COPPE
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia
UFRJ

A PROBABILISTIC APPROACH TOWARDS IDENTIFYING THE SOURCE OF RANDOM EPIDEMICS ON FINITE NETWORKS

Danielle Alves Castelo Branco da Silva

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientadores: Daniel Ratton Figueiredo
Giulio Iacobelli

Rio de Janeiro
Março de 2020

A PROBABILISTIC APPROACH TOWARDS IDENTIFYING THE SOURCE OF RANDOM EPIDEMICS ON FINITE NETWORKS

Danielle Alves Castelo Branco da Silva

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Daniel Ratton Figueiredo, Ph.D.

Prof. Giulio Iacobelli, Ph.D.

Prof. Valmir Carneiro Barbosa, Ph.D.

Prof. João Batista de Moraes Pereira, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2020

Silva, Danielle Alves Castelo Branco da

A Probabilistic Approach Towards Identifying the Source of Random Epidemics on Finite Networks/Danielle Alves Castelo Branco da Silva. – Rio de Janeiro: UFRJ/COPPE, 2020.

XII, 58 p.: il.; $29,7$cm.

Orientadores: Daniel Ratton Figueiredo
Giulio Iacobelli

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2020.

Referências Bibliográficas: p. 57 – 58.

1. Source identification. 2. Random trees. 3. Epidemic models. I. Ratton Figueiredo, Daniel *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

A Deus, pela dádiva da vida e
por me permitir realizar tantos
sonhos nesta existência.

Agradecimentos

A minha família, por todo auxílio e suporte fornecido, pelo aconchego de um abraço e incentivo dados em momentos difíceis. Por serem meus exemplos de perseverança, cuidado, amor e por sempre terem acreditado em mim.

Aos meus orientadores, que tanto me ensinaram ao longo da trajetória do mestrado. E ao corpo docente do PESC que forneceu ao longo de todas as disciplínas realizadas o conhecimento necessário para construção desta tese de mestrado.

A UFRJ que me permitiu o sonho de estudar e concluir cursos numa das melhores Universidades do país. Ao CNPQ e CAPES, pelo auxílio financeiro através de bolsa de estudos permitindo este mestrado se tornar realidade.

Aos amigos, pelos momentos de descontração e grupos de estudos por cada uma das bibliotecas do CT.

Por fim, a todos aqueles que contribuíram, direta ou indiretamente, para a realização desta tese de mestrado, o meu sincero agradecimento.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA ABORDAGEM PROBABILÍSTICA PARA IDENTIFICAR A ORIGEM DE EPIDEMIAS ALEATÓRIAS EM REDES FINITAS

Danielle Alves Castelo Branco da Silva

Março/2020

Orientadores: Daniel Ratton Figueiredo
Giulio Iacobelli

Programa: Engenharia de Sistemas e Computação

Os modelos de redes epidêmicas são estudados há mais de 20 anos e são usados para representar diferentes processos de difusão em redes. Por exemplo, a propagação de uma doença através do contato físico entre indivíduos numa população ou a propagação de notícias falsas por meio de uma rede social online. Nesse contexto, surge o problema de identificar o vértice da rede que iniciou a epidemia a partir da observação parcial do processo epidêmico. Esta tese considera um modelo probabilístico para uma epidemia de rede finita e pressupõe a observação da árvore epidêmica ao final do processo. A partir de uma análise probabilística, apresentamos um algoritmo eficiente para encontrar o vértice com maior probabilidade de ser a origem da epidemia. Os resultados numéricos ilustram o potencial da abordagem proposta. Usaremos o modelo SI de propagação de epidemias em redes e métodos probabilísticos baseados em análise combinatória para identificar a origem da propagação de um boato numa rede quando recebermos informações parciais sobre o processo de propagação e sobre a própria rede.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

A PROBABILISTIC APPROACH TOWARDS IDENTIFYING THE SOURCE OF RANDOM EPIDEMICS ON FINITE NETWORKS

Danielle Alves Castelo Branco da Silva

March/2020

Advisors: Daniel Ratton Figueiredo
Giulio Iacobelli

Department: Systems Engineering and Computer Science

Epidemic network models have been studied for over 20 years and are used to represent different diffusion processes across a network. For example, the spread of a disease through physical contact between individuals in a population or the spread of fake news through an online social network. In this context, the problem of identifying the network vertex that initiated the epidemic from the partial observation of the epidemic process arises. This thesis considers a probabilistic model for a finite network epidemic and assumes observation of the epidemic tree at the end of the process. From a probabilistic analysis, we present an efficient algorithm to find the vertex most likely to be the epidemic source. Numerical results illustrate the potential of the approach being proposed. We will use the SI network epidemic spreading model and probabilistic methods based on combinatorial analysis to identify the source of the spread of a rumor in a network when we receive partial information about this propagation process and about the network itself.

Contents

List of Figures **x**

List of Tables **xii**

1 Introduction **1**

1.1 Motivation 2

1.2 Objective and Contribution 3

1.3 Organization 4

2 Epidemic Models **5**

2.1 Classic Epidemic Models 5

2.1.1 SI Model 6

2.1.2 SIS Model 8

2.1.3 SIR Model 9

2.2 Network Epidemic Models 11

2.2.1 SI model on Networks 13

2.2.2 SIS model on a Network 14

2.3 Epidemics Through Recursive Random Trees 15

2.3.1 Classical Recursive Random Trees 17

2.3.2 Recursive Random Spanning Trees 18

2.3.3 Recursive Random Weighted Spanning Trees 19

3 Source Identification in Recursive Random Trees **21**

3.1 Computing the Probability of the Epidemic Source 22

4 Source Identification in Recursive Random Trees: a Special Case with Complete Graphs **26**

4.1 Probability of an Edge Sequence 26

4.1.1 Numerical Example 27

4.2 Probability of a Tree by Fixing a Root 28

4.3 Counting the Number of Sequences in $B_{(\tau,v_0)}$ 29

4.3.1 Line and Star Example 32

4.4 The Most Likely Source 33
4.5 Source Identification Algorithm 37

5 Numerical evaluation 40

5.1 Numerical evaluation of relative frequency of identified sources 41
5.2 Distance distribution on source identification 45
5.3 Runtime 48

6 Related Work 50

7 Conclusion 53

7.1 Future Work 54

Bibliography 57

List of Figures

2.1 States of SI model and classic logistic growth curve (BARABÁSI, 2017). 8

2.2 States of SIS model and classic logistic growth curve (BARABÁSI, 2017). 9

2.3 States of SIR model and classic logistic growth curve (BARABÁSI, 2017). 10

2.4 Exponential regime of those three epidemic models (BARABÁSI, 2017). 11

2.5 Degree Block Approximation (BARABÁSI, 2017). 12

3.1 (a) Arbitrary Graph, (b) Resulting spanning tree, (c) Resulting spanning tree without labels. 21

