactly s white neighbours, the average number of transmissions is:

$$\sum_{k=0}^{s} k \left(\begin{array}{c} s \\ k \end{array} \right) p^k (1-p)^{s-k} = ps$$

So we find that: the average lifetime of a red node is $\sum_{x=0}^{\infty} xb(1-b)^{x-1} = \frac{1}{b}$ and the average number of transmission for a node with s white neighbours is $\sum_{k=0}^{s} k \left(\begin{array}{c} s \\ k \end{array} \right) p^k (1-p)^{s-k} = ps$

Given the estimations above we can say that the average number of transmissions that a red node makes in its lifetime can be estimated as:

$$R = s\frac{p}{b} \tag{5.41}$$

where s is the number of white neighbours of the node, p is the probability to transmit the red colour, b is the probability to spontaneously change from red to white state.

In the mathematical model that describes our Symmetric Graph the number of white neighbours on a red node is estimated as the proportion of white nodes in the system, times the average number of edges per node, $s = (1-r)D$. The estimation of the proportion of red nodes in these models is $r = 1 - \frac{b}{pD}$ so, the number of white neighbours for a red node can be estimated as $s = (1-r)D = \left(1 - (1 - \frac{b}{pD})\right) D = \frac{b}{p}$. Given that the number of transmissions that a red node makes in its lifetime is estimated as $R = s\frac{p}{b}$, and substituting the value of s that we have for our models, we get: $R = s\frac{p}{b} = \frac{b}{p}\frac{p}{b}$ which is obviously 1.

In the mathematical model that describes our ER-Random Graph we follow exactly the same logic: the number of white neighbours on a red node is estimated

as the proportion of white nodes in the system, times the average number of edges per node, $s = (1-r)A$. The estimation of the proportion of red nodes in these models is $r = 1 - \frac{b}{pA}$ so, the number of white neighbours for a red node can be estimated as $s = (1-r)A = \left(1 - (1 - \frac{b}{pA})\right) A = \frac{b}{p}$. Given that the number of transmissions that a red node makes in its lifetime is estimated as $R = s\frac{p}{b}$, and substituting the value of s that we have for our models, we get: $R = s\frac{p}{b} = \frac{b}{p}\frac{p}{b} = 1$.

For the scale-free network models we are going to also use similar analysis to each other. In all models we are going to consider that the average number of transmissions in a lifetime for a red node of degree n is

$$R_n = s_n \frac{p}{b} \tag{5.42}$$

where s_n is the number of white neighbours of a node of degree n, p is the probability to transmit the red colour, and b is the probability to spontaneously change from red to white state. This is the same model used in 5.41. We then calculate the total average number of transmissions in a lifetime in the whole model as:

$$R = \sum_n R_n \frac{r_n d_n}{\sum_n r_n d_n} = \frac{1}{r} \sum_n R_n r_n d_n \tag{5.43}$$

5.7.2 Calculating R for Scale-Free networks - First Heterogeneous Model

For the first heterogeneous model we have the asymptotic behaviour of the system described as: $r_n = r_n - br_n + (1 - r_n)prn$, which leads to $r_n = \frac{nr}{m+nr}$. We also need to make use of the following lemma:

Lemma 1. $\sum_n \frac{n}{m+rn} d_n = 1$

Proof. Following the fact that $r_n = \frac{nr}{m+nr}$ and having defined r as $r = \sum_n r_n d_n$ we see that:

$$r = \sum_n r_n d_n = \sum_n \frac{nr}{m+nr} d_n \Leftrightarrow$$

$$1 = \sum_n \frac{n}{m + nr} d_n$$

□

So, by definition we have:

$$R = \frac{1}{r} \sum_n R_n r_n d_n = \frac{1}{r} \sum_n R_n r_n d_n = \frac{1}{r} \sum_n s_n \frac{p}{b} r_n d_n = \frac{1}{br} \sum_n s_n p r_n d_n$$

let N be the total number of nodes in the network:

$$R = \frac{1}{Nbr} \sum_n s_n p N r_n d_n$$

Given that a red node of degree n is estimated to have s_n white neighbours and it can transmit its red colour to each one of them with probability p then, at each time step it will transmit its colour to $s_n p$ nodes. I can consider $s_n p$ to be the total number of new transmission of each red node of degree n in the network. Also, we know that $N r_n d_n$ is the total number of red nodes of degree n in my network so we have:

$$R = \frac{1}{Nbr} \sum_n \underbrace{\begin{bmatrix}\text{total number of}\\ \text{transmissions}\\ \text{made by a red}\\ \text{node of degree } n\end{bmatrix}}_{s_n p} \underbrace{\begin{bmatrix}\text{total number of}\\ \text{red nodes of de-}\\ \text{gree } n\end{bmatrix}}_{N r_n d_n}$$

$$= \frac{1}{Nbr} \sum_n \begin{bmatrix}\text{total number of}\\ \text{transmissions}\\ \text{made by all red}\\ \text{nodes of degree}\\ n\end{bmatrix} = \frac{1}{Nbr} \begin{bmatrix}\text{total number of}\\ \text{transmissions}\\ \text{made by all red}\\ \text{nodes in the}\\ \text{network}\end{bmatrix}$$

$$= \frac{1}{Nbr} \sum_k \begin{bmatrix}\text{total number of}\\ \text{new red nodes of}\\ \text{degree } k\end{bmatrix} = \frac{1}{Nbr} \sum_k \begin{bmatrix}\text{total number of}\\ \text{white nodes of}\\ \text{degree } k\end{bmatrix} \begin{bmatrix}\text{probability for a}\\ \text{white node of}\\ \text{degree } k \text{ to get}\\ \text{transmited}\end{bmatrix}$$

$$= \frac{1}{Nbr}\sum_k [(1-r_k)Nd_k][krp] = \frac{1}{Nbr}\sum_k \left[(1-\frac{rk}{m+rk})Nd_k\right] krp$$

$$= \frac{1}{Nbr}\sum_k \frac{m}{m+rk} Nd_k krp = \sum_k \frac{k}{m+rk} d_k = 1$$

In the calculations above, and also in the calculations made for the next three models, we have estimated the average transmissions in a lifetime of a red node of degree n as $R_n = s_n \frac{p}{b}$, where s_n is the number of white neighbours the red node has. In this model we calculate r as the probability for a random neighbour of a random node to be red, therefore, $(1-r)$ would be the probability for a random neighbour of a random node to be white. That makes $s_n = n(1-r)$ a good estimate of the number of white neighbours for a node of degree n. That would make $R_n = n(1-r)\frac{p}{b}$ a good estimate of number of transmissions in a lifetime for a red node of degree n.

5.7.3 Calculating R for Scale-Free networks - Second Heterogeneous Model

For the first heterogeneous model we have the asymptotic behaviour of the system described as: $r_n = r_n - br_n + (1-r_n)pMn$, which leads to $r_n = \frac{nM}{m+nM}$. In this case we need to make use of the following lemma:

Lemma 2. $\sum_n \frac{1}{r}\frac{nM}{m+Mn} d_n = 1$

Proof. Following the fact that $r_n = \frac{nM}{m+nM}$ and having defined r as $r = \sum_n r_n d_n$ we see that:

$$r = \sum_n r_n d_n = \sum_n \frac{nM}{m+nM} d_n$$

$$1 = \frac{1}{r}\sum_n \frac{nM}{m+nM} d_n$$

□

So we again have:

$$R = \frac{1}{Nbr} \sum_n s_n p N r_n d_n$$

and in this case it follows:

$$R = \frac{1}{Nrb} \sum_n \underbrace{\begin{bmatrix}\text{total number of}\\ \text{transmissions}\\ \text{made by a red}\\ \text{node of degree } n\end{bmatrix}}_{(1-M)np} \underbrace{\begin{bmatrix}\text{total number of}\\ \text{red nodes of de-}\\ \text{gree } n\end{bmatrix}}_{Nr_n d_n}$$

$$= \frac{1}{Nrb} \sum_n \begin{bmatrix}\text{total number of}\\ \text{transmissions}\\ \text{made by all red}\\ \text{nodes of degree}\\ n\end{bmatrix} = \frac{1}{Nrb} \begin{bmatrix}\text{total number of}\\ \text{transmissions}\\ \text{made by all red}\\ \text{nodes in the}\\ \text{network}\end{bmatrix}$$

$$= \frac{1}{Nrb} \sum_k \begin{bmatrix}\text{total number of}\\ \text{new red nodes of}\\ \text{degree } k\end{bmatrix} = \frac{1}{Nrb} \sum_k \begin{bmatrix}\text{total number of}\\ \text{white nodes of}\\ \text{degree } k\end{bmatrix} \begin{bmatrix}\text{probability for a}\\ \text{white node of}\\ \text{degree } k \text{ to get}\\ \text{transmitted}\end{bmatrix}$$

$$= \frac{1}{Nrb} \sum_k [N(1-r_k)d_k][Mkp] = \frac{1}{Nrb} \sum_k \left[N(1 - \frac{Mk}{m+Mk})d_k\right] Mkp$$

$$= \frac{1}{Nrb} \sum_k \frac{m}{m+Mk} NkMpd_k = \frac{1}{r} \sum_k \frac{kM}{m+Mk} d_k = 1$$

In this case we have that $R_n = n(1-M)\frac{p}{b}$ is an estimate of the number of transmissions in a lifetime for a red node of degree n.

5.7.4 Calculating R for Scale-Free networks - Extension of the Second Heterogeneous Mathematical Model

For the first heterogeneous model we have the asymptotic behaviour of the system described as: $r_n = r_n - br_n + (1-r_n)pMn$ for $n > 1$ and $r_1 = r_1 - br_1 + (1-r_1)pM_1$, which leads to $r_n = \frac{nM}{m+nM}$ for $n > 1$ and $r_1 = \frac{M_1}{m+M_1}$. where:

$$M_1 = \frac{1}{\langle k \rangle - d} \sum_{k>1} k d_k r_k$$

$$M = \frac{1}{\langle k \rangle} \sum_k k d_k r_k$$

The total average number of red nodes in the network is given by:

$$r = \sum_n r_n d_n = \frac{M_1}{m + M_1} d_1 + \sum_{n>1} \frac{nM}{m + nM} d_n$$

For the analysis of R in this model we need to prove the following lemma:

Lemma 3. $\frac{1}{r}\left(\frac{M}{m+M} d_1 + \sum_{n>1} \frac{nM}{m+Mn} d_n\right) = 1$

Proof. Following the fact that $r_n = \frac{nM}{m+nM}$ for $n > 1$ and $r_1 = \frac{M_1}{m+M_1}$ and having defined r as $r = \sum_n r_n d_n$ we see that:

$$r = \sum_n r_n d_n = \frac{M_1}{m + M_1} d_1 + \sum_{n>1} \frac{nM}{m + nM} d_n \Leftrightarrow$$

$$1 = \frac{1}{r}\left(\frac{M}{m + M} d_1 + \sum_{n>1} \frac{nM}{m + Mn} d_n\right)$$

□

So we have,

$$R = \frac{1}{Nbr} \sum_n s_n p N r_n d_n$$

and it follows:

$$R = \frac{1}{Nrb} \sum_n \underbrace{\begin{bmatrix}\text{total number of}\\ \text{transmissions}\\ \text{made by a red}\\ \text{node of degree } n\end{bmatrix}}_{(1-M)np} \underbrace{\begin{bmatrix}\text{total number of}\\ \text{red nodes of de-}\\ \text{gree } n\end{bmatrix}}_{N r_n d_n}$$

$$= \frac{1}{Nrb} \sum_n \begin{bmatrix}\text{total number of}\\ \text{transmissions}\\ \text{made by all red}\\ \text{nodes of degree}\\ n\end{bmatrix} = \frac{1}{Nrb} \begin{bmatrix}\text{total number of}\\ \text{transmissions}\\ \text{made by all red}\\ \text{nodes in the}\\ \text{network}\end{bmatrix}$$

$$= \frac{1}{Nrb} \sum_k \begin{bmatrix} \text{total number of} \\ \text{new red nodes of} \\ \text{degree } k \end{bmatrix} = \frac{1}{Nrb} \sum_k \begin{bmatrix} \text{total number of} \\ \text{white nodes of} \\ \text{degree } k \end{bmatrix} \begin{bmatrix} \text{probability for a} \\ \text{white node of} \\ \text{degree } k \text{ to get} \\ \text{transmitted} \end{bmatrix}$$

$$= \frac{1}{Nrb} \left([N(1-r_1)d_1][M_1 p] + \sum_{k>1} [N(1-r_k)d_k][Mkp] \right)$$

$$= \frac{1}{Nrb} \left(N(1 - \frac{M_1}{m+M_1})d_1 \sum_k \left[N(1 - \frac{Mk}{m+Mk})d_k \right] Mkp \right)$$

$$= \frac{1}{Nrb} \left(\frac{m}{m+M_1} N M_1 p d_1 + \sum_{k>1} \frac{m}{m+Mk} N k M p d_k \right)$$

$$= \frac{1}{r} \left(\frac{M_1}{m+M_1} d_1 + \sum_{k>1} \frac{kM}{m+Mk} d_k \right) = 1$$

In this case $s_n = n(1-M)$ is an estimate of the number of white neighbours for a node of degree $n > 1$ and $s = (1-M)$ for the nodes of degree 1. That would make $R_n = n(1-M)\frac{p}{b}$ a good estimate of the number of transmissions in a lifetime for a red node of degree $n > 1$. In the same way $R_1 = (1-M_1)\frac{p}{b}$

5.8 Chapter Summary

In this chapter we studied a model of SIS dynamics on a network. The SIS dynamics can be understood as a representation of different processes in a population (not only related to epidemics) such as the spread of a common cold in a population, the spread of a computer virus on a network, or the phenomenon of extinction and recolonisation in different graphical areas. We implement the individuals as a set of nodes and the interaction between the individuals as edges between the nodes. The nodes and edges form a network in which the spread process takes place. We define the process in terms of colours as follows: Nodes can be in either red or white state. A red node will transmit its colour to each

of its white neighbouring nodes with probability p. A red node will turn spontaneously to white with probability b. The aim of these models is to investigate how the topology of a social network affects the "spread of the colour red" in the network.

We implement networks with three different topologies in which the process takes place. The first network is a Symmetric Graph 5.4, for which we give a formal definition in Section 5.2. The mathematical model says that the steady-state of the red colour is given by equation 5.2 which has two fixed points, $r_1 = 0$ and $r_2 \approx 1 - \frac{b}{Dp}$. Using the *Lemma of the Fixed Points* 4 we get to the conclusion that when $ln\frac{1}{(1-p)^D} > b$ the fixed point $r_1 = 0$ is unstable and the fixed point r_2 is stable. On the other hand, if $ln\frac{1}{(1-p)^D} <= b$ the fixed point $r_1 = 0$ is stable and r_2 unstable. That means that when $ln\frac{1}{(1-p)^D} > b$ the system is in an "endemic" state (i.e. there are always red nodes in the system) and the number of red nodes in the system is given by r_2. However, if $ln\frac{1}{(1-p)^D} <= b$ the red colour would disappear form the system, regardless of the initial number of red nodes in the network.