3.2 Evolution of the edge cut of an epidemic 24

4.1 Complete graph with eight vertices. 27

4.2 Line tree - τ^{Line} 32

4.3 Star tree - τ^{Star} 33

5.1 Relative frequency with which a node is identified as the epidemic source. 41

5.2 Relative frequency with which a node is identified as the epidemic source by the algorithm (average over 100000 rounds). 42

5.3 Relative frequency with which a node is identified as the epidemic source by the algorithm (average over 100000 rounds). 43

5.4 Frequency with which node zero (the actual source) is identified as epidemic source for different network sizes. 44

5.5 Relative frequency of the distance of the identified epidemic source to the real epidemic source for a tree with size 10. 46

5.6 Relative frequency of the distance of the node identified as the source and the real epidemic source. 47

5.7 Barplot of the average runtime. 49

5.8 Barplot of the average runtime. 49

7.1 Hypercube Q_3 55

7.2 2D lattice with 9 nodes. 56

List of Tables

1.1 Spread of diseases versus informations - DALEY and KENDALL (1964) 3

2.1 Main formulas of epidemic models in networks. 15

3.1 Sequence of infection of the nodes. 22

3.2 Sequence of infection of the nodes when starting in node x. 24

5.1 Frequency with which the nodes are returned as source in networks with different sizes. 44

5.2 Identification and distances of nodes in a network of size 10. 46

5.3 Mean and SD of the distance distribution. 48

5.4 Mean and standard deviation of runtime of source identification algorithm in the simulation. 49

Chapter 1

Introduction

Various biological, social and communication systems can be described through complex networks where nodes represent individuals, or groups of individuals, and edges represent their respective connections or interactions. In biological systems, edges can represent some kind of biological interaction such as the bonding between pairs of proteins. In social networks, edges represents social relationships such as friendships or collaborations. The structure of these networks is often complex and highly asymmetric, as nodes in these networks often have very different degrees.

Among several processes that occur in such networks, a common and important process are diffusion or epidemics. These processes model the spread of something that depends on the kind of network. An example in online social networks such as Facebook or WhatsApp is the propagation of fake news. Another example is the spread of a disease in a population through the contact network, such as the spread of HIV. In this context, it is interesting to understand how epidemics unfold on networks independently of what is spreading, an approach know as network epidemics.

The study of network epidemics is not a recent subject in academy, epidemic models have been studied for centuries and a pioneer paper was written by BERNOULLI (1766). More recently, spreading processes based on a particular epidemic model was proposed by HARRIS and WILSON (1978), epidemic models in directed-graphs by Kephart and White (1991) and epidemics on scale free networks by Pastor-Satorras and Vespignani (2001). Recent works in network epidemics focus on the behavior of the diffusion process to characterize its duration or intensity. In particular, many works focus on the input that the network structure has on the epidemic process as well as consider different epidemic models.

At a very high level, the spreading process of a rumor can be regarded as similar to the spreading process of a disease. An individual initiates a rumor, spreading it to a friend or a group of friends. Each of them can repeat this spreading process of the rumor to more people giving rise to an epidemic in a network of friends. This process is often random and depends on the network of friends and on the rumor. The study of such epidemic models, reveals characteristics of the rumor, the time required for the rumor to reach a given fraction of the population, or the identification of the node that started the rumor. This thesis, focus in this last problem when partial information about the epidemic process is available.

1.1 Motivation

In all of humanity's history several epidemics have happened and continue to happen, some of which with drastic consequences to society such as *The Great Influenza Epidemic of 1918* (BARRY, 2004). In order to understand the spread of these diseases in the population, various mathematical models have been developed to predict the behavior of an epidemic as a function of its parameters and a common approach is to use graphs.

When considering the population of a city, for example, each individual can be represented by a node in a graph, and each edge on this graph represents that two individuals have been in direct contact for a certain period of time. Moreover, this population can be divided into groups, more specifically, in three groups. Those who are sick, called infected group; those who are healthy, called susceptible and those who were sick but have been treated and cured, and cannot be contaminated by the disease again, called recovered. These groups allow the epidemic model to capture the infection process in the population over time.

Another epidemic process on a network is the diffusion of computer viruses. Modern computer viruses are programs capable of transmitting a copy of itself from one computer to another. Its process of contamination often resembles that of dissemination of pathogens in a population. In this process the nodes represents the computers and the edges represents the exchange of information between two computers. A specific kind of computer virus infect smartphones and propagate through Bluetooth connections. Thus, Bluetooth viruses can infect smartphones found within Bluetooth range of the infected phones, which is about 10-30 meters. The spread of Bluetooth viruses requires physical proximity, and is therefore similar to the spread of diseases that also require physical proximity.

While digital viruses can spread in networks of digital devices, rumors can spread in online social networks such as Facebook. Intuitively, the spread of the rumor depends on the structure of the network which is defined by a given relationship. For example, a network of students in a school will not have the same structure as the network induced by some class of this school or a group of friends in this class. The analogy between the spread of rumors and disease has been recognized for over 50 years. DALEY and KENDALL (1964) propose and discuss the connection between the two kinds of epidemics, an analogy between the spreading of an infectious disease and dissemination of information as shown in Table 1.1. This connection between these two kinds of epidemics allow us to use epidemic models to represent rumors in a more general abstraction.

Table 1.1: Spread of diseases versus informations - DALEY and KENDALL (1964)

Class	**Rumor Interpretation**	**Epidemic Interpretation**
$\boldsymbol{x}$	Has not heard the rumor	Susceptible to disease
$\boldsymbol{y}$	Actively Spreading Rumor	Infectious Case
$\boldsymbol{z}$	No Longer Spreading Rumor	Dead, Isolated or Immune

An important problem when considering epidemics is the identification of the first node to be infected, often called patient zero. This node is the source or origin of the epidemic and identifying it is not trivial if only partial information about the unfolding of the epidemic process is available. However, in many scenarios only partial information concerning the unfolding of the epidemic is available, such as the spread of a rumor through a real social network. This motivates algorithms and models that can effectively be used to identify the source of an epidemic and is the main topic addressed in this thesis.

1.2 Objective and Contribution

This thesis addresses the problem of identifying the source of an epidemic that unfolds on a finite network. The recursive random spanning tree model is used to represent the epidemic process. This model generates a tree that represents an epidemic spread across the entire population.

Thus, the information observed is a tree through which the epidemic spread. Given this tree and the underlying network, the goal is to determine the source of the epidemic. No temporal information is available such as the time nodes were infected. Thus, the structure of the tree and the underlying graph must be used to identifying the source. Intuitively, since the epidemic unfolds uniformly, its source is more likely to be at the "center" of the tree.

This work assumes a single node is infected at time zero and all other nodes will be infected and be part of the spanning tree. The main contributions of this thesis is the analysis of the probabilistic epidemic model and the characterization of the probability that a node is the source given the observed tree. This expression is solved analytically for when the underlying network is the complete graph. In this case, an efficient linear time algorithm that finds the most probable source is presented. Finally, numerical simulation assess the accuracy of the proposed methodology, and indicate that the most probable source is often the epidemic source. To calculate the probability of each node to be the epidemic source (Chapter 3) of our algorithm which returns the node with maximum probability (Chapter 4).

1.3 Organization

This thesis is organized in 6 Chapters. Chapter 2 presents a literature review on classic epidemic models, network epidemic models and recursive random trees. Chapter 3 presents the source identification in recursive random trees. In Chapter 4 investigates recursive random tree model on complete graphs and describes the algorithm to find the most probable source, as well as, numerical evaluation of the methodology. In Chapter 5 presents works related to epidemic source identification. Chapter 6 concludes the paper and discuss future works.

Chapter 2

Epidemic Models

The study of models that capture the nature of an epidemic is quite an old topic. One of the first studies about rumor spreading was in 1950's and give rise to the small world theory. Bernoulli (1766) made a study about network epidemic models, then, Kephart and White (1991) studied the spreading of computer viruses and Pastor-Satorras and Vespignani (2001) studied epidemic on scale free networks, and since then many other authors made several contributions to the area.

When a person catches a disease, depending on the pathological agent contracted, this person may recover, die or remain in a chronic disease state depending on the nature of the pathogen. In addition, this disease can spread, depending on the form of transmission and become an epidemic in the population. In order to understand how diseases spread in populations beside considering the connections that characterize the connections of this population, we must take this biological factor into account. Over the years, several researchers have studied models that could capture information so that we can mathematically understand the spreading process of a disease.

In this Chapter, we will introduce some important concepts considered for the understanding of the content presented in this thesis as classic epidemic models, the three principal epidemic models (SI, SIS and SIR); network epidemic models and recursive random trees, that will be used in this thesis.

2.1 Classic Epidemic Models

The epidemic models classify individuals according to their epidemic state that can be: *susceptible*, *infected* and *recovered*. The susceptible individuals are those who have not had contact with the pathogen/infected individuals, the infected individuals are those who have already been contaminated by the pathogen and the recovered individuals are those who have been contaminated but were healed. Recovered individuals may be able to contract the disease again never contract the disease again and each one of these types are represented in the models that will be present in this Section.

Consider a population with n individuals with an amount I of infected individuals and an amount S of healthy individuals considered susceptible to the pathogen. The infected individuals may contaminate any susceptible individuals as long as exist an environment for contamination as direct contact or other form that the pathogen needs for contamination.

After being contaminated, an individual may or may not recover according to the pathogen in question. When the pathogen in question does not allow cure, diseases such as HIV for example, the model that captures the behavior of this epidemic is the SI model. When the pathogen allows the individual to recover, such as influenza, measles and others, two models can be used, the SIS model or the SIR model. For a disease like measles, once cured, the individual does not contract this disease again, we can use the SIR model. When it comes to a disease that the individual may have again like the flu, we use the SIS model.

2.1.1 SI Model

The simplest mathematical representation of an epidemic, the SI model, has only two states: *susceptible* and *infected* - see Figure 2.1. In the SI model, the individuals, once infected, remained in this state, that is, they can not recover from the disease. So at the beginning of the epidemic, all individuals belong to susceptible state, which is, healthy individuals who have not had contact with the disease-causing pathogen. Once infected, an individual, changes its state from susceptible to infected and can contaminate other individuals in the susceptible state and remain on this state until the end of the epidemic. This way, the fraction of infected individuals starts in zero and grows until it stabilizes and reaches one when all individuals are infected - see Figure 2.1.

The traditional approach considers that all individuals in the network have an equal chance to have direct contact and be contaminated if in contact with infected individuals. Once infected, this individual can contaminate any other susceptible individual. Obviously, this is not a realistic representation of the world, as we know that not all individuals are equally related to the other individuals in the network (NEWMAN, 2012).

Consider a disease spreading through a population with n individuals and the homogeneity in the process of contamination. At each unit of time, only one individual of the population is infected. Let $S(t)$ be the number of individuals who are healthy (susceptible state) and $I(t)$ the number of infected individuals of the population at time t, the probability of contamination of a new vertex at time t is

$\frac{1}{S(t)}$, that is choose one of the susceptible individuals to infect.

At the beginning of the epidemic, all individuals in the population are susceptible and there are no infected individuals (the infected set is empty). At time $t = 0$ a first individual is infected then $I(0) = 1$ and $S(0) = n - 1$. Suppose that the probability of the disease being transmitted from an infected to a susceptible individual at a unit of time is equal to β.

Since there are *I(t)* infected individuals transmitting the pathogen at time t, each at rate β, we can write a differential equation for the rate of new infections *I(t)* and to facilitate the notation we will use the fraction of infected and susceptible individuals respectively as $i = {}^{I(t)}/_{n}$ and $s = {}^{S(t)}/_{n}$, thus (NEWMAN, 2012) then, the rate of new infections at time t is:

$$\frac{di}{dt} = \beta si \tag{2.1}$$

At the same time as the number of infected individuals increases, the number of susceptible individuals decreases at the same rate:

$$\frac{ds}{dt} = -\beta si \tag{2.2}$$

These equations characterize mathematically the SI model. But, as we know that there is only two possible states in this model, each individual can only belong to one of the two states. Therefore, the sum of the number of infected and susceptible individuals must be equal to the size of the population. Mathematically, $S(t)+I(t) = n$, this way we can rewrite Equation 2.1 using this information, we have that $s = 1-i$, so:

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

Using standard methods we can solve this equation to give the *logistic growth equation*, were i_0 is the value of ${}^{I(t)}/_{n}$ at $t = 0$.

$$i(t) = \frac{i_0\, e^{\beta t}}{1 - i_0 + i_0 e^{\beta t}} \tag{2.3}$$

Next, we can see in the Figure 2.1, the graph of the logistic growth curve and the states that the model presents.