The second network topology we implement is the called *ER-Random Graph* 5.5, which is formed by first creating N nodes and then choosing each possible edge in the network with independent probability d. This makes the average number of neighbours per node in the network as $A = d(N-1)$. The mathematical analysis says that the dynamics of the system is given by equation 5.7, which has two fixed points $r_1 = 0$ and $r_2 \approx 1 - \frac{b}{Ap}$. Using the *Lemma of the Fixed Points* 4 we get to the conclusion that when $ln\frac{1}{(1-p)^A} > b$ the fixed point $r_1 = 0$ is unstable and the fixed point r_2 is stable, and if $ln\frac{1}{(1-p)^A} <= b$ the fixed point $r_1 = 0$ is stable and r_2 unstable. That means that if $ln\frac{1}{(1-p)^A} > b$ then the system is in an "endemic" state (i.e. there are always red nodes in the system) and the number of red nodes in the system is given by r_2. However, if $ln\frac{1}{(1-p)^A} <= b$ the red colour would disappear form the system, regardless of the initial state of the system.

The third topology we studied is the scale-free network 5.6. The network is built using the Barabási Albert network model (or BA model) (15) using a

preferential attachment rule. The network will have a total of N nodes, and each time we add a new node to the network, it will initially connect to c neighbours. For analysing mathematically the process in this network topology we create four different mathematical models.

- The first model is the same as in ER-Random Graph model in which the average number of neighbours per node is given by $A = 2c$, where the fixed points of the system are the same as before $r_1 = 0$ and $r_2 \approx 1 - \frac{b}{Ap}$, and the stability analysis is also the same as the models stated before.
- First Mathematical Heterogeneous model is presented in Section 5.6.3 and considers the fact that nodes of different degrees have different probabilities to get the colour red transmitted. The dynamics of the system in its steady-state is estimated by equations 5.9 and 5.10. We have $r = 0$ as one fixed point but the solution for the second fixed point cannot be solved analytically. It has been solved numerically with Maple 12, but only for some parameter values. The analysis of the stability of the fixed points says that when $\frac{2cp}{b} < 1$ we have that $r = 0$ is unstable and $r = \sum_n r_n d_n$ with $r_n = \frac{nr}{m+nr}$ is stable so the system is in an endemic state, and when $\frac{2cp}{b} > 1$ the fixed point $r = 0$ is stable, and we will get to a *red free* state.
- The Extension of the First Mathematical Heterogeneous model is presented in Section 5.6.4. It is a model of a network in which the nodes has a minimum number of neighbours $c = 1$. This mathematical model considers the fact that nodes of degree 1 are not connected to nodes of degree 1, so the probability to get transmitted the colour red from another node of degree 1 is 0. This makes the probability to get the colour transmitted (for a node of degree 1) as: $\sum_{i>1} \frac{d_k}{1-d_1}$ where d_k is the probability that a randomly chosen node has degree n. This gives us a mathematical system defined by the equations 5.16, which gives $r = 0$ as one fixed point. The second fixed point should be deduced from equations system, but we have not been able to solve the system analytically. However, we have been able to solve it numerically with Maple 12 for all relevant values. We have not made an analysis of the stability of the fixed points.

- The fourth mathematical model is called the Second Mathematical Heterogeneous model 5.6.5 and it comprises the fact that nodes with different degrees have different probability to get transmitted the colour red (as in the two previous models), and the fact that red nodes of higher degrees have more probabilities to transmit its colour because they are connected to more neighbours. The mathematical system is given by equation 5.23. One fixed point is $r = 0$ and the second has to be deduced from the equations above. We have not been able to solve this system analytically or numerically. Pastor-Satorras and Vespignani (72; 73) have solved this system analytically by approximating the degree distribution $d_k \approx \frac{2c^2}{k^3}$ and by approximating the infinite sum in equation 5.23 to an integral. The result they get is an estimation of the number of red nodes as $\rho = 2e^{\frac{-b}{cp}}$. In their analysis they conclude that for scale-free network models which have a degree distribution with exponent $2 < \gamma \leq 3$ the fixed point $r = 0$ is always unstable, so once there is colour red in the network the colour is going to remain in the network forever.

- The last mathematical model is an Extension of the Second Mathematical Heterogeneous model 5.6.6. It is a model of a network with $c = 1$ and it takes into account the fact that nodes of degree 1 are not connected to other nodes of degree 1. The system is defined by equations 5.30. One of the fixed points is $r = 0$ and the other has to be deduced from the system. Unfortunately we have not been able to solve it analytically or numerically. However, we have successfully studied the stability of the fixed points coming to the same conclusion as Pastor-Satorras and Vespignani; for our scale-free network models, which have a degree distribution with exponent $2 < \gamma \leq 3$, the fixed point $r = 0$ is always unstable. This implies that once the colour red is in the network, it remain in the network forever regardless of the probabilities to become white, or to transmit the colour.

As part of our interests we also investigate the effective reproductive rate R in our systems 5.7. We make a mathematical analysis for the three different network topologies. All the mathematical models estimate the average number

of transmissions in a lifetime per red node in the network $R = 1$, regardless of the values for p, b or even the network topology. We have also estimated the average number of transmissions in a lifetime for nodes of degree n in a scale-free network with $c = 1$ as $R_n = n(1 - M_n)\frac{p}{b}$, with M_n defined as in the Extension of the Second Mathematical Heterogeneous model 5.6.6.

One of the first questions of this thesis is to understand how accurate our mathematical analysis is and what information we can infer from the analysis about the dynamics of the models. In the next Chapter 4.1 we present several experiments and contrast the results of this simulations with the outcomes of the mathematical analysis of each model. The experiments are normally set with a large number of nodes in the networks, as we know that the mathematical model reflects the behavior of an infinite population.

We intend to test the mathematical models created for the Symmetric Graph model and the ER-Random model, investigating its accuracy for different values of the probabilities b and p and different number of neighbours per node. We also want to investigate the accuracy of the analysis of the stability of the fixed points.

We will also run experiments for scale-free networks with minimum number of neighbours per node being $c = 5$ and $c = 1$. We want to study the accuracy of our mathematical estimations by the First Mathematical Heterogeneous model and its Extension. We are looking both at the estimation of the total number of red nodes in the network r, and the number of red nodes of different degrees r_n. We also compare the values of r estimated by different mathematical models, including the approximation by Pastor-Satorras and Vespignani, for a network with minimum number of neighbours per node being $c = 1$.

We finally run experiments in scale-free networks with different minimum number of neighbours per node and different values b and p studying the effective reproductive rate R in all of them. We also test our estimation of R_n for a scale-free network with $c = 1$.

Chapter 6

Experiments and Results

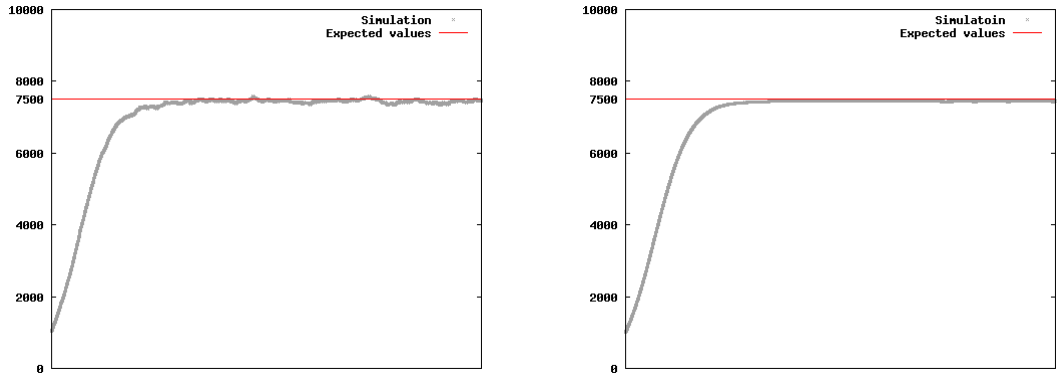
This chapter presents a set of experiments that we run in order to investigating how accurate our mathematical models are and to study how some parameters affect the asymptotic behaviour of the systems modelled.

6.1 Symmetric Graph - Experiments and Results

We have created a symmetric graph of 100×100 nodes with the form of a 2 dimensional lattice, (so each person is connected to people that are right on top, below, right or left of him) with the depth of the connections being 4, (so the degree of connections D of each person is 16). We set the probability to pass the colour between two connected nodes as $p = 0.002$ and the probability to recover back to white as $b = 0.0005$. The starting number of red nodes is 1,000, i.e. 10% of the set of nodes. The mathematical model that corresponds to the simulation used for this experiment 5.4 estimates that, when the simulation reaches its steady

state, the proportion of red nodes under these conditions is 0.75, or a total of 7.500 red nodes in the network.

We can see that the mathematical solution is very accurate in this case. As the simulation runs, the number of red nodes in the simulation increases. At some point the values don't increment any more, but oscillate smoothly around a value, which we will call attractor. This attractor seems to be very close to the expected value given by the result of the mathematical model. We can also observe the result of running 100 runs of this experiment and taking the average. The line that follows this graph is very smooth and it clearly stabilises at the expected point.



(a) Single run

(b) Average of 100 runs

Figure 6.1: The horizontal axis represents time ranging from 0 to 5,000. The vertical axis represents the number of red nodes in the system. Experiment in a 100×100 lattice with depth = 4, b=0.0005, p=0.002. The initial proportion of red nodes in the simulations is 10%. In both graphs we show two lines, one representing the estimated number of red nodes when the model is in its steady state, given by the mathematical model, and the solution of the simulations in a) one single run and b) the average of 100 runs.

Figure 6.2 shows what happens if we change the values of b and p leaving the ratio b/p fixed (leaving the expected proportion of infected people in r=0.75). Parameters b and p will vary according to table 6.1. The number of red nodes in

the simulation oscillate around different attractors depending on the values for b and p and as we increase the values of b and p, the distance between the model's attractor and the mathematical solution increases.

	p	b
experiment 1	0.0005	0.002
experiment 2	0.005	0.02
experiment 3	0.01	0.04
experiment 4	0.02	0.08
experiment 5	0.03	0.12
experiment 6	0.04	0.16
experiment 7	0.05	0.2

Table 6.1: Combinations of values for b and p used in the experiment below.

This means that the mathematical model would be accurate only if both the probability p to pass the colour red from one node to another and the probability b to go back to white are very small. As the probabilities increment in value the mathematical model looses its predictive power and the number of red nodes will be smaller than expected by the mathematical solution. The reason for this may be the use of approximation $1-(1-p)^{rD} \approx prD$ in the analysis of this model, since this approximation works best when p is very small. It is interesting to see that, even though the mathematical model does not predict well in some cases, it seems to be an upper bound of the number of red nodes in the system.

In Figure 6.3 we can also observe that as the values of b and p increase, the oscillations of the models' results around its respective attractors get wider, larger and more sharp and sudden as b and p increase. An explanation for this phenomena may be that, as the probabilities to change state (white to red or red to white) get larger, it is more common to have greater differences in the values of the simulation between consecutive time steps, so we can observe there are sudden changes in the simulation in short time periods. This phenomenon can also be observed in Figure 6.4 where we show the standard error of the average

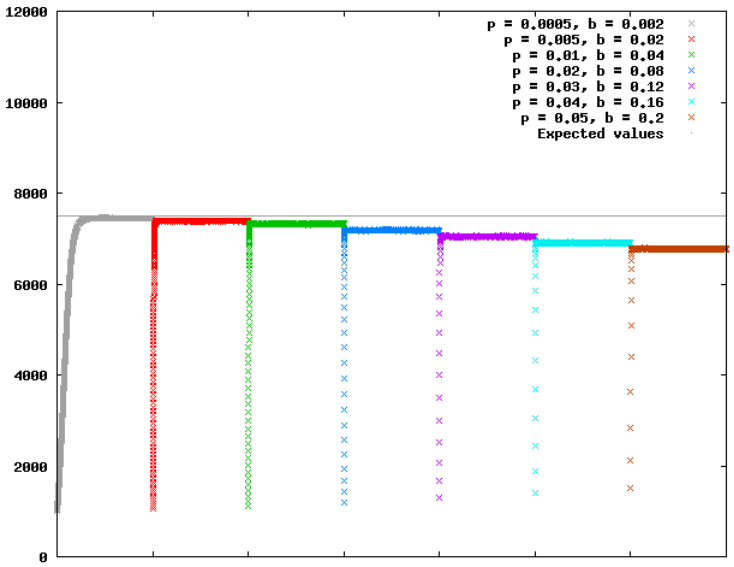


Figure 6.2: Experiment in a 100×100 lattice with depth = 4, b/p = 4. Average of 100 runs. The vertical axis represents the total number of red nodes in the system. The horizontal axis represents time for different values of b and p. We change the values of b and p, leaving b/p constant. The expected number of red nodes in the system, given by the mathematical model, is 7,500 in all cases. As b and p increase, the mathematical model looses its predictability power.

of 100 runs for experiments with p=0.0005, p=0.005 and p=0.05; it shows how the standard error reaches higher values as p increases.

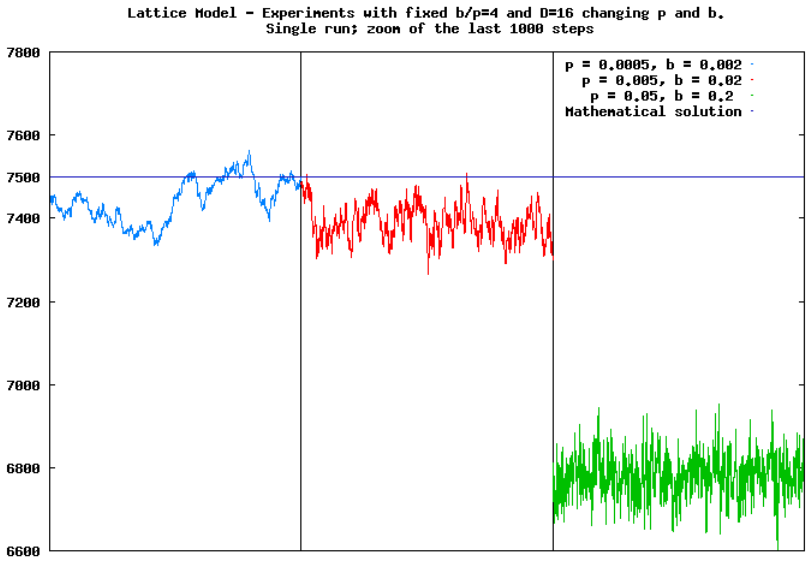


Figure 6.3: The vertical axis represents the total number of red nodes in the system. The horizontal axis represents time t for three different sets of values for b and p. The parameter values are: depth = 4, and: b=0.002, 0.02, 0.2 and p = 0.0005, 0.005, 0.05 respectively. t goes from 4,000 to 5,000 (i.e. the last 1,000 of each run). As b and p increase, the oscillations of the number of red nodes around its expected value (given by the mathematical model) get larger and sharper.