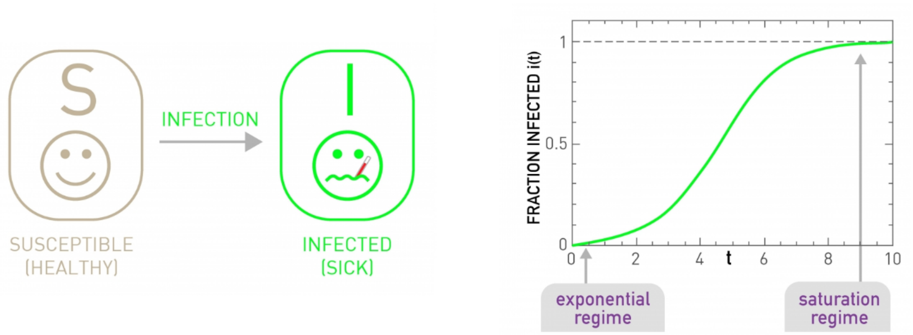


Figure 2.1: States of SI model and classic logistic growth curve (BARABÁSI, 2017).

The logistic growth curve increases exponentially for a short time as we can see in Figure 2.1, corresponding to the initial phase of the spreading process when the most part of the population is susceptible and the final phase when the disease spreads in the population and most of the population is already infected (NEWMAN, 2012). As individuals cannot recover, the fraction of infected individuals will always reach one.

2.1.2 SIS Model

There are many ways to extend the SI model in order to obtain a more realistic model or a model which is more appropriate for a specific disease. The SIS model, is a extension of the SI model that allows *reinfection*, that is, allows the individual to recover and become susceptible again, being able to be reinfected by the pathogen.

As the SI model, the SIS model presents only two states: *Susceptible* and *Infected*, but the main difference between these two models is that in SIS model the infected individuals can return to the susceptible state after recovery - see Figure 2.2. The infection between individuals are assumed to happen at average rate β per individual and infected individuals return to susceptible state at some constant average rate μ, that is, the recovery rate. As in SI model, the number of infected individuals at time t is $I(t)$ and the number of susceptible individuals at time t is $S(t)$. The differential equations for this model will be present below and to facilitate the notation we will use $i = {}^{I(t)}/_{n}$ and $s = {}^{S(t)}/_{n}$ (NEWMAN, 2012):

$$\frac{ds}{dt} = \mu i - \beta s i; \tag{2.4a}$$

$$\frac{di}{dt} = \beta s i - \mu i. \tag{2.4b}$$

Note that we have the same equation of the SI model if the recovery rate is zero. As we know, because there are only two states in the model, $s + i = 1$ we can transform Equation 2.4b of infected individuals in:

$$\frac{di}{dt} = (\beta - \mu - \beta i)i$$

Solving the previous equation using standard methods, we have the *logistic growth equation* present below.

$$i(t) = i_0 \, \frac{(\beta - \mu)\, e^{(\beta-\mu)t}}{\beta - \mu + \beta\, i_0\, e^{(\beta-\mu)t}} \tag{2.5}$$

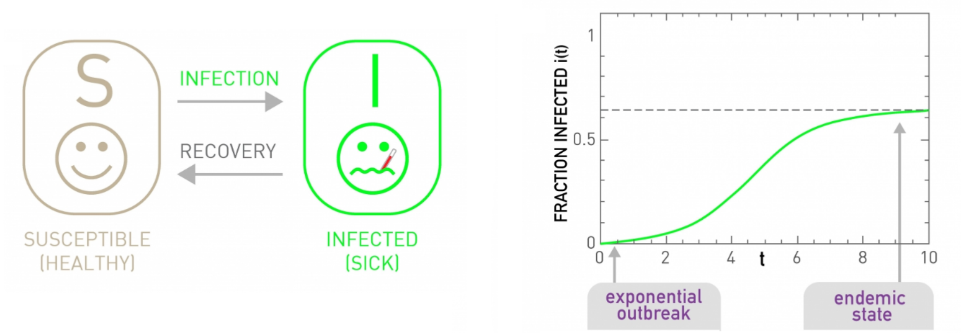


Figure 2.2: States of SIS model and classic logistic growth curve (BARABÁSI, 2017).

Note that if $\beta < \mu$ then Equation 2.5 predicts that the disease will die out exponentially and because of the fact that the individuals can recover from the disease, the logistic growth curve will never achieve one to the fraction of infected individuals if β is greater than μ (NEWMAN, 2012). And if $\beta > \mu$ the logistic growth curve will be similar of the SI model - see Figure 2.2 - but differs in a important aspect: the population will never be fully infected by the disease.

2.1.3 SIR Model

In some diseases, when a individual recovers, he will never contract it again, this characterizes the SIR model, where there are three states, which are: *Susceptible*, *Infected* and *Recover* (or *Removed*) - see Figure 2.3 (BARABÁSI, 2017).

Different than the SI and SIS model, in the SIR model, once recovered, the individual is immune to the pathogen and will not return to the susceptible state anymore.

Just like in SI and SIS models, infection between individuals are assumed to happen at average rate β per individual and infected individuals recover (or die) at some constant average rate μ (NEWMAN, 2012). The number of infected individuals at time t is represented by *I(t)*, the number of susceptible by *S(t)* and the number of recovered individuals at time t is represented by *R(t)*. The differential equations which represents this model are presented below and to facilitate the notation we will use $i = {}^{I(t)}/_{n}$, $s = {}^{S(t)}/_{n}$ and $r = {}^{R(t)}/_{n}$:

$$\frac{ds}{dt} = -\beta si; \tag{2.6a}$$
$$\frac{di}{dt} = \beta si - \mu i; \tag{2.6b}$$
$$\frac{dr}{dt} = \mu i. \tag{2.6c}$$

Again, if the recovery rate is zero, we have the same equations of the SI model. As there is only those three possible states, as in the previous models, it is necessary to satisfy that the sum of the number of individuals in each state be equal to the population size and by dividing all these quantities by the population size, we have that: $s + i + r = 1$. Now, solving these equations in 2.6

$$\frac{ds}{dt} = -\beta i\left[1 - r - i\right]; \tag{2.7a}$$
$$\frac{di}{dt} = -\mu i + \beta i\left[1 - r - i\right]; \tag{2.7b}$$
$$\frac{dr}{dt} = \mu i. \tag{2.7c}$$

In Figure 2.3 we can see the curves of the states of this model. Note that the number of susceptible individuals decreases exponentially and the number of infected increases very quickly until the individuals start to recover, than the number of infected decreases and never reaches one.