We are also interested in studying the effect of parameter D (number of neighbours of each nodes) in the simulation. We are again working with a 100×100 lattice in which the nodes are connected to other nodes situated next above, below, left and right with depth D'. The degree of each node is $D = 4 * D'$. We will run experiments for values of D' from 2 to 20 every 2. We run this experiment 100 times and we calculate the average. For this experiment we fix the probabilities to $b = 0.03$ and $p = 0.0025$. The result of this experiment can be observed in Figure 6.5. We can see the results of the simulation in different colours for different values of the depth D'. We can also see, for each value of D', two lines that represent 1.- the average of each run in the last 3,000 steps,

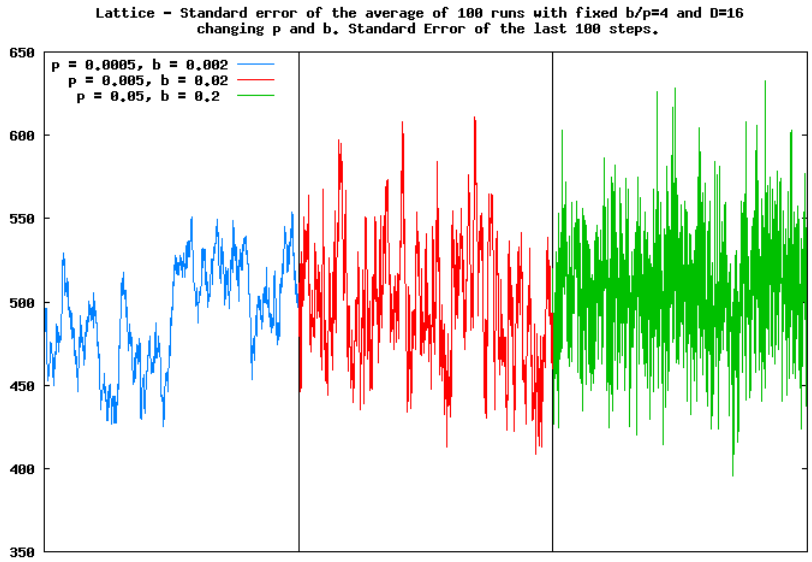


Figure 6.4: The vertical axis represents the standard error of the average number or red nodes, when we average 100 runs. The horizontal axis represents time t for three different sets of values for b and p. The parameter values are: depth = 4, and: b=0.002, 0.02, 0.2 and p = 0.0005, 0.005, 0.05 respectively. t goes from 4,000 to 5,000 (i.e. the last 1,000 of each run). As b and p increase, the standard error of the simulation, in respect to the average of 100 runs, gets larger.

which can be considered as the attractor of the simulated system, and 2.- the mathematical solution for each D. As we can see in this figure the solution given by the mathematical model is more accurate as D increases in value.

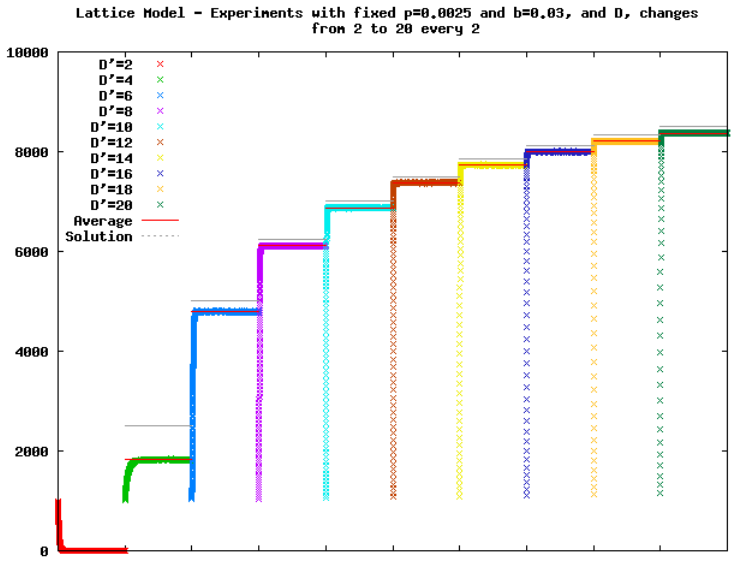


Figure 6.5: The vertical axis represents the number of red nodes. The horizontal axis represents time (t from 0 to 1,000) for different values of D' (D' form 2 to 20 with increments of 2). The simulations are set with 10% initial number of red nodes, $p = 0.0025$ and $b = 0.03$. We can see that as D increases, the predictability power of the mathematical solution looses its power.

In Figure 6.6 we can see a better representation of this phenomena. The values of Figure 6.6 are the absolute value of the difference between the mathematical solution and the attractor of the simulation, divided by the mathematical solution[1] so these values are dimensionless and easier to compare between them for different values of D. We call this value *Error*. When the value *Error* is 0 that means that the mathematical solution and the simulation value are the same, and when its value is 1 it means that the mathematical model is positive and the simulation result is 0. For value $D = 2$ the mathematical solution is the same as

[1] $Error = \left| \frac{\text{Math - Sim}}{\text{Math}} \right|$ where "Math" is the mathematical solution and "Sim" is the average result of 100 runs of the simulation.

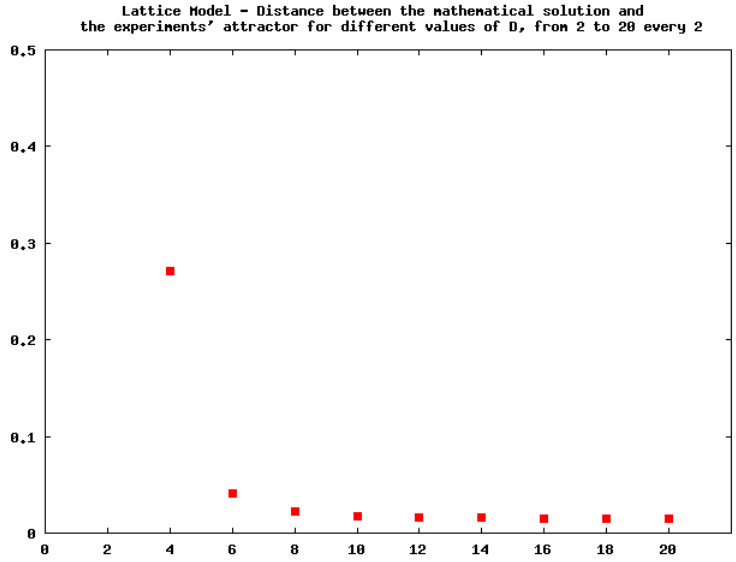


Figure 6.6: The vertical axis represents the *Error* or difference between the mathematical solution and the experiments attractor when the simulations are in its steady state. The horizontal axis represents the depth of connections in the lattice D'. We can see that the *Error* decreases as D' increases its value.

the simulation run because the degree D is too small to let the colour red spread in the network (as predicted by the mathematical model). However, in the rest of the runs we can see the relation between the value D' and the accuracy of the mathematical solution.

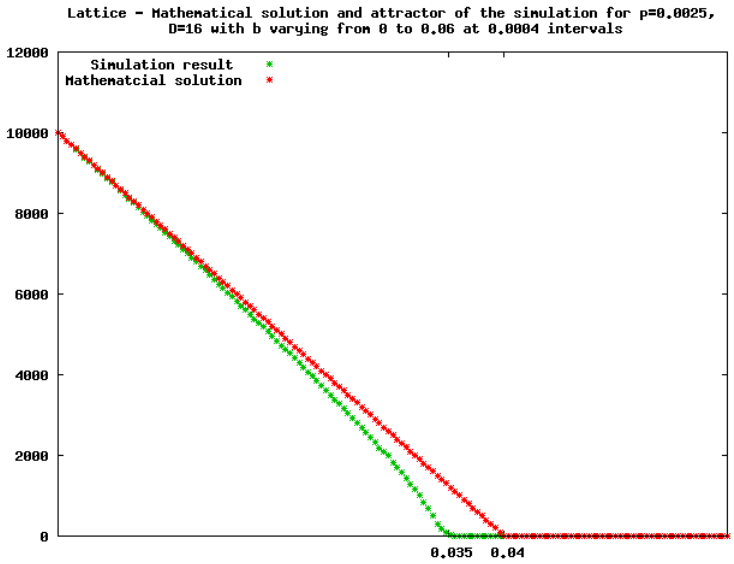


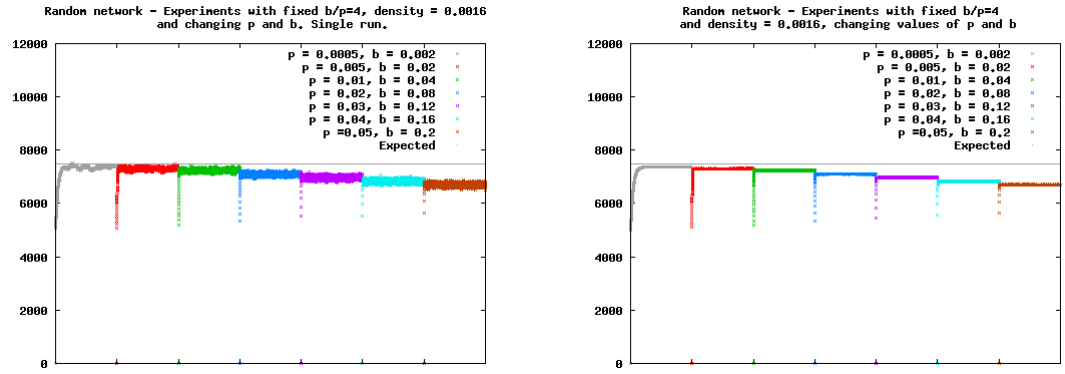
Figure 6.7: The vertical axis represents the total number of red nodes when the simulation is in its steady state. The horizontal axis represents the value of the parameter b in each simulation. The value of b changes from 0 to 0.06. at 0.0004 intervals. The simulations are 100×100 lattice models with D=16 and p=0.0025. The simulations are run for 5,000 time steps in order to make sure that they are in its steady state. We compare the attractor of the simulations results with mathematical solutions.

We have run several experiments in a 100×100 lattice with parameter values for $D' = 4$ ($D = 16$) and $p = 0.0025$, making one run for each value of b between 0 and 0.8 increasing b at 0.004 intervals. For each value of b we have run 56 experiments and calculated the average. The result of this average is shown in Figure 6.7 in green. The propose of this experiment is to compare the results of this simulation and the mathematical result regarding the stability of the fixed points $r_1 = 0$ and $r_2 \approx \frac{b}{Dp}$. The mathematical model says that, if

$b >= Dp$, then the fixed point $r_1 = 0$ is stable and, when $b < Dp$, the fixed point $r_2 \approx \frac{b}{Dp}$ is stable. In Figure 6.7 we can see in red the different expected fixed points of the mathematical model as the parameter b changes. We can see that, at first, the values in these averages are very close to the fixed points of the mathematical model. As b increases, the averages of the simulation drop faster than the fixed points of the mathematical model. We have seen this effect in previous experiments shown in Figure 6.2. We can see that the averages of the simulation runs reach the value 0 at around $b = 0.035$. As we said previously, we can observe that the number of red nodes in the simulation is equal or less than the number of red nodes expected from the mathematical model, so the fixed point of the mathematical model could be used as an upper bound of this value. The analysis of the stability of the fixed points $r = 0$ and $r = 1 - \frac{b}{Dp}$ said that, for $b < \ln \frac{1}{0.9975^{16}} = 0.0400500835$, the fixed point $r = 0$ is unstable and $r = 1 - \frac{b}{Dp}$ stable. However, looking at the experiments results, the asymptotic value of the average number of red nodes seems to drop faster than expected as we approach the value $b \approx 0.035$, and the point $r = 0$ seems to be stable for values $b > 0.035$.

6.2 ER-Random Graph - Experiments and Results

In the first set of experiments we created an ER-Random Graph of 10,000 nodes with density $d = 0.0016$, which give us an average number of edges of approximately $A = 16$. We use different probabilities b and p according to the same table used in the Symmetric Graphs 6.1. These values keep the ratio $b/p = 4$ and therefore, following the mathematical model, we have that the expected number of red nodes in the simulation is 7.500 nodes, for ever set of values b, p. In Figures 6.8 we can see the results of this experiments. In the first figure we show one single run for each set of values b and p. In the second graph we show the average of 100 runs for each set b, p. The results are very similar to the ones for Symmetric Graphs, as we increase the probabilities b and p, the simulation



(a) Single runs

(b) Averages of 100 runs

Figure 6.8: Experiment in a ER-Random Graph of 10,000 nodes with density $d = 0.0016$. The vertical axis represents the total number of red nodes in the system, for single runs in the first graph and for the average of 100 runs in the second graph. The horizontal axis represents time for different values of b and p. We change the values of b and p, leaving b/p constant. The expected number of red nodes in the system, given by the mathematical model, is 7,500 in all cases. We compare the result of the simulations with the expected number of red nodes in the system when it is in its steady state, given by the mathematical model. As b and p increase, the mathematical model looses its predictability power.

results get more distant from the mathematical solution, being the mathematical solution an upper bound of the simulation results.

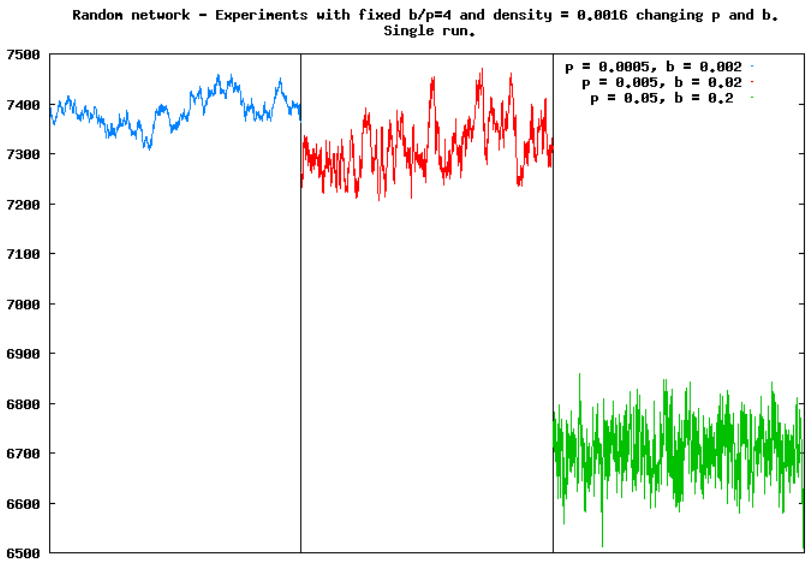


Figure 6.9: Results of simulations run in a ER-Random graph with 10,000 nodes and density $d = 0.0016$. The vertical axis represents the total number of red nodes in the system. The horizontal axis represents time t for three different sets of values for b and p. The values of b and p are: b=0.002, 0.02, 0.2 and p = 0.0005, 0.005, 0.05 respectively. t goes from 4,000 to 5,000 (i.e. the last 1,000 of each run). As b and p increase, the oscillations of the number of red nodes around its expected value (given by the mathematical model) get larger and sharper.

We can also observe in single runs (Fig. 6.9) that, as values for b and p get larger, the oscillations of the simulation results around the different attractors get larger and more sudden. We also see in Figure 6.10 that the standard error of the average of the 100 simulation results is higher for larger values of b and p. These are the same results we have seen for the Regular Symmetric Graph experiments.