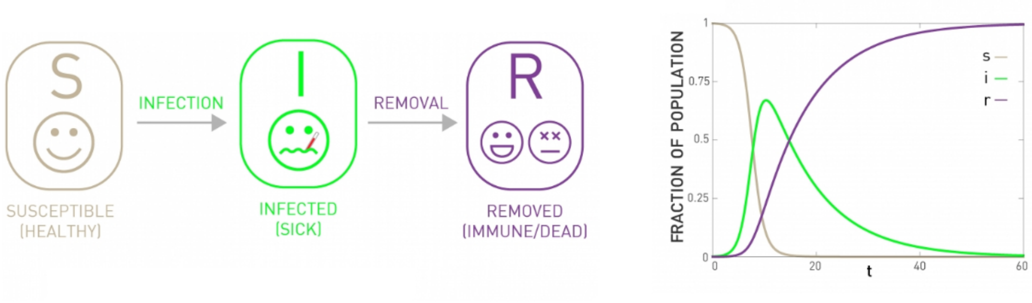


Figure 2.3: States of SIR model and classic logistic growth curve (BARABÁSI, 2017).

The difference between these three models is clear when we observe the fraction of infected individuals in each of them - see Figure 2.4. The outcomes are different for large times but at the beginning of the epidemic, in the exponential regime, all increases exponentially. In the SI model everyone becomes infected; in the SIS model either reaches an endemic state, in which a finite fraction of individuals are always infected, or the infection dies out; and in the SIR model everyone recovers at the end. Note that in SI and SIS models, when the fraction of infected reaches the maximum, stabilizes and remains until the end of the epidemic. Already in SIR, when it reaches the maximum, decreases until returning to zero.

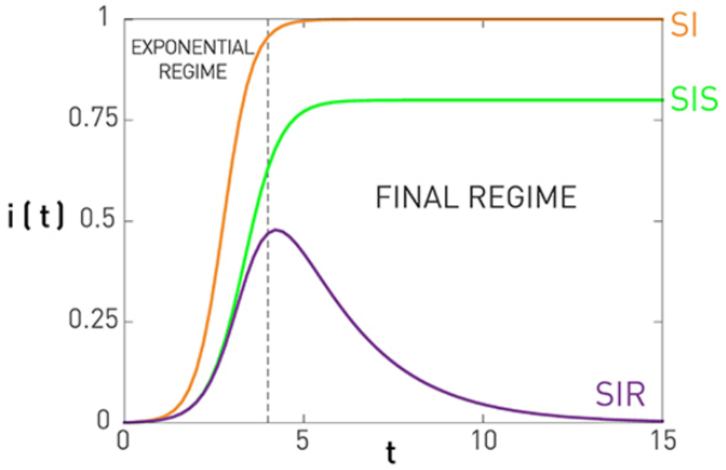


Figure 2.4: Exponential regime of those three epidemic models (BARABÁSI, 2017).

In Figure 2.4, we can see the rate of infected individuals throughout the epidemic considering finite population. Observing these curves we can have different questions:

1. If we use finite graphs, this epidemic will last or die?

2. If the epidemic will last, how soon will the entire population be infected?

3. If the epidemic will die, how soon will that happen?

Note that if an epidemic last or die, depends on the rate of infection and recovery of the population and if we where in infinite graphs the epidemic can last forever because the population does not stop growing making them always susceptible individuals.

2.2 Network Epidemic Models

In Section 2.1, we present the three most used epidemic models of literature. But these models have the assumption of “full mixing” of the population, this means that each individual may have contact with any other individual who may be contaminated in the network and may then be infected. In addition to the assumption that an individual can infect anyone else in the network, these models also carry the assumption that all individuals have a comparable number of contacts $\langle k \rangle$. In real word, individuals can transmit a pathogen only to those they come into contact with and we cannot say that two individuals, or two groups of individuals, could potentially have contact with everyone else. This way the pathogens spread on a complex contact network.

It is important to note that in the network epidemic models the structure of the network is taken into account. This type of models began to be used approximately 20 years ago and one of the main papers of the subject was published by PASTOR-SATORRAS e VESPIGNANI (2001). They define a dynamical model for the spreading of infections on scale-free networks finding the absence of an epidemic threshold and its associated critical behavior. This paper is one of the most important and cited in network epidemic models.

In this Section, we will show some differences with the classic epidemic models presented in Section 2.1 for epidemic processes on networks and will focus in the model with only two states, SI and SIS. The approach used in this thesis will consider the SI model.

Network epidemic models take into account the network structure in equations. As a network model, we need to understand what means first the motivation to use structure information in the model equations.

As the network models aim to be as close as possible to the real word, we cannot consider that all individuals are connected, because the probability of two individuals randomly chosen in a very large sample meeting each other is very small to the point of being neglected. So we have to consider that one individual can infect only the individuals with whom it is connected in some way. Another thing to consider is that individuals with the same number of connections tend to behave in the same way in the spreading process and can be grouped with the individuals with the same number of connections.

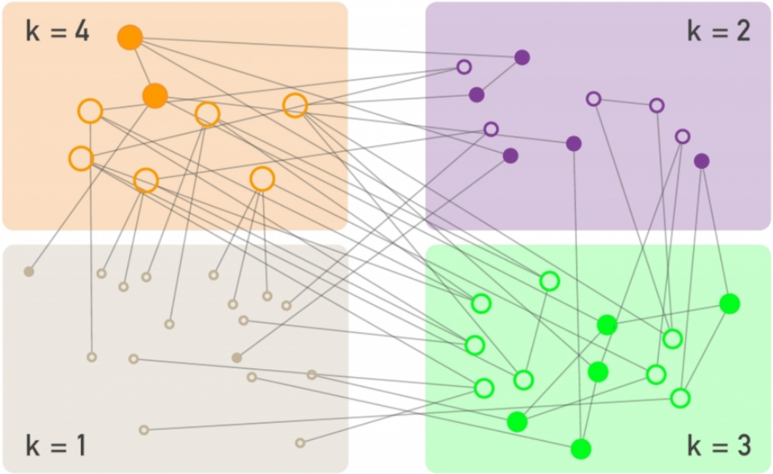


Figure 2.5: Degree Block Approximation (BARABÁSI, 2017).

In network epidemic models, the degree of each node is regarded as an implicit variable and this is achieved by the *degree block approximation* that distinguishes nodes by their degree and makes the assumption that nodes with the same degree are statistically equivalent and behave similarly - see in the Figure 2.5 (BARABÁSI, 2017).