In this experiment we study what are the effects on the simulation when we change the density of the model d. By changing the density d, the total number

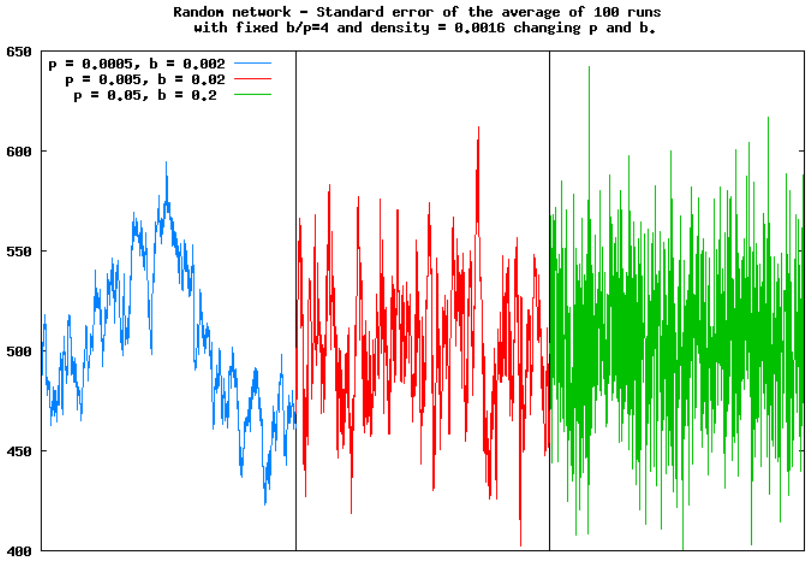


Figure 6.10: Results of simulations run in a ER-Random graph with 10,000 nodes and density $d = 0.0016$. The vertical axis represents the standard error of the average number or red nodes, when we average 100 runs. The horizontal axis represents time t for three different sets of values for b and p. The values of b and p are b=0.002, 0.02, 0.2 and p = 0.0005, 0.005, 0.05 respectively. t goes from 4,000 to 5,000 (i.e. the last 1,000 of each run). As b and p increase, the standard error of the simulation, in respect to the average of 100 runs, get larger.

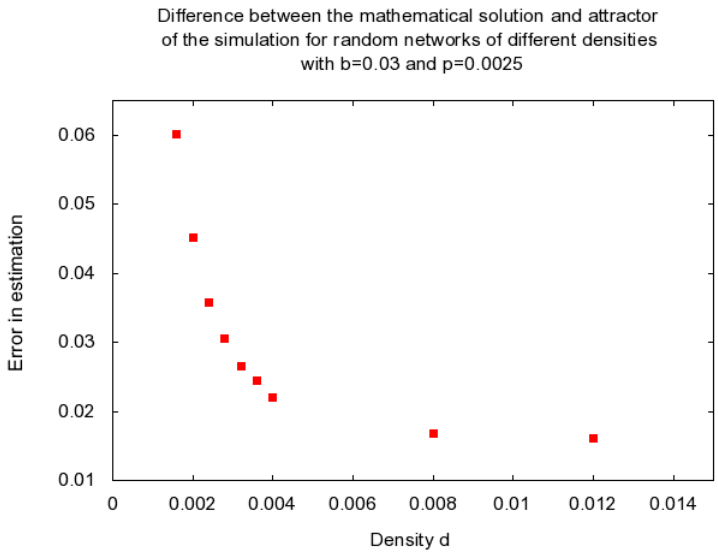


Figure 6.11: Results of simulations run in a ER-Random graph with 10,000 nodes $b = 0.03$ and $p = 0.0025$. The vertical axis represents the *Error* or difference between the mathematical solution and the experiments attractor when the simulations are in its steady state. The horizontal axis represents the density of the graph d (d varies form 0.002 to 0.012. We can see that the *Error* decreases as d increases its value.

of edges in the network will increase and, therefore, the average number of edges per node will be higher. This would correspond in the Regular Symmetric Graph to an increase in the number of edges per node D. The result of increasing d in this model are almost the same as increasing D in the Regular Symmetric Graph model. We can see the same results as in the previous model. As we increment the degree of the network d from 0.002 to 0.012 we see a decrease in the difference between the mathematical result and the simulation result.

We have run several experiments in a 100×100 ER-Random Graph with parameter values for $d = 0.0016$ and $p = 0.0025$, making one run for each value of b between 0 and 0.8 increasing b at 0.004 intervals. For each value of b we have run 45 experiments and calculated the average. The result of this average is shown in the Figure 6.12, in green dots. The red dots are the different expected fixed points of the mathematical model as the parameter b changes. We can see that in general, all the values in these averages are very close to the fixed points of the mathematical model. The average of the simulations is a bit lower than the mathematical solution in the middle points of the interval [0,0.4]. However, they seem to be very accurate when b is close to 0 or close to 0.4. For those middle points of the interval [0,0.4] we can still consider the mathematical model as an upper bound of the simulation result.

One of the results of the mathematical analysis was the stability of the fixed points of the model fixed points $r = 0$ and $r = 1 - \frac{b}{Ap}$. This analysis said that

- When $\ln \frac{1}{(1-p)^A} > b$, the fixed point $r = 0$ is unstable and the second fixed point, given by running the equation 5.1, (which can be approximated with $r = 1 - \frac{b}{Dp}$) is stable.
- When $\ln \frac{1}{(1-p)^A} < b$, the fixed point $r = 0$ is stable, and there are no other fixed points in the system.

That means that for values $p = 0.0025$ and $A = 16$ we would find that, if $b < \ln \frac{1}{0.9975^16} = 0.0400500835$, the fixed point $r = 0$ is unstable and $r = 1 - \frac{b}{Dp}$

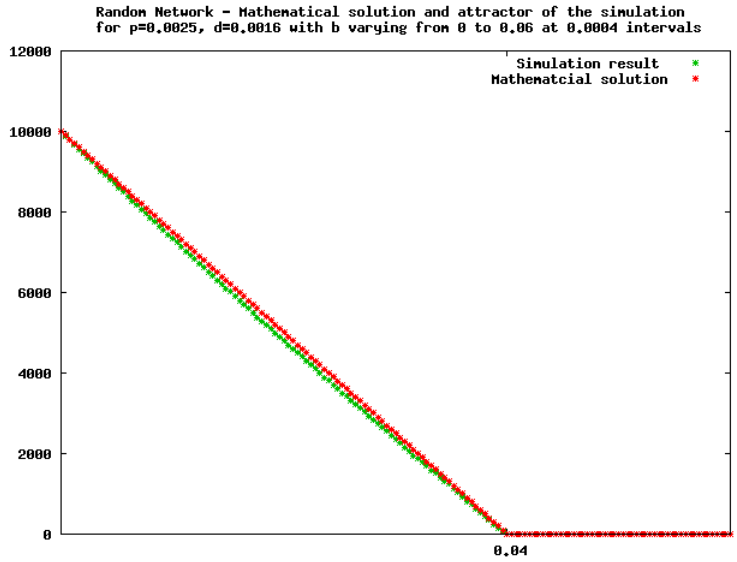


Figure 6.12: Results form a 10,000 ER-Random Graph model with d= 0.0016 and p=0.0025. We compare the attractor of the simulations results after a long run with mathematical solutions. The value of b changes from 0 to 0.06.

stable. Looking at the experiments results, the point $r = 0$ seems to be stable from about $b = 0.4$ as expected by the mathematical model.

In Figure 6.13 we can compare the results of this experiment in the two network topologies we have studied. We can see clearly that, although for small values of b the results for an ER-Random Graph model seem to drop faster than the Regular Symmetric Graph, over all, the results for an ER-Random Graph seem to be better as they get closer to the mathematical solution as b increases, while the results for a Symmetric Graph drop faster and don't get close to the mathematical solution for larger values of b. This is a surprising result because, although we are virtually using the same mathematical model, the ER-Random Graph is more complex than the Symmetric Graph. Also, in the mathematical model of the ER-Random Graph we were not considering the exact number of neighbour per node, but just an average, as in this graph different nodes have different number of neighbours. An explanation of this phenomenon may be the

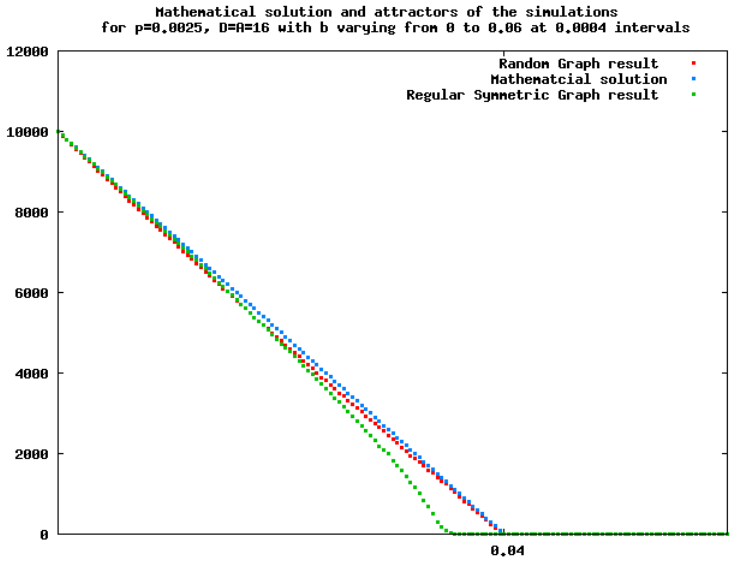


Figure 6.13: Comparing results form a 10,000 ER-Random Graph model with d= 0.0016 and p=0.0025, a 100×100 lattice network with D=16 and p=0.0025 and the mathematical solution. The value of b changes from 0 to 0.06.

following: We know that an ER-Random Graph has a short path connectivity (99) and that may help the process to be transmitted differently around the network (68), and it is in that sense perhaps more similar to a "perfect mixing" in which all individuals interact with each other with the same probability. Also, the Symmetric Graph has a high clustering coefficient (99). The proportion of red nodes around a node that is red itself is much higher than in the total population. Therefore, the number of white nodes in the immediate neighbourhood of a red node is strongly constrained. That makes the transmission process to be much slower, and the probability for the simulation to get into the absorbing state $r = 0$ is much higher than in a network with low clustering coefficient (94).

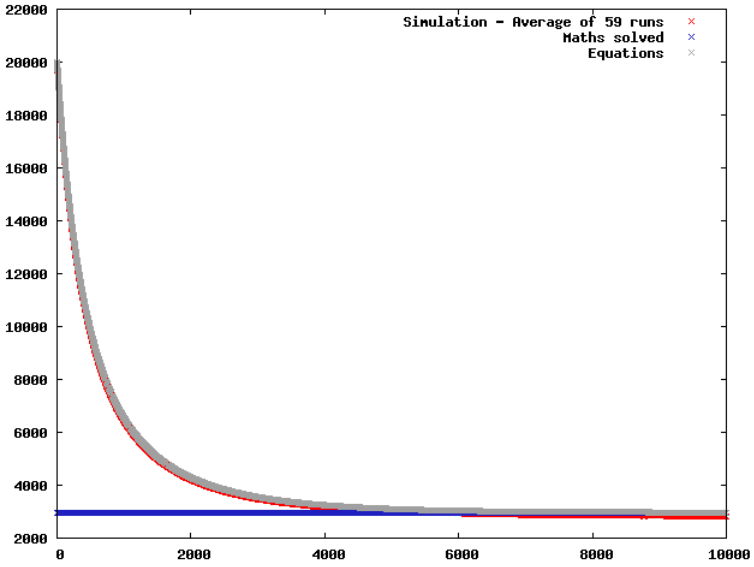
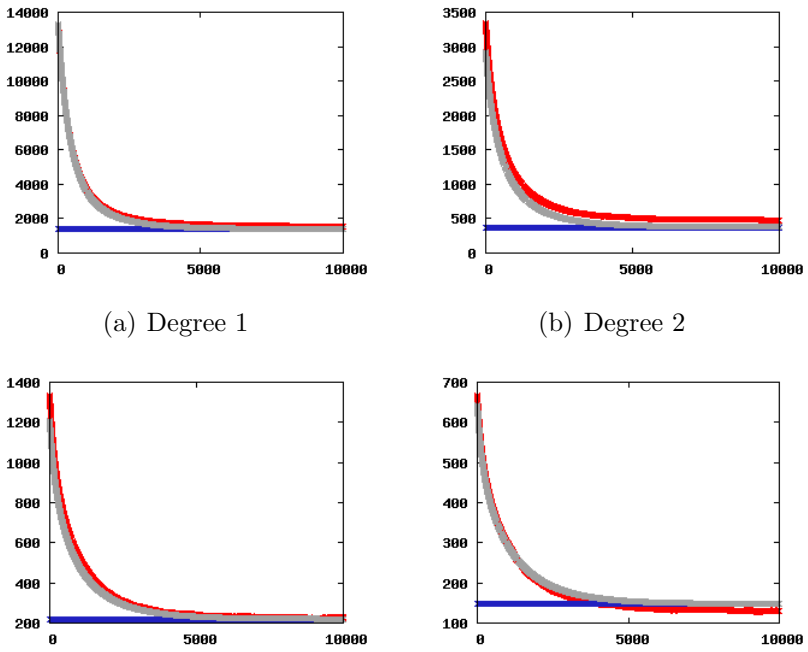


Figure 6.14: These graphs shows solutions for a scale free network with minimum number of edges being $c = 1$ and having a total of 20,000 nodes. The values of p and b are $p = 0.001$ and $b = 0.002$. The vertical axis represents the total number of red nodes in the system and the horizontal axis represents time. In the graph we compare 1.- the results of a simulation run over 10,000 time steps with an initial number of red nodes of 100%, 2.- the expected number of red nodes when the system is in its steady state, given as a solution of the analysis of the First Mathematical Heterogeneous model, and 3.- the results of dynamical equations given by the First Mathematical Heterogeneous model.



(a) Degree 1

(b) Degree 2

(c) Degree 3

(d) Degree 4

Figure 6.15: These graphs shows solutions for a scale free network with minimum number of edges being $c = 1$ and having a total of 20,000 nodes. The graphs represent the same as in Figure 6.14 but specifically for the total number of red nodes of degrees 1, 2, 3 and 4.

6.3 Scale-Free networks - Experiments and Results

In this experiment we create a network of 20,000 nodes using the *rich get richer* algorithm and adding one node and one edge at a time. The probability to transmit the red colour is p=0.001 and the probability to go back to white colour is 0.002. We run this experiment for 10,000 time steps.

We also run the mathematical equations that we analysed for this model (5.16) given by the extension of the First Mathematical Heterogeneous model, so we can compare how the number of red nodes changes over time in the two cases. In Figure 6.14 we have the simulation run in red colour and the mathematical equations running in grey colour. We also have the expected asymptotic value of the number of red nodes in the system (solution of the mathematical equations 5.21) in blue. We can see the results of the simulation are very close to the mathematical model. In Figure 6.15 we have four different subgraphs showing the same results for four different degrees. We can see that the mathematical model's results are very accurate in both the run of the mathematical equations, and the calculation of the asymptotic value of the red nodes in the system.

In this experiment the scale-free network was created by adding 5 edges for each new node, following the *rich get richer* rule. The mathematical model is slightly different, but it is still very accurate in its prediction. In Figure 6.16 we present the results of a simulation of a scale-free network of 20,000 nodes, with p=0.001 and b=0.002. The initial number of red nodes is 20,000 and the simulation is run for 10,000 time steps. In Figure 6.17 we show the results of the same experiment for nodes of degrees 5, 6, 7 and 8. We compare these results with the First Mathematical Heterogeneous model. We can see how the mathematical equation and the mathematical solution of the asymptotic value of the system are very accurate.