2.2.1 SI model on Networks

The calculation of the fraction of infected nodes in network epidemic models is very similar to the classic epidemic models, differing only in taking the degree of the node into account. By the *degree block approximation* the calculation of the number of infected individuals is partitioned because nodes with the same degree tend to behave similar so the calculation of infected nodes considers the number of infected nodes with degree $\langle k \rangle$

$$i_k = \frac{I_k}{n_k} \tag{2.8}$$

where n_k is the number of degree-k nodes in the network.

Given all the different nodes degrees, the total fraction of infected nodes can be calculated by the sum of all infected degree-k nodes multiplied by the fraction of nodes with degree equal to k, denoted ny p_k, as we can see in Equation 2.9.

$$i = \sum_k p_k \, i_k \tag{2.9}$$

Calculated the fraction of infected nodes in the network, we can write the differential equations that represent the SI model on networks for each degree separately. Then, since we have $i_k(t)$ infected individuals with degree equal to k are transmitting the pathogen, each at rate β at time t, a density function that represents the fraction of infected neighbors of a susceptible node k Θ_k, the average number of new infections $di_k(t)$ during a timeframe dt is calculated as shown in Equation 2.10.

$$\frac{di_k}{dt} = \beta(1 - i_k)k\Theta_k \tag{2.10}$$

The Equation 2.10 is similar to Equation 2.1 in the classic epidemic model where the infection rate is proportional to β and the fraction of degree-k nodes that are not yet infected $(1 - i_k)$ and the node degree k. After some calculations, the fraction of infected nodes with degree k presented in 2.10, becomes:

$$\frac{di_k}{dt} \approx \beta k i_0 \, \frac{\langle k \rangle - 1}{\langle k \rangle} \, e^{\left(\frac{t}{\mathscr{T}}\right)}. \tag{2.11}$$

This is because we can approximate $\beta(1 - i_k)k\Theta_k$ by the factor $\beta k \Theta_k$ and transform the Θ_k function in $i_0 \, \frac{\langle k \rangle - 1}{\langle k \rangle} \, e^{\left(\frac{t}{\mathscr{T}}\right)}$ which are demonstrated in BARABÁSI (2017). Beside that, we can calculate the characteristic time $\mathscr{T}$ for the spread of the pathogen in the network which takes into account the average degree in this network and the spreading rate β.

$$\mathscr{T} = \frac{\langle k \rangle}{\beta(\langle k \rangle^2 - \langle k \rangle)} \tag{2.12}$$

Integrating Equation 2.11 under all k, we get the fraction of infected nodes and in the same way as in 2.9 the total fraction of infected nodes grows with time as follows:

$$\begin{aligned} i &= \int_0^{k_{max}} i_k p_k \, dk \\ &= i_0 \left(1 + \frac{\langle k \rangle^2 - \langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \left(e^{\left(\frac{t}{\mathscr{T}}\right)} - 1 \right) \right) \end{aligned} \tag{2.13}$$

According to the spreading time equation (2.12) the characteristic time $\mathscr{T}$ depends not only on $\langle k \rangle$, but also on the network's degree distribution through $\langle k^2 \rangle$. In the SI model with time the pathogen reaches all individuals. Consequently the degree heterogeneity affects only the characteristic time, which in turn determines the speed with which the pathogen sweeps through the population (BARABÁSI, 2017).

2.2.2 SIS model on a Network

As in the SI model, equations of the SIS model are a direct extension of the classic SIS model of epidemics. So, as the previous Section, the differential equation to describe the infected individuals in the model is represented below and the difference between the classic and the network model is the presence of the recovery term $-\mu i_k$. This difference impacts on the characteristic time equation (2.15). Note that the characteristic time in SI model is different of the characteristic time on SIS model.

$$\frac{di_k}{dt} = \beta(1 - i_k)k\Theta_k(t) - \mu i_k \tag{2.14}$$

$$\mathscr{T} = \frac{\langle k \rangle}{\beta\langle k^2 \rangle - \mu\langle k \rangle} \tag{2.15}$$

It is important to highlight that depending on the value of μ, if it is large enough, i_k decays exponentially and this condition does not depend only on the recovery rate and $\langle k \rangle$. Another thing important is that in order to predict whether a disease will persist in a population, the *spreading rate* is defined. This rate depends only on the probability of transmission β and the recovery rate μ. The higher the spreading rate, the more likely it is that the disease will spread in the population (BARABÁSI, 2017).

$$\lambda = \frac{\beta}{\mu} \tag{2.16}$$

Table 2.1 presents the continuum equation for the three basic epidemic models (SI, SIS, SIR) on a network with arbitrary $\langle k \rangle$ and $\langle k^2 \rangle$, this continuum equation gives the average number of new infections during a certain timeframe. Beside that, the table contains the corresponding characteristic time $\mathscr{T}$ and the epidemic threshold λ_c for these three models.

Table 2.1: Main formulas of epidemic models in networks.

Model	**Continuum Equation**	$\mathscr{T}$	λ_c
SI	$\frac{di_k}{dt} = \beta[1 - i_k]k\Theta_k$	$\frac{\langle k \rangle}{\beta(\langle k^2 \rangle - \langle k \rangle)}$	0
SIS	$\frac{di_k}{dt} = \beta[1 - i_k]k\Theta_k - \mu i_k$	$\frac{\langle k \rangle}{\beta\langle k^2 \rangle - \mu\langle k \rangle}$	$\frac{\langle k \rangle}{\langle k^2 \rangle}$
SIR	$\frac{di_k}{dt} = \beta s_k\Theta_k - \mu i_k$ $\mathrm{s}_k = 1 - i_l - r_k$	$\frac{\langle k \rangle}{\beta\langle k^2 \rangle - (\mu+\beta)\langle k \rangle}$	$\frac{1}{\frac{\langle k^2 \rangle}{\langle k \rangle} - 1}$

In the SI model, the epidemic threshold λ_c is equal to zero and there is no recovery, that is, $\mu = 0$, so the pathogen spreads until all susceptible individuals in the network are infected and the characteristic time considers only the average degree in the network and the spreading rate. In the SIS model, the continuum equation differs of the SI equation for one term which is discounted related to the recovery rate and the characteristic time also considers the recovery rate in Equation. For further details, we refer the reader to BARABÁSI (2017).

2.3 Epidemics Through Recursive Random Trees

Recursive random trees models (RRT) are a special class of models used to describe an epidemic process. We shall see three different types of RRT: classic recursive random trees, recursive random spanning trees and recursive random weighted spanning trees.