We have also tested what is the effect of changing the values b and p leaving

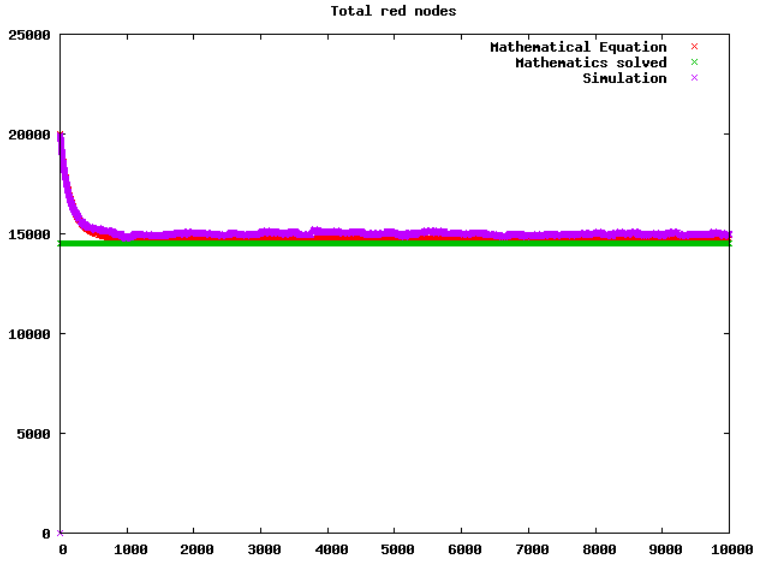
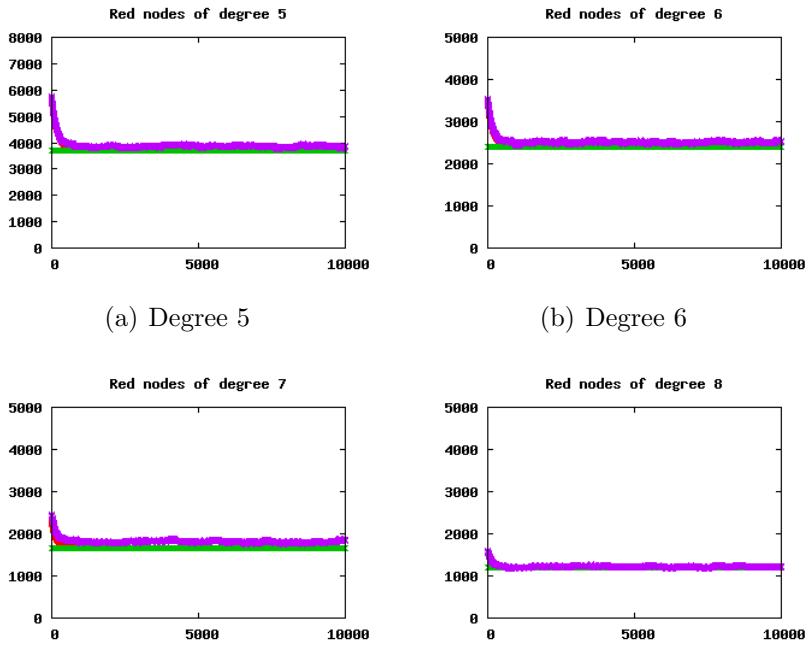


Figure 6.16: These graphs shows solutions for a scale free network with minimum number of edges being $c = 5$ and having a total of 20,000 nodes. The values of p and b are $p = 0.001$ and $b = 0.002$. The vertical axis represents the total number of red nodes in the system and the horizontal axis represents time. In the graph we compare 1.- the results of a simulation run over 10,000 time steps with an initial number of red nodes of 100%, 2.- the expected number of red nodes when the system is in its steady state, given as a solution of the analysis of the First Mathematical Heterogeneous model, and 3.- the results of dynamical equations given by the First Mathematical Heterogeneous model.



(a) Degree 5

(b) Degree 6

(c) Degree 7

(d) Degree 8

Figure 6.17: These graphs shows solutions for a scale free network with minimum number of edges being $c = 1$ and having a total of 20,000 nodes. The graphs represent the same as in Figure 6.16 but specifically for the total number of red nodes of degrees 5, 6, 7, and 8.

$m = b/p$ (and therefore the estimated number of red nodes) constant. We have seen that, unlike for the Symmetric Graph and ER-Random Graph models, in the case of scale-free networks this changes do not affect the behaviour of the system. We have tested this effect in both networks created by adding one edge per node and networks created by adding five edges per node.

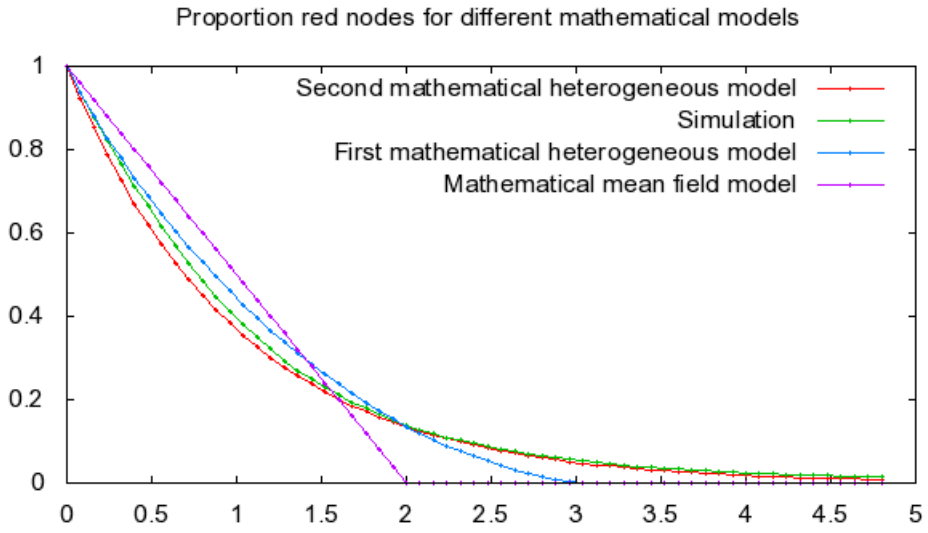


Figure 6.18: Results of a scale free model of 20,000 nodes formed by adding 1 node at a time (c=1), when $p = 0.001$. The vertical axis represents the proportion of red nodes in the model when it is in its steady state, and the horizontal axis represents the value of $m = b/p$ which ranges from 0 to 4.8 (as b ranges from 0 to 0.0048). In this graph we compare three different mathematical models and the simulation results.

In Figure 6.3 we can compare the results for three different mathematical models and the simulation results. The models represent a network formed by adding one edge at a time. The mathematical models are 1.- the mathematical mean-field model, 2.- the extension of the First Heterogeneous model, and 3.- the approximation of the Second Heterogeneous model used by Pastor-Satorras and Vespignani in (73) and (72) (we have been unable to solve analytically or numerically the so called second heterogeneous model, or the extension of this second

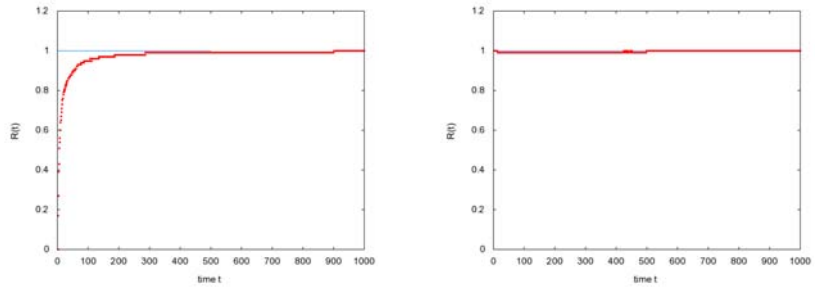
heterogeneous model). We can see that the mean-field mathematical model is not a good estimation. The first homogeneous mathematical model is a good approximation of the disease dynamics in this network, only for low values of $m = b/p$. We can see how the stable fixed point of this mathematical model reaches 0 at a value around 3. However, experiments show that the simulation does not reach the fixed point 0. The last mathematical model comes to an agreement with this phenomena.

6.4 Effective Reproductive Rate

In this section we present the results of calculating the average number of transmissions in a lifetime of the red nodes in different network structures, and for different parameters p and b. In order to calculate this value we look at each red node during its whole lifetime, i.e. until it becomes white again, counting how many times a red nodes transmits the colour red to its white neighbours. At each time step we look at all those nodes that go back to white state, and calculate the average number of transmissions that those nodes have made during their lifetime.

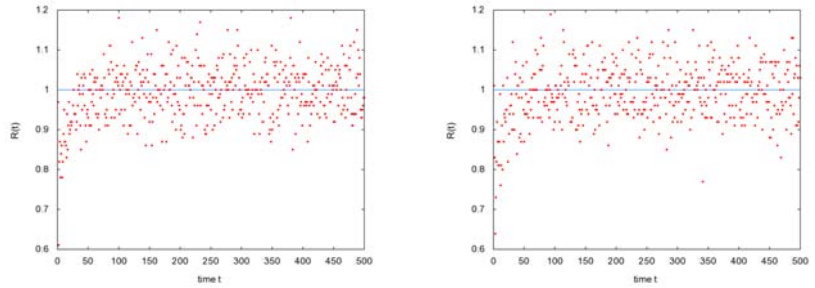
In Figures 6.19 and 6.20 we show the results obtained for running simulations in an ER-Random Network of size 10,000 and density $d = 0.0016$ (Figure 6.19) and a Symmetric Graph of 100×100 nodes, each node having 16 neighbours (Figure 6.20). The initial number of red nodes in both cases is 20%. Also in both cases we have tried different parameters p=0.02 and b=0.05 and p=1, b=0.05. In all cases, even using the extreme values $p = 1$ the average number of transmissions per node in a lifetime, after the simulation has run a certain time, is 1 (or very close to it) in all cases.

In Figure 6.21 we show the results obtained by running a simulation of a scale-free network built by adding 2 edges per new node following the *rich get richer* rule. We can see 4 Sub-figures showing results of different parameter values. In



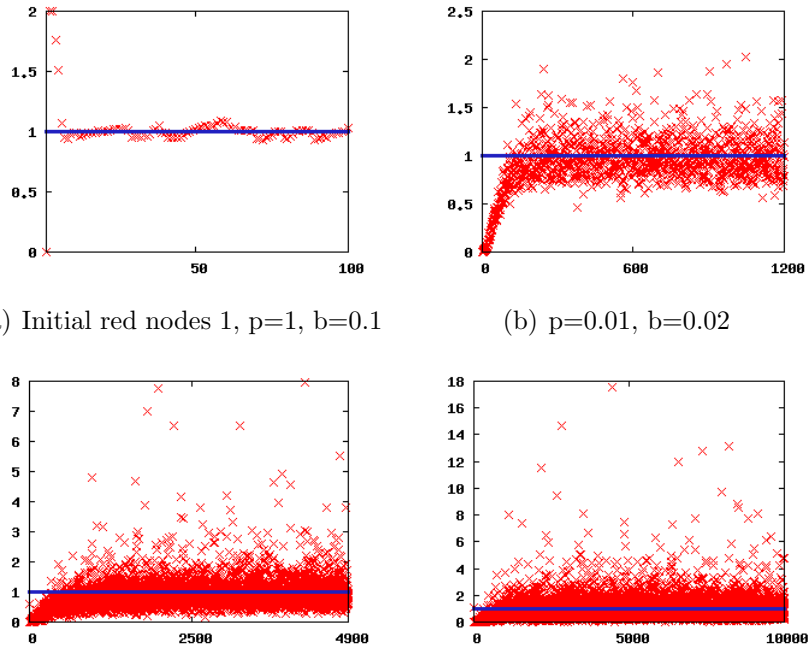
(a) p=0.02, b=0.05 (b) p=1, b=0.05

Figure 6.19: Simulation run in a ER-Random Network of size 10,000 and density $d = 0.0016$. Initial number of red nodes is 20%. The vertical axis represents the average number of transmissions for each red node that has been just transformed to white. The horizontal axis represents time. The red dots represent the results of the simulation at each time step and the blue line is the average of these results over the last 500 steps.



(a) p=0.02, b=0.05 (b) p=1, b=0.05

Figure 6.20: Symmetric Graph of size 100×100 with 16 number of neighbours per node. Initial number of red nodes is 20%. The vertical axis represents the average number of transmissions for each red node that has been just transformed to white. The horizontal axis represents time. The red dots represent the results of the simulation at each time step and the blue line is the average of these results over the last 500 steps.



(a) Initial red nodes 1, p=1, b=0.1

(b) p=0.01, b=0.02

(c) p=0.01, b=0.002

(d) p=0.001, b=0.002

Figure 6.21: Reproductive rate of the simulation for different ps and bs in a Scale Free network with 500,000 nodes and c = 2. Initial number of red nodes is 20%. The vertical axis represents the average number of transmissions for each red node that has been just transformed to white. The horizontal axis represents time. The red dots represent the results of the simulation at each time step and the blue line is the average of these results over the last 50% of the total run time.

the first sub-figure we set up the model so that there is only one red node in the system. The probability to transmit the red colour red p is 1, and the probability to go back to white is b=0.1. As we can see in the graph, the system starts with an average of 2 transmissions per node each time step, but then soon it goes down and oscillates around the value 1. In the sub-figures 6.17 we show in blue the average number of transmissions in a lifetime. This average is created using the last 50% steps in the simulation run.

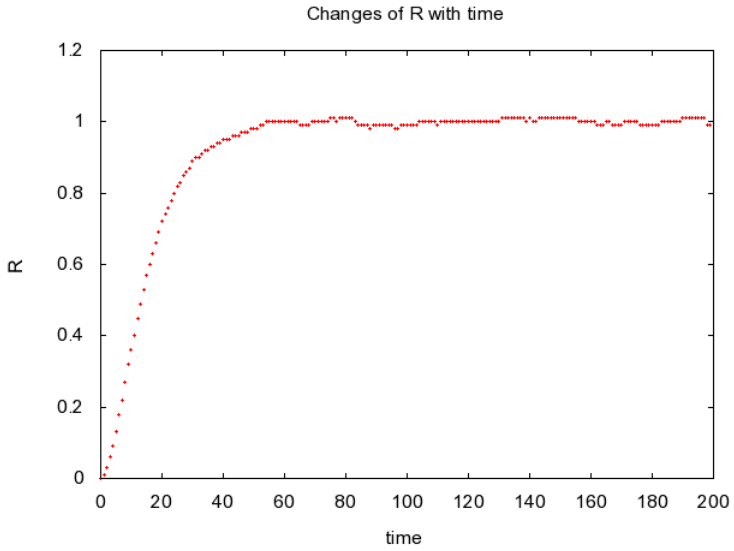


Figure 6.22: Reproductive rate of the simulation for different ps and bs in a Scale-Free network with 500,000 nodes, and c = 1, b=0.001 and p=0.001. Initial number of red nodes is 20%. The vertical axis represents the average number of transmissions for each red node that has been just transformed to white. The horizontal axis represents time. The red dots represent the results of the simulation at each time step, clearly approaching 1.

For the scale-free network models we also run a simulation with 500,000 nodes created by adding 1 edge per node. The probability to transmit the colour red is $p = 0.001$ and the probability to go back to white is $b = 0.001$. We run this experiment for 20,000 time steps, although we have only recorded the data every 100 steps. The initial number of red nodes in the simulation was 20%.