This models represent an epidemic spreading on the population and the representation of this population is made as follows. Let G be a general connected graph where V and E are the sets of vertices and edges of this graph and the total number of its vertices and edges are denoted by $|V|$ and $|E|$ respectively. An edge connecting two vertices belonging to the population is represented by a pair of nodes, $e = \{u, v\}$. As usual, the nodes represents the individuals that may be infected and the edges the channel through which the epidemic can be spread.

In the spreading process of the epidemic, only some edges are used. These edges represents the "path taken by the epidemic", that is, the channels used to transmit some pathogen through the population and are represented by pairs of vertices with the labels linked to the time were they where contaminated in the epidemic, $\{v_k, v_{k+t}\}$. In these edges one vertex is the contaminant and the other is contaminated. To identify the contaminant and the contaminated just look the label of them, the vertex with the smaller label is the contaminant vertex and the other vertex is the contaminated.

Considering a population of size n, the population is partitioned in two subsets: *susceptible* and *infected*, named as S and I, respectively. At each contamination takes a unit of time (time discrete) and there is no chance of two individuals be infected at the same time. As only one individual is infected at each unit of time, at time n, all individuals of the population will be infected. At time zero, that is, before the epidemic starts, all population is susceptible and the infected set is empty.

Let *S(t)* and *I(t)* be the sets of susceptible and infected individuals at time t of the epidemic. At the beginning of the epidemic we have that the susceptible set has all individuals, $|S| = n$, and the infected set is a empty set, $|I| = 0$. After that, at the first infection in the population, the number of nodes in each set change and the first vertex infected receive the label v_0.

$$I(0) =$$

$$S(0) = V$$

Throughout the epidemic, the number of vertices in the susceptible and infected sets will change respecting the fact that the sum of the number of nodes in these two sets must be equal to the population size regardless of the period of the epidemic ($|S(t)| + |I(t)| = |V|\ \forall t$).

The epidemic starts in a root vertex labeled as v_0. This initiate the set of infected vertices and decreases the susceptible set in a vertex. When a vertex is infected it can then infect other vertices that are still susceptible. However, an infected vertex may only contaminate vertices with which it has some connection, that is, its neighbors in the graph. As the epidemic continues, new vertices are infected changing these two sets so that the relationship between S and I is $S(t) = V \backslash I(t)$ and $|S(t)| + |I(t)| = |V|$, $\forall t$. The labels received by the vertices correspond to the order in which they were contaminated. So, if a vertex has the label v_k we know that this vertex is the k-th infected vertex.

When the epidemic process ends, that is, when all individuals in the population are contaminated, it can be characterized by a spanning tree τ since the population is represented by a graph and all nodes in this graph will be included in this tree and there will be no cycles since the base epidemic model is the SI model. Thus, this spanning tree can be described by edge sequences b that describes the channels used in the process of spreading the epidemic. Beside that, an edge sequence represent, in a unique way, a spanning tree.

2.3.1 Classical Recursive Random Trees

The recursive random trees model is a procedure to construct trees that can be interpreted as an epidemic process based in the SI model. The classical recursive random trees model will represent an epidemic of SI type in a population of size n that can be understood in this model as a complete graph.

As explained before, this population is partitioned in two subsets, S and I and considering that the model has discrete times $t = 0, 1, \cdots, n$, that is, ate each unit of time in the epidemic, one and only one, new vertex is infected. The labels received by the new infected are referent to the time in which it was contaminated, that means the k-th infection in the epidemic will occur in time $t = k$ and will receive the label v_k.

Let $S(t)$ and $I(t)$ be the set of susceptible and infected vertices, respectively in time t. At the beginning of the epidemic, the susceptible set has all individuals and the infected set is empty. The first infection in the epidemic occurs by selecting uniformly at random one vertex from S. This means that all vertices has equal probability of being infected and this probability is equal to ${}^{1}/_{|S(0)|}$, where $|S(0)|$ is the number of vertices in S at time $t = 0$. This first infected will receive the label v_0 and the subsets S and I will be now

$$\begin{aligned} &I(0) = \{v_0\} \\ &S(0) = V \backslash I(0) \end{aligned}$$

The next contamination will occur by selecting uniformly at random an infected vertex from I and connecting the new infected vertex to the first vertex infected. The probability of choosing one vertex uniformly at random in the sets I is: ${}^{1}/_{|I(t)|}$, that means that all vertices in the subset at time t will have the same probability of being selected.

As defined before $b = (e_1, e_2, \cdots, e_{n-1})$ is the sequence of edges used in the spreading process and each edge is represented by an unordered $e = (u, v)$ where u belong to the susceptible set and v to infected set. Considering the construction of the *infected* and *susceptible* sets considering each new edge added in b we have that,

$$\begin{aligned} &S(t+1) = S(t) \backslash \{u\} \\ &I(t+1) = I(t) \cup \{u\}. \end{aligned}$$

To keep track of who infected whom, the RRT model build a tree where the vertex set is $\{v_0, \cdots, v_{n-1}\}$ and there is an edge between v_i and v_j if v_i infect v_j. Note that the process defines a tree that is given from the set of pairs of nodes chosen at every step.

2.3.2 Recursive Random Spanning Trees

The recursive random spanning tree generalizes the classical model assuming the existence of specific channels of interactions between the population. In particular, an individual may infect another only if they can interact. To encode the presence of these channels of interactions the model assumes the existence of an underlying graph where the vertices are the individuals and the edges, channels of interaction. We shall always assume that the graph which describes the population in connected. Note that, if we use as underling graph a complete graph we have exactly the classical recursive random trees. Just as in classical recursive random trees, the epidemic process will generate a spanning tree of the underlying graph.

Definition 2.3.1. *A spanning tree of a connected graph G is a sub-graph of G, that is connected and has no cycles.*

Let G be a connected graph representing a finite population. Suppose that a disease begins to spread through this population. In this epidemic process, we consider that each vertex can contaminate only their neighbors in the graph, making the spreading process of this epidemic directly dependent on the structure of this network. Once contaminated, the vertex does not recover remaining infected until the end of the epidemic, which occurs when all the individuals are contaminated.

The spreading process occurs as follows: At the start of the epidemic process, the number of infected is zero, and the number of susceptible is equal the size of the population.

When the first vertex of G was infected by being selected uniformly at random from S, the sets of infected and susceptible will change, that is, $|I(0)| = 1$ and $|S(0)| = n - 1$. The initial infected vertex v_0 can now infect susceptible individuals to whom it is connected to.

$$I(0) = \{v_0\}$$
$$S(0) = V \backslash I(0).$$

Observe that the sets *I(0)* and *S(0)* are also partitions of G. Because a node can only belong to one of these states, the edges that separates these two sets compose an edge cut of the graph G.