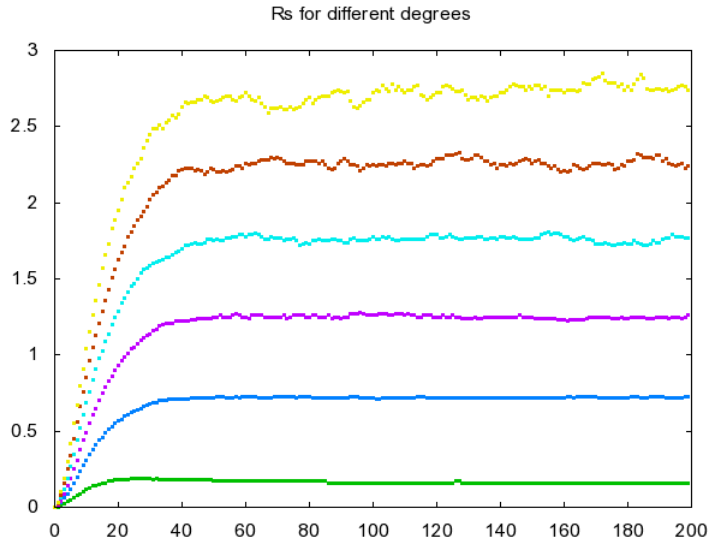


Figure 6.23: Different values R_s for a run made in a scale-free network with minimum number of edges = 1. The network has 500,000 nodes, b=0.001 and p=0.001. The green line represents the reproductive rate for nodes of degree 1, R_1, the blue line represents the reproductive rate for nodes of degree 2, R_2, and so on.

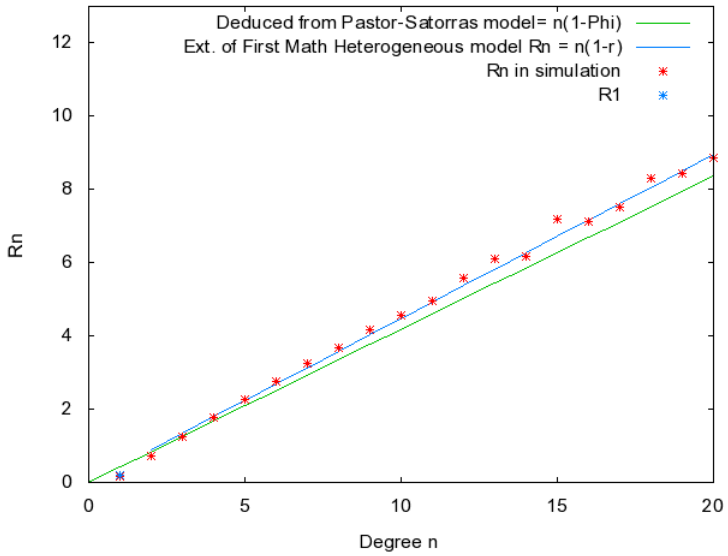


Figure 6.24: Asymptotic values for each R_n, in red, compared to the estimated value of this R_s given by 1.- the Extension of the First Heterogeneous model in blue and 2.- an estimation based on the solutions of the analytical approximation to the Second Mathematical Heterogeneous model used by Pastor-Satorras and Vespignani, in green.

The results of this experiments show how the average number of transmissions per red node during its lifetime, R, approaches 1 as the simulation runs (see Figure 6.22). We can also see in Figure 6.23 how nodes of different degrees transmit the colour to different number of white nodes (in average) Figure 6.23.

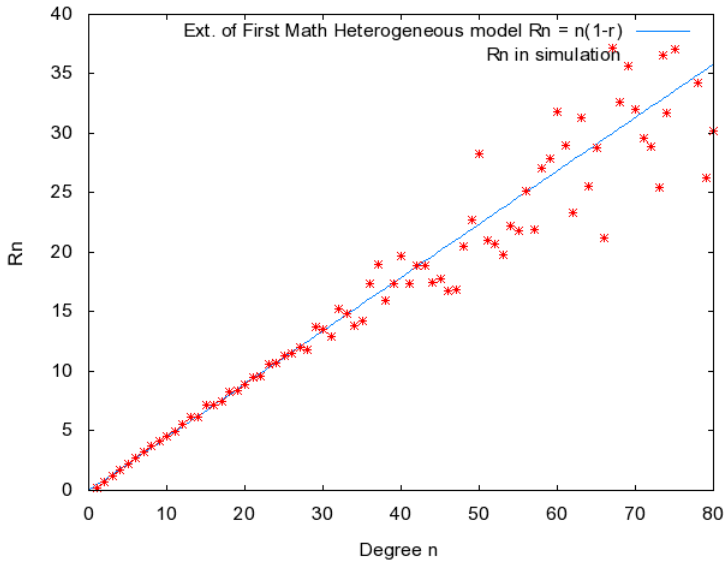


Figure 6.25: Asymptotic values for each R_n, in red, compared to the estimated value of this R_s given by the Extension of the First Heterogeneous model, in blue.

In Figure 6.24 we can see a comparison of two estimations of R given by two mathematical models. The firs one, in blue, is deduced from our Extension of the First Mathematical Heterogeneous model and the second, in green is deduced from the model created by Pastor-Satorras. We can see how, in this case, the estimation made by the mathematical models is very accurate although our estimation gets closer to the results in the simulation, even for degree 1, which in our model it follows a different trend from the rest of the degrees. We can also see in Figure 6.25 that, although higher degrees tend to have less predictable value R_n, they still follow a trend very close to our mathematical estimation.

We make this same experiment for different probabilities p to transmit the colour red and for different probabilities b to go back to white state. In all of

them the effective reproductive rate R is one after the simulation has run for enough time. All these results come to an agreement with the value $R = 1$ suggested by (94).

6.5 Chapter Summary

In this chapter we have run several experiments to test the estimations made by different mathematical models of the SIS dynamics implemented in different network topologies.

In Section 6.1 we run experiments of a 2 dimensional lattice forming a 100×100 grid of nodes with a torus shape. We calculated the fixed points of the system and analysed their stability. The experiments show that the mathematical estimation is really good when probabilities b and p are very small, but it is not so accurate when they get higher values. We have also observed that if the number of neighbours per node D is high enough, the estimation is again very accurate, but not so much as D gets smaller. We have finally contrasted the stability of the fixed points of the mathematical model with the behaviour of the simulation. We have observed that the simulation tends to have $r = 0$ as steady-state for values of $m = b/p$ lower than estimated by the mathematical analysis.

In Section 6.2 we run experiments with graphs of 10,000 nodes and density $d = 0.0016$. We have observed that the mathematical model is very accurate when probabilities b and p are very small and looses accuracy as p and b increase in value. We have also seen that the mathematical estimation of red nodes is much better for large values of d than for lower values. We have then compared the capacity to predict the steady-state of the model as we fix all parameters except b and the stability of the fixed points. We have observed that the mathematical model is very accurate in this case. Surprisingly the mathematical model is much more accurate in the ER-Random Graph than in the Symmetric Graph. The reason for this may be the contrast of short path connectivity property of the

ER-Random Graph model, which accelerates the spread of the disease, and the high clustering of the Symmetric Graph model which makes it slower.

In Section 6.3 we first run a simulation of a network with minimum number of neighbours per node being 1. We compare the results with the estimations given by the Extension of the First Heterogeneous Mathematical model. We can see that the estimations are very accurate both for the total number of red nodes in the network and for the number of red nodes of different degrees 1, 2, 3, and 4. We then run experiments for a network with minimum number of neighbours per node being 5 and compare it with the First Heterogeneous Mathematical model obtaining again very accurate results. We have then run an experiment in which we fix all parameters except b which ranges from 0 to 0.0045. Here we compare three mathematical models and the simulation in order to investigate the predictive power of each of them to estimate the total number of red nodes and to observe the stability of the fixed points. We can see that the mean-field approximation is the less accurate of all. We can also see that the Extension of the First Mathematical model is accurate when b is very low, but it looses its predictable power for values of $b/p > 2$. The best estimation in terms of the total number of red nodes in the network is given by the approximation used by Pastor-Satorras and Vespignani (72; 73), mainly for high values of b/p. We can see that the simulation seems to remain in an endemic state for any value of $m = b/p$, as predicted by Pastor-Satorras and Vespignani's model and also predicted by the Extension of the Second Mathematical Model.

In Section 6.4 we run experiments to investigate the estimation of the effective reproductive rate R in the simulation. We can see that in all cases, regardless of the topology of the network, and regardless of the probabilities p and b, the value of R reaches 1 when the simulation has run for long enough. In the case of scale-free network we have also observed that the effective reproductive rate for different degrees R_n has different values, so nodes of higher degree transmit the disease to more neighbours. We also compare how the estimated value for different R_n follows the trend of the simulation results, and it is a better approximation

than the estimation we made using the expectations from Pastor-Satorras and Vespignani's model.

6.6 Conclusions

We have been able to create very accurate mathematical models of the SIS dynamics implemented in different networks and formulate very precisely the properties of the systems. We have been able to make an analysis of the stability of the fixed points of most systems, which has helped us understand the behaviour of the models in its steady-state. We have seen that the approach of grouping the nodes according to its number of neighbours has great advantages in calculating the overall number of red nodes in the network, as well as the individual number of red nodes for each degree. We have also been able to understand how network properties such as short connectivity path or high clustering influence the transmission process in the network.

During this process we have been able to appreciate the increment of complexity in the mathematical models as the topologies became more complex, specially for the scale-free network models. We have been unable to solve analytically some of our mathematical systems such as our two mathematical heterogeneous models and the respective extensions 5.9, 5.16, 5.23 and 5.30. We have only been able to solve numerically (and for all relevant parameter values) the Extension of the First mathematical model 5.9.

We have shown empirically and mathematically that, regardless of the probability to transmit the red colour p or the probability for a red node to go back to white state, the average number of transmissions that a red node will do in its lifetime is just 1. This happens in any of the network topologies studied in this thesis. This is a remarkably surprising finding and not easy to understand. We have also shown empirically and mathematically that in a scale-free network, red

nodes of different degrees transmit its colour to different number of nodes (on average).

In the process of building these models we have again benefited from taking a step by step approach mainly in the development and understanding of the mathematical models for the scale-free networks.

Part IV

Conclusions

Chapter 7

Conclusions and Further Work

7.1 Conclusions

In this thesis we have presented models and studied the properties of two different systems using both simulation and mathematical analysis. In the first part of this thesis we implemented a model of vector-borne disease with heterogeneous landscape, adding movement to the population in the model, in order to understand how the heterogeneity and the movement of the population has an effect in the spread of the disease in the population as a whole. In terms of the mathematical analysis we have successfully used the Heuristic Random Search approach in order to understand the macro-behavior of the model focusing on its steady-state and the stability of the fixed points of the mathematical system. In the second part of this thesis we have implemented a SIS model on different network topologies studding how the this topologies have an effect in the transmission of the process in the network. We have also been able to make very accurate predictions of the steady-state of the network models and understand how different characteristics of the system affect this process such as clustering, short path connectivity, and degree distributions. We have also shown empirically and mathematically that in our models the *effective reproductive rate* is 1. We have also shown empirically

and mathematically that the number of transmissions made by a red individual in a scale-free network is directly proportional to the number of neighbours this individual has in the network.

In both cases, vector-borne disease model and SIS model on a network, we have been able to create very accurate mathematical models formulating very precisely the properties of the systems. We have been able to make an analysis of the stability of the fixed points of most systems, which has helped us understand the behaviour of the models in its steady-state. However, we have also observed that the complexity of the mathematical models increases rapidly with the complexity of the model. In some cases we have been unable to solve the motion equations analytically and/or numerically and in other cases we have been unable to carry out a mathematical analysis of the stability of the fixed points.

In both models it has been clear that the strategy of carrying out an incremental implementation with increasing complexity of the models has great advantages for the understanding of the more complex models. By creating the simulations and the mathematical models in parallel we have found that having two parallel representations of the system has been crucial for the understanding of the behavior of the model. Both approaches have been used to verify the outcome of one another and to give an insight of what they should look like.

In this thesis we have also presented the *Lemma of the Fixed Points* that was found to be very useful in the analysis of the stability of fixed point of many of our models.

7.2 Further Work

There are many possible areas in which the work in this thesis could be developed further. In the vector-borne disease model it would be interesting to investigate

how immunity would affect the dynamics of the disease. It would also be of great interest how vaccination may modify the spread of the disease. It can be a vaccination targeted to only one sort of patch, or to all of them with different intensities, given that vaccination is a restricted resource in every country.

Regarding the models of SIS dynamics on a network, it would be interesting to investigate other characteristics present in social networks such as clustering, and how that affects the dynamics of the system in general, also paying also attention to the *basic reproductive rate* and the *effective reproductive rate* of the transmitted process. It would also be very desirable to extend the mathematical models in parallel, trying to incorporate the effects of the clustering coefficient of a network.

Another very interesting question would be to investigate why an epidemic, in a scale-free networks with exponent $2 < \gamma \leq 3$, is always endemic when it is not the case in other topologies such as Symmetric Graphs or ER-Random Graphs. It would also be interesting to create a scale-free networks with exponent $3 < \gamma$ using a modification of the *preferential attachment* rule and observe if, in this case, the process is not endemic as predicted by the mathematical analysis of the fixed points of the Extension of the Second Heterogeneous model.

Another interesting aspect of the SIS process would be to add a constant probability that a white node gets red spontaneously (i.e. without transmission from a neighbour). This would correspond to someone getting a disease from the environment (e.g. the first case of swine flu). Or in the "populating the galaxy" scenario, it corresponds to the probability that a planet evolves intelligent life by itself.

I find very interesting the idea of creating a model that involves both the metapopulation approach and the contact network approach. The idea would be to divide a population into a set of sub-populations located in different patches (as in our first model) and, in each patch, model the sub-population as a network (as in the second model). That would give us a better understanding of how

movement of people affects the transmission process around a set of local social networks. However, we would first need to understand how a transmission process, such as the SIS dynamics, works in a dynamic network structure in which we can temporarily remove and add nodes from the network.

In terms of mathematics, it would be desirable to extend the *Lemma of the Fixed Points* stated in Appendix A to a multi dimensional motion function $G : \mathbb{R}^n \to \mathbb{R}^n$. This Lemma could be used for the analysis of the stability of the fixed points of the vector-borne disease models presented in the first part of this thesis.

Appendix A

Lemma of the Fixed Points

In this thesis we have presented some models of dynamic systems for which we have been able to write a master equation describing the dynamics of the system. The master equations for some of our models are of the form $i = f(i)$ where f is an operator from $[0, A]$ to $[0, A]$ with $A > 0$, in some cases $A = 1$. Our interest in this thesis is to find out about the fixed points of these systems and their stability (note that 0 is always a fixed point in the systems described by f). Lemma 4 presented below states that if the function f satisfies certain requirements then:

1 The system has at most two fixed points in $[0, A]$ one of which is 0.

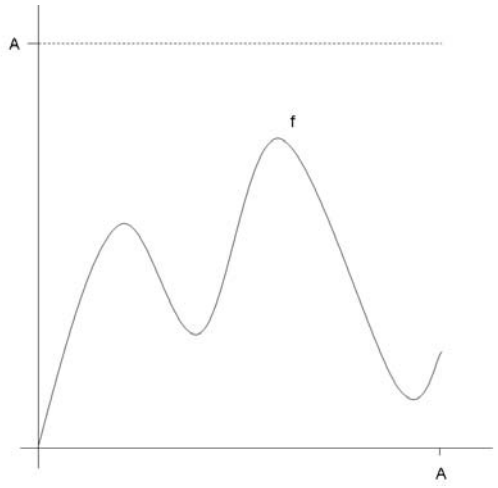
2 If the fixed point 0 is stable then there are no other fixed points in $[0, A]$.

3 If the fixed point 0 is not stable then there is one more fixed point and it is stable.

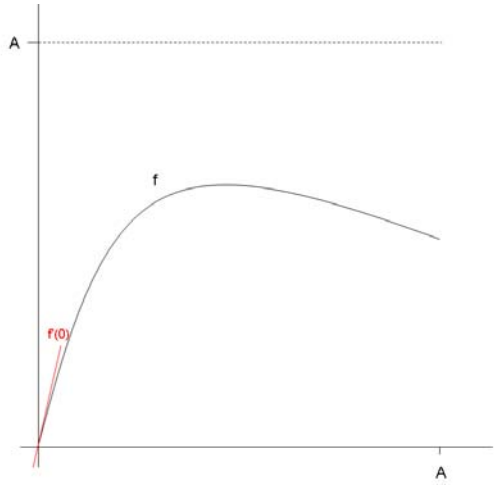
Below I explain this lemma in plain English showing some graphs as I explain it. After that I enunciate and prove this lemma mathematically.

A. LEMMA OF THE FIXED POINTS

Let us suppose that f is continuous in $\mathbb{R}$. Let's suppose that $f(0) = 0$, $0 \leq f \leq A$ in $[0, A]$ and $f(A) < A$.

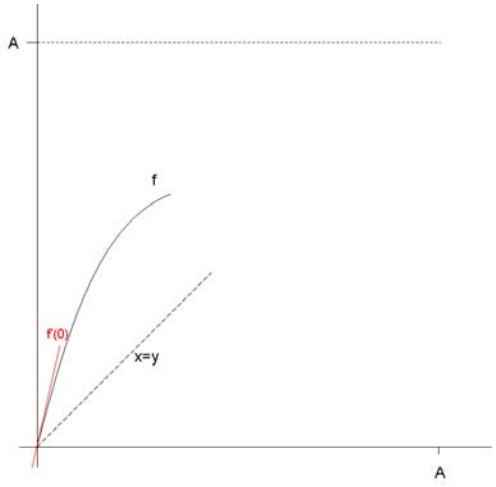


Let us also suppose that f' and f'' exist in $[0, A]$ and $f'' < 0$ in $[0, A]$. That makes the function concave downward in the interval.

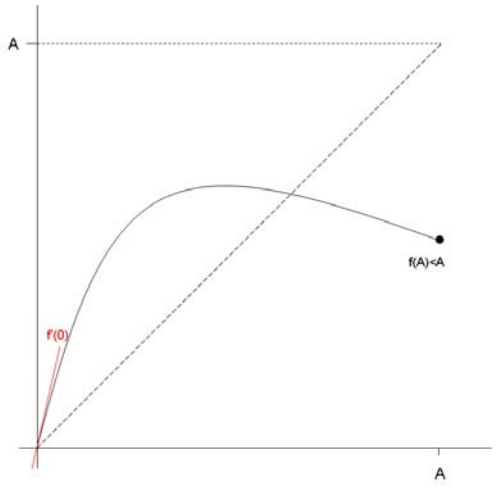


Then we have:

1 If $f'(0) > 1$ the functions grows faster than the line $x = y$ at 0, so the function f must be above the line $x = y$ for positive values close to 0.

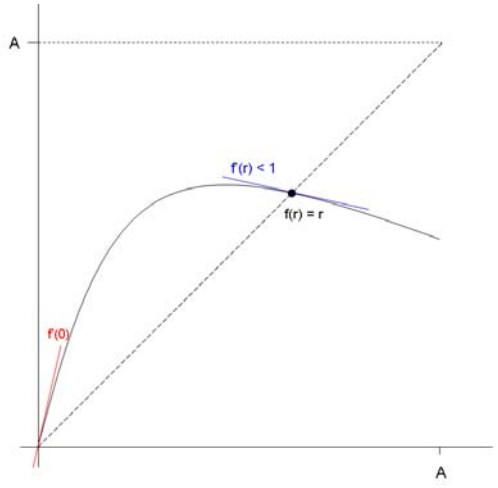


We know that $f(A) < A$, so function f is below the line $x = y$ at A, therefore f must cross the line $x = y$ at some point in $(0, A)$. Because f is concave downwards in the interval, the function can only cross the line $x = y$ once.

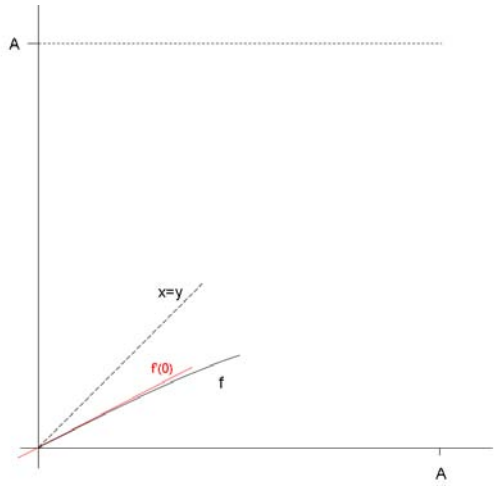


Let r be the point where f crosses the line $x = y$. We say that r is a fixed point for f. Looking at the graph it is easy to realise that the growth of f at the crossing point r must be slower than the growth of the line $x = y$. That means that the derivative of f at r must be lower than 1. Therefore if f satisfies that $f'(r) > -1$ then we would have that $|f'(r)| < 1$ and therefore

r would be a stable fixed point for f.

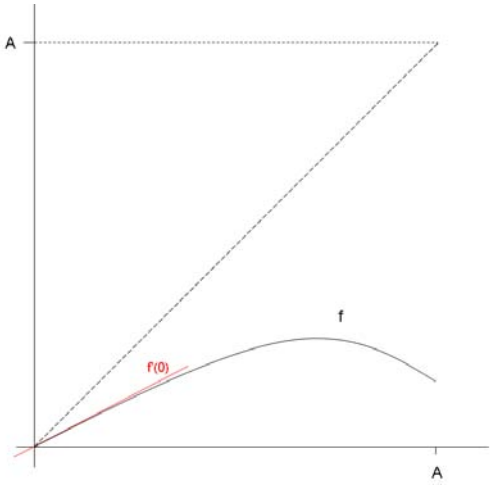


2 If $f'(0) \leq 1$ the growth of f at 0 may be the same as the line $x = y$ or slower. However, because $f'' < 0$ the values of f' will be less than 1 for positive points close to 0 (though not necessarily at 0). That means that as soon as x start increasing the growth of $f(x)$ is slower than the growth of the line $x = y$ around 0 so f will be below the line $x = y$ for positive values close to 0.



Because f is concave downwards in $[0, A]$ and f grows slower than $x = y$

at 0, then f will grow slower than $x = y$ in the whole interval $[0, A]$ and therefore f will be below $x = y$ in $(0, A]$.



This would mean that f does not have fixed points in $(0, A]$.

In mathematical terms we would say:

Lemma 4. ***Lemma of the Fixed Points:*** *Let us suppose that f is continuous in $\mathbb{R}$. Let's suppose that $f(0) = 0$, $0 \leq f \leq A$ in $[0, A]$ and $f(A) < A$. Let us also suppose that f' and f'' exist in $\mathbb{R}$ and $f'' < 0$ in $[0, A]$. That makes the function concave downward in the interval.*

1 If $f'(0) > 1$ there is one and only one point r in $(0, A)$ such that $f(r) = r$. For this point r we get $f'(r) < 1$, so if f' satisfies that $f'(r) > -1$ we would

have that r is a stable fixed point of f.

2 If $f'(0) \leq 1$ then there is not a point r in $(0, A)$ such that $f(r) = r$. I.e. there is not a fixed point of f in $(0, A)$.

For the proof of this theorem we use the Mean Value Theorem, the Bolzano's Intermediate Value Theorem and another Lemma. These three are stated below. The proof for the two theorems can be found in any Mathematical Analysis book. The Lemma and its proof can be found at (17) page 174.

Theorem 1. ***Mean Value Theorem:*** *Suppose that f is continuous on a closed interval $I := [a, b]$, and that f has a derivative in the open interval (a, b). Then there exists at least one point c in (a, b) such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

Theorem 2. ***Bolzano's Intermediate Value Theorem:*** *Let I be an interval and let $f : I \to \mathbb{R}$ be continuous on I. If $a,\ b \in I$ then for any $k \in \mathbb{R}$ such that $f(a) < k < f(b)$ there is a point $c \in I$ between a and b such that $f(c) = k$.*

Lemma 5. *(from (17), page 174). Let $I \subseteq \mathbb{R}$ be and interval, let $f : I \to \mathbb{R}$, let $c \in I$, and assume that f has a derivative at c. Then:*

a) If $f'(c) > 0$ then there is a number $\delta > 0$ such that $f(x) > f(c)$ for $x \in I$ such that $c < x < c + \delta$.

b) If $f'(c) < 0$ then there is a number $\delta > 0$ such that $f(x) > f(c)$ for $x \in I$ such that $c - \delta < x < c$.

Proof. We have to prove two parts:

1 Let us have $f'(0) > 1$

- First we will prove that $\exists a \in (0, A)$ such that $f(a) > a$:

  We have that $f'(0) > 1$. Therefore, using Lemma 5 we see that there exists $\delta > 0$ such that if $a \in (0, \delta)$ then $f(a) > a$.

- Next we prove that $\exists\ r$ with $0 < a < r < A$ such that $f(r) = r$:

  We have $0 < a < A$ with $f(a) > a$ and $f(A) < A$. Let's define $g(x) = f(x) - x$. g is continuous in $[0, A]$ and $g(a) > 0 > g(A)$. For *Bolzano's Theorem* (2) we have that $\exists r$ with $a < r < A$ such that $g(r) = 0$ and therefore we have that $\exists r \in (0, A)$ with $f(r) = r$.

- Now we prove that r is unique, i.e. that $\exists! r' \neq r$ in $(0, A)$ such that $f(r') = r'$. We prove it by contradiction:

  We know that $f'' < 0$ in $[0, A]$, therefore f' is strictly decreasing. This means that for two different points x and x' in $[0, A]$ then $f'(x)$ and $f'(x')$ cannot be equal.

  Let us assume that r with $f(r) = r$ is not unique in $[0, A]$ so $\exists r' \neq r$ in $(0, A)$ such that $f(r') = r'$. Let's suppose $0 < r < r' < A$. For the *Mean Value Theorem* (1) we know that $\exists t \in (0, r)$ such that $f'(t) = \frac{f(r)-f(0)}{r-0} = 1$ and also $\exists t' \in (r, r')$ such that $f'(t') = \frac{f(r')-f(r)}{r'-r} = 1$. That makes $f'(t)$ and $f'(t')$ equal, which leads to a contradiction.

Therefore there is only one r in $(0, A)$ that satisfies $f(r) = r$, and there is $t \in (0, r)$ such that $f'(t) = 1$.

- Now we prove that $|f'(r)| < 1$:

  We proved above that there is one $t \in (0, r)$ such that $f'(t) = 1$. Because $f'' < 0$ then f' is strictly decreasing in $[0, A]$. Therefore, as $t < r$ then $1 = f'(t) > f'(r)$. By hypothesis we know that $f'(r) > -1$ and therefore $|f'(r)| < 1$.

2 Let us have that $f'(0) \leq 1$.

We prove that $\nexists r \in (0, A)$ with $f(r) = r$ and we prove it by contradiction:

We know that $f'' < 0$ in $[0, A]$ which makes f' strictly decreasing. Therefore for any $0 < t < A$ we get $1 \geq f'(0) > f'(t)$. This means that there cannot be a number $t \in (0, A]$ with $f'(t) = 1$.

Let us assume there exists r in $(0, A)$ such that $f(r) = r$. We know that $f(0) = 0$. Using again the *Mean Value Theorem* (1) we have a $t \in (0, r)$ such that $f'(t) = \frac{f(r)-f(0)}{r-0} = 1$. This leads to a contradiction, therefore $\nexists r \in (0, A)$ with $f(r) = r$.

□

Bibliography

[1] The Transims Project Website. http://transims-opensource.net/. [Online; accessed 03-March-2010]. 36

[2] L. A. Adamic and B. A. Huberman. Power-Law Distribution of the World Wide Web. *Science*, 287, 2000. 27

[3] Réka Albert and Albert-László Barabási. Statistical Mechanics of Complex Networks. *Reviews of Modern Physics*, 74(1):47–97, January 2002. 24, 25

[4] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the World-Wide Web. *Nature*, 401:130–131, September 1999. 24, 26, 27

[5] L. A. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of Small-World Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149–11152, October 2000. 26, 27

[6] Roy M. Anderson and Robert M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford Science Publications, 1991. 36, 118

[7] R. Aspinall and D. Pearson. Integrated Geographical Assessment of Environmental Condition in Water Catchments: Linking Landscape Ecology, Environmental Modelling and GIS. *Journal of Environmental Management*, 59(4):299–319, 2000. 35

[8] Robert Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, August 1997. 4

[9] Raul Bagni, Roberto Berchi, and Pasquale Cariello. A Comparison of Simulation Models Applied to Epidemics. *Journal of Artificial Societies and Social Simulation*, 5, 2002. 6

[10] M. Baguette and G. Mennechez. Resource and Habitat Patches, Landscape Ecology and Metapopulation Biology: A Consensual Viewpoint. *OIKOS*, 106(2), 2004. 22

[11] Philip Ball. Utopia Theory. *Physics World*, 16(10):29–33, October 2003. 4

[12] Michael Banton. *The Social Anthropology of Complex Societies*, page 156. Tavistock, London, 1966. 4

[13] Yaneer Bar-Yam. *Dynamics of Complex Systems*. Perseus Books, 1997. 3, 4

[14] A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the Social Network of Scientific Collaboration. *Physica A*, 311:590–614, 2002. 27

[15] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, October 1999. 26, 94, 103, 111, 116, 127

[16] Rocío Gómez Bardón. Modeling Complex Systems: Movement of Crowds in Cities. Master's thesis, School of Compuer Science, University of Birmingham, December 2003. 45, 51

[17] Robert G. Bartle and Donald R. Sherbert. *Introduction to Real Analysis*. John Wiley and Sons Inc, September 1999. 176

[18] Eric D. Beinhocker. *Origin of Wealth: Evolution, Complexity, and the Radical Remaking of Economics*. Harvard Business School Press, June 2006. 4

[19] N. Boccara and K. Cheong. Critical Behaviour of a Probabilistic Automata Network SIS Model for the Spread of an Infectious Disease in a Population of Moving Individuals. *Journal of Physics A Mathematical General*, 26:3707–3717, August 1993. 17, 43, 100

[20] Peter C. Bull and Kevin Marsh. The Role of Antibodies to Plasmodium Falciparum-Infected-Erythrocyte Surface Antigens in Naturally Acquired Immunity to Malaria. *Trends in Microbiology*, 10(2):55–58, 2002. 42

[21] Jean-Philippe Cointet and Camille Roth. How Realistic Should Knowledge Diffusion Models Be? *Journal of Artificial Societies and Social Simulation*, 10(3):5, 2007. 6, 16, 90

[22] Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron, and Alessandro Vespignani. Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. *PLoS Medicine*, 4(1):e13, 01 2007. 22

[23] Paul Davidsson. Agent Based Social Simulations: A Computer Science View. *Journal of Artificial Societies and Social Simulations*, 5(1), 2002. 5

[24] Dominic Kwiatkowski. *Management of Severe and Complicated Malaria. A Practical Handbook: H. M. Gilles. Geneva: World Health Organization, 1991.*, volume 87. 1993. 42

[25] Bruce Edmonds. Review of Philp Ball: Critical Mass: One Thing Leads to Another. *The Journal of Artificial Societies and Social Simulation*, 9(3), July 2006. 4

[26] P. Erdös and A. Rényi. On Random Graphs I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959. 24, 30

[27] Jennifer H. Fewell. Social Insect Networks. *Science*, 301:1867–1870, September 2003. 24

[28] R. P. Freckleton and A. R. Watkinson. Large-Scale Spatial Dynamics of Plants: Metapopulations, Regional Ensembles and Patchy Populations. *Journal of Ecology*, 90:419–434, 2002. 22

[29] Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The Human Disease Network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, May 2007. 4

[30] Rocío Gómez-Bardón and Jonathan E. Rowe. An Agent-based Simulation of Crowd Movement. In D. Al-Dabass, editor, *Proceedings of Uk Simulation Confrence*, Cambridge, UK, April 2003. Uk Simulation Society. 45, 51

[31] Andrzej Grabowski and R.A. Kosiński. The SIS Model of Epidemic Spreading in a Hierarchical Social Network. *Acta Physica Polonica B*, 36:1579–1593, October 2005. 33

[32] Andrzej Grabowski Grabowski and Magdalena Rosińska. The SIS Model for Assessment of Epidemic Control in a Social Network. *Acta Physica Polonica B*, 37:1521, May 2006. 34

[33] I. Hanski and O. E. Giaggiotti, editors. *Ecology, Genetics, and Evolution of Metapopulations*. Elsevier Academic Press, 2004. 23, 91

[34] I. Hanski and O. Ovaskainen. Metapopulation Theory for Fragmented Landscapes. *Journal of Theoretical Population Biology*, 64:119–127, November 2002. 22, 23

[35] Ilkka Hanski. Spatially Realistic Theory of Metapopulation Ecology. *Publicacin Naturwissenschaften*, 88(9):372–381, September 2001. 23

[36] Ilkka Hanski and Mats Gyllenberg. Two General Metapopulation Models and the Core-Satellite Species Hypothesis. *The American Naturalist*, 142(1):17–41, July 1993. 22

[37] Ilkka Hanski and Otso Ovaskainen. The Metapopulation Capacity of a Fragmented Landscape. *Nature*, 404:755–758, April 2000. 23

[38] Rainer Hegselmann and Andreas Flache. Understanding Complex Social Dynamics: A Plea For Cellular Automata Based Modelling. *Journal of Artificial Societies and Social Simulation*, 1(3), 1998. 5, 16, 17

[39] S. Janson, T. Luczak, and A. Rucinski. John Wiley, New York, 1999. 25

[40] N. R. Jennings and M. J. Wooldridge. *Applications of Intelligent Agents*, pages 3–28. Springer-Verlag, 1998. 7

[41] Bin Jiang. SimPed: Simulating Pedestrian Flows in a Virtual Urban Environment. *Journal of Geographic Information and Decision Analysis*, 3(1):21–30, 1999. 35

[42] Dr. B.S. Kakkilaya. Complications of P. Falciparum Malaria. http://www.malariasite.com/malaria/Complications2.htm, June 2009. [Online; accessed 03-March-2010]. 42

[43] Dr. B.S. Kakkilaya. Malaria Life Cycle. http://www.malariasite.com/malaria/LifeCycle.htm, June 2009. [Online; accessed 03-March-2010]. 42

[44] Alex Kamenev and Baruch Meerson. Extinction of an Infectious Disease: A Large Fluctuation in a Nonequilibrium System. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 77(6):061107, 2008. 37

[45] Stuart A. Kauffman. *At Home in the Universe - The Search for the Laws of Self-Organization and Complexity*. Oxford University Press, 1995. 4, 24

[46] M. J. Keeling and L. Danon. Mathematical Modelling of Infectious Diseases. *British Medical Bulletin*, page 038, 2009. 29

[47] Jeffrey O. Kephart and Steve R. White. Directed-Graph Epidemiological Models of Computer Viruses. In *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, pages 343–359, Los Alamitos, CA 90720-1264, May 1991. 4, 91

[48] Judith S. Kleinfeld. Could It Be a Big World After All? The 'Six Degrees of Separation' Myth. *Psychology Today Magazine*, Mar/Apr 2002. 26

[49] Geoffrey A. Landis. The Fermi Paradox: An Approach Based on Percolation Theory. *Journal of the British Interplanetary Society*, 51:163–166, 1998. 119

[50] Reinhard C. Laubenbacher, Abdul Salam Jarrah, Henning S. Mortveit, and S. S. Ravi. A Mathematical Formalism for Agent-Based Modeling. *ArXiv e-prints*, Dafecember 2008. 7

[51] Simon B. Laughlin. Communication in Neural Networks. *Science*, 301, September 2003. 24

[52] Richard Levins. Some Demographic and Genetic Consequences of Environmental Heterogeneity for Biological Control. *Bulletin of the Entomological Society of America*, (15):237240, 1969. 22

[53] Jonathan Luk. GIS, Scale, and Landscape Ecology. http://crssa.rutgers.edu/courses/lse/Web_scale/Webproject/html/jONGISfinal.htm; [Online: accessed 21-July-2009]. 35

[54] John Paul Marney and Heather F. E. Tarbert. Why Do Simulation? Towards a Working Epistemology for Practitioners of the Dark Arts. *Journal of Artificial Societies and Social Simulation*, 3(3), 2000. 4

[55] Pim Martens and Lisbeth Hall. Malaria on the Move: Human Population Movement and Malaria Transmission. *Emerging Infectious Diseases*, 6(2), 2000. 43

[56] R. M. May and R. M. Anderson. Transmission Dynamics of HIV Infection. *Nature*, 326:137–142, March 1987. 36, 37

[57] R. M. May, R. M. Anderson, and M. E. Irwin. The Transmission Dynamics of Human Immunodeficiency Virus (HIV) [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 321(1207):565–607, 1988. 36

[58] L.E. Merrin, C.R. Todd, D.G. Williams, E.S.G. Schreiber, and C. James. Coupling Meta-Population Models with GIS to Predict Freshwater Biotic Invasions. In Andre Zerger and Robert M. Argent, editors, *MODSIM05*, 2005. 35

[59] Gabriela Minigo, Tonia Woodberry, Kim A. Piera, Ervi Salwati, Emiliana Tjitra, Enny Kenangalem, Ric N. Price, Christian R. Engwerda, Nicholas M. Anstey, and Magdalena Plebanski. Parasite-Dependent Expansion of TNF Receptor II-Positive Regulatory T Cells with Enhanced Suppressive Activity in Adults with Severe Malaria. *PLoS Pathogens*, 5(4):e1000402, April 2009. 42

[60] Atte Moilanen. Patch Occupancy Models of Metapopulation Dynamics: Efficient Parameter Estimation Using Implicit Statistical Inference. *Ecology*, 80(3):1031–1043, 1999. 23

[61] J. M. Montoya and R. V. Sole. Small World Patterns in Food Webs. *Journal of Theoretical Biology*, 214:405–412, 2002. 24

[62] Joël Mossong, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo S. Tomba, Jacco Wallinga, Janneke Heijne, Malgorzata Sadkowska-Todys, Magdalena Rosinska, and W. John Edmunds. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. *PLoS Medicine*, 5(3):e74, March 2008. 29

[63] M. E. J. Newman. Assortative Mixing in Networks. *The American Physical Society*, 89(20), November 2002. 26

[64] M. E. J. Newman. Random Graphs as Models of Networks. *Proceedings of the Natural Academy of Science*, 99(1):2566–2572, February 2002. 24, 26, 27, 32, 61

[65] M. E. J. Newman. Properties of Highly Clustered Networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 68, March 2003. 27, 30

[66] M. E. J. Newman. The Structure and Function Of Complex Networks. *SIAM Review*, 45:167–256, March 2003. 24, 25, 28, 29

[67] M. E. J. Newman. Random Graphs with Clustering. *Physical Review Letters*, 103(5):058701, July 2009. 29, 33

[68] M. E. J. Newman, I. Jensen, and R. M. Ziff. Percolation and Epidemics in a Two-Dimensional Small World. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 65(2), January 2002. 26, 33, 147

[69] M. E. J. Newman and Juyong Park. Why Social Networks Are Different From Other Types of Networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 68, May 2003. 28, 30

[70] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random Graphs with Arbitrary Degree Distribution and Their Applications. *Physical Review E*, 64, 2001. 27, 30, 31

[71] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random Graph Models of Social Networks. *Proceedings of the National Academy of Sciences*, 99(1):2566–2572, February 2002. 27, 31

[72] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic Dynamics and Endemic States in Complex Networks. *Physical Review E*, 63:066117, 2001. 111, 116, 129, 153, 162

[73] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.*, 86(14):3200–3203, Apr 2001. 111, 116, 129, 153, 162

[74] Romualdo Pastor-Satorras, Alexei Vzquez, and Alessandro Vespignani. Dynamical and Correlation Properties of the Internet. *The American Physical Society*, 87(25), December 2001. 27, 29

[75] D. J. S. Price. Networks of Scientific Papers. *Science*, 149:510–515, 1965. 24, 27

[76] Colin R. Reeves and Jonathan E. Rowe. *Genetic Algorithms: Principles and Perspectives - A Guide to GA Theory*. Kluwer Academic Publishers, 2003. 11

[77] Jonathan E. Rowe and Rocío Gómez-Bardón. El Botellón: Modelling the Movement of Crowds in a City. *Complex Systems*, 14(4):363–370, 2003. 45, 51

[78] Thomas C. Schelling. Models of Segregation. *The American Economic Review*, 59(2):488–493, 1969. 4

[79] Thomas C. Schelling. Dynamic Models of Segregation. *Journal of Mathematical Sociology*, 1:143–186, 1971. 4

[80] Rilando B. Schinazi. On the Role of Social Clusters in the Transmission of Infectious Diseases. *Journal of Theoretical Population Biology*, 61(2):163–169, 2002/3. 20

[81] Rilando B. Schinazi. On the Role of Reinfection in the Transmission of Infectious Diseases. *Journal of Theoretical Biology*, 225(1):59–63, November 2003. 20

[82] Rilando B. Schinazi. On the Spread of Drug Resistant Diseases. *Journal of Statistical Physics*, 97:409–417, October 1999. 19

[83] Rinaldo B. Schinazi. Mass Extinctions: an Alternative to the Allee Effect. *Annals of Applied Probability*, 15:984, 2005. 21

[84] Rinaldo B. Schinazi. On the Role of Spatial Aggregation in the Extinction of a Species. *Progress in Probability*, 60:551–557, 2008. 21

[85] I. B. Schwartz, L. Billings, M. Dykman, and A. Landsman. Predicting Extinction Rates in Stochastic Epidemic Models. *Journal of Statistical Mechanics: Theory and Experiment*, 1:5, January 2009. 37

[86] Small World Project. http://smallworld.columbia.edu/description.html. [Online; accessed 06-June-2009]. 26

[87] University of Chicago Social Science Research Computing. RePast. http://repast.sourceforge.net/, 2008. [Online; accessed 03-March-2010]. 11

[88] Complex Systems Society. What Are Complex Systems. http://cssociety.org/. [Online; accessed 03-March-2010]. 3

[89] Susanne H. Sokolow, Patrick Foley, Janet E. Foley, Alan Hastings, and Laurie L. Richardson. Disease Dynamics in Marine Metapopulations: Modelling Infectious Diseases on Coral Reefs. *Journal of Applied Ecology*, 46(3):621–631, 2009. 22

[90] Starlogo. The StarLogo Website. http://education.mit.edu/starlogo/. [Online; accessed 03-March-2010]. 35

[91] Keith Still. *Crowd Dynamics*. PhD thesis, University of Warwick, Mathematics Institute, 2000. 36

[92] Rob Stocker, David Cornforth, and Terry R. J. Bossomaier. Network Structures and Agreement in Social Network Simulations. *Journal of Artificial Societies and Social Simulation*, 5(4), October 2002. 27

[93] Steven H. Strogatz. Exploring Complex Networks. *Nature*, 410:268–276, March 2001. 24, 26

[94] Balázs Szendrői and Gábor Csányi. Polynomial Epidemics and Clustering in Contact Networks, 2004. 37, 118, 119, 147, 161

[95] Robert Tobias and Carole Hofmann. Evaluation of Free Java-Libraries for Social-Scientific Agent Based Simulation. *The Journal of Artificial Societies and Social Simulation*, 7(1), January 2004. 11

[96] Robert W. Ulanowicz. *A Third Window: Natural Life Beyond Newton and Darwin*. Templeton Press, May 2009. 4

[97] Robert W. Ulanowicz, Sally J. Goerner, Bernard Lietaer, and Rocío Gómez Bardón. Quantifying Sustainability: Resilience, Efficiency and the Return of Information Theory. *Ecological Complexity*, 6(1):27–36, 2009. 4

[98] Michael Vose. *The Simple Genetic Algorithm - Foundations and Theory*. The MIT Press, Cambridge, Massachusetts, 1999. 11

[99] D. J. Watts and S. H. Strogatz. Collective Dynamics of 'Small-World' Networks. *Nature*, 393:440–442, June 1998. 25, 28, 147

[100] Eric W. Weisstein. Binomial Distribution. http://mathworld.wolfram.com/BinomialDistribution.html. From MathWorld–A Wolfram Web Resource. [Online; accessed 3-March-2010]. 120

[101] Eric W. Weisstein. Geometric Distribution. http://mathworld.wolfram.com/GeometricDistribution.html. From MathWorld–A Wolfram Web Resource. [Online; accessed 3-March-2010]. 119

[102] John C. Wierman and David J. Marchette. Modeling Computer Virus Prevalence with a Susceptible-Infected-Susceptible Model with Reintroduction. *Computational Statistics and Data Analysis*, 45(1):3–23, 2004. Computer Security and Statistics. 4, 91

[103] Wikipedia. Average Path Length — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Average_path_length&oldid=192788785, 2008. [Online; accessed 30-October-2009]. 25

[104] Wikipedia. Computer Virus — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Computer_virus&oldid=322777635, 2009. [Online; accessed 30-October-2009]. 8, 91

[105] Wikipedia. Landscape Ecology — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Landscape_ecology&oldid=303258359, 2009. [Online; accessed 21-July-2009]. 35

[106] Wikipedia. Metapopulation — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Metapopulation&oldid=277791851, 2009. [Online; accessed 17-March-2009]. 22

[107] Herbert S. Wilf. *Generatingfunctionology*. Academic Press, London, 2nd edition, 1994. 30

[108] Eric Winsberg. Simulations, Models, and Theories: Complex Physical Systems and Their Representations. *Proceedings of the Philosophy of Science Association*, 2001(3), 2001. 4

[109] H. Peyton Young. Social Dynamics: Theory and Applications. *Handbook of Computational Economics II (Forthcoming)*, February 2004. 24

[110] Nikos Zaboutis and Nicolas Marmaras. Searching Efficient Plans for Emergency Rescue Through Simulation: The Case of a Metro Fire. *Cognition, Technology and Work*, 6(2):117–126, May 2004. 36