Definition 2.3.2. *The edge cut C of a graph G corresponding to the partition S, I of V, is a set of edges defined as*

$$C = \{e = (u, v) \in E : u \in S,\ v \in I\} \tag{2.17}$$

From this edge cut, an edge is uniformly chosen and the corresponding susceptible node is infected. The probability of choosing an edge from the edge cut is proportional to the edge cut size in the moment of infection of the vertex, that is, for the k-th infection we have that

$$P(e_k = e) = \frac{1}{|C_k|}\mathbb{I}_{\{e \,\in\, C_k\}}$$

This indicator function implies that the probability is positive only if e belong to the edge cut, otherwise, is equal to zero because this edge cannot be used at this moment.

Note that the edge cut is different in every contamination, because the set of infected and susceptible vary, and assuming that every contamination is independent, the probability of a sequence of contamination is calculated by the product of each contamination.

It is not difficult to see that the sequence of edges induced by the spreading process describe the construction of a spanning tree of the underlying graph. At the end of the spreading process, the probability of an exact sequence of edges that form the spanning tree will be equal to the product of the probabilities of contamination in each one of the steps of the epidemic, that is, the probability of choosing each of the selected edges.

2.3.3 Recursive Random Weighted Spanning Trees

Recursive random weighted spanning trees is very similar to the recursive random spanning trees. In both models there is an underlying graph that will be used as base to generate the spanning tree. However, while in recursive random spanning trees the weight of all edges is equal to one and the decision of which node is going to be infected is uniformly at random in the edge cut, in recursive random weighted spanning trees the edges have weights.

Let $G = (V,\ E)$ be a connected weighted graph, where to each $e \in E$ is associated a positive value, $w_e > 0$ and C_t be the edge cut at time t and W the sum of the weights of the edges belonging to the edge cut.

$$W = \sum_{e\,\in\, C_t} w_e$$

The contamination in recursive random weighted spanning trees is made by the weight of the edges. The first node is selected uniformly at random as all the other models of recursive random trees. But, from the second node, the way that the infection occurs is different. An edge belonging to the edge cut will be selected with probability to the corresponding weight and the vertex of the end of this edge that belongs to the susceptible set is infected.

$$P(e_t = e) = \frac{w_e}{W}\, \mathbb{I}_{\{e \,\in\, C_t\}}. \tag{2.18}$$

Those weights will represent a proportion of times that an edge belonging to the edge cut at time t is chosen instead others. If we have in a edge cut two edges with weights 7 and 3, this means that if we select one edge from the edge cut 10 times, one of them will be selected 7 times and another 3.

After every infection the node, the edge cut will change because the set of infected nodes change and after that, the next infection will occurs. This process keep repeating until all nodes are infected.

Chapter 3

Source Identification in Recursive Random Trees

Using recursive random trees as the epidemic model. Yields a labeled spanning tree where the labels represent the order of infection. Now, suppose that the labels are removed and what is observed is only a tree structure without any temporal information that characterizes the order of infection. The epidemic source can be identified by calculating the probability of a node being the epidemic source (given the observed tree) and finding the node with maximal probability.

Example

Consider the graph shown in Figure 3.1(a) and an epidemic. The nodes of the graph represent the individuals of the population, which we assume to have labels *a,b,c,d,...*, and the edges the connections among them. At the end of the epidemic the spanning tree resulting is labeled with the moments when each node was infected, $v_0, \cdots, v_{n-1}$ - see Figure 3.1(b). Now, suppose that we cannot observe those labels (times) and the only information is the spanning tree resulting from this epidemic and the original labels *a,b,c,d,..* - see Figure 3.1(c). The main problem is to determine which node was first infected and the underlying graph.

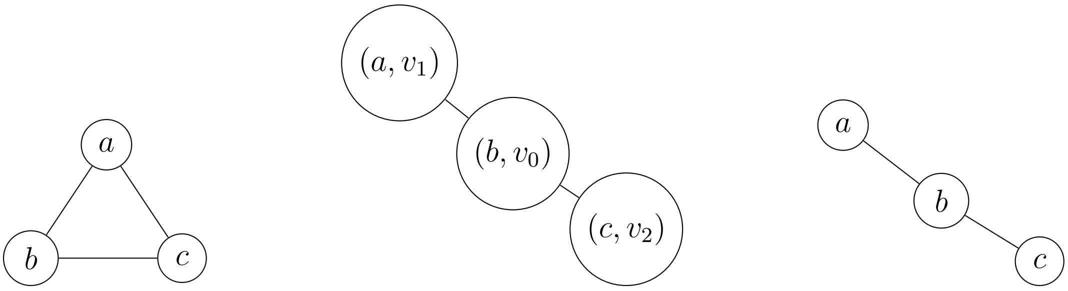


Figure 3.1: (a) Arbitrary Graph, (b) Resulting spanning tree, (c) Resulting spanning tree without labels.

Note that the tree in Figure 3.1(c) can be obtained in several ways. In Table 3.1 we show the edge sequences that can describe this tree. For example, the first line an epidemic with source in node a which contaminate node b and node b contaminate node c.

Table 3.1: Sequence of infection of the nodes.

Sequence of infection of nodes given the source
$a : (a, b), (b, c)$
$c : (c, b), (b, a)$
$b : (b, a), (b, c)$
$b : (b, c), (b, a)$

We can observe from Table 3.1 that there are twice as many sequences starting with node b than the other two nodes. This gives the intuition that node b is the most probable source of the epidemic, because the chance to build this tree is larger than the others.

The calculations to find the most probable source of the tree must result in the node b. In this example is easy to see that there are more sequences that start in node b, but when the tree is very large the most probable source must be identified without enumerating every possible sequence.

3.1 Computing the Probability of the Epidemic Source

There are many alternatives to infer an epidemic source from the epidemic tree. One of them is the method proposed in this thesis, which is simply to find the node that has the largest probability to be the epidemic source given the tree and the underlying graph.

To characterize an epidemic tree, we use rooted edge sequences that allow us to represent the course of the epidemic in the network starting from a fixed node. Each one of these rooted sequences has a probability associated and the probability that a node is the source is the sum of the probability of all rooted sequences describing the epidemic tree.

Definition 3.1.1. *Given a rooted tree (τ, v_0) were the root is v_0, an edge sequence rooted in v_0, $b = (e_1, \cdots, e_{n-1})$ generates (τ, v_0) if $e_1 \cap \{v_0\} \neq \emptyset$ and for all $k = 2, \ldots, n$, $e_k = \{u_k, v_k\} \in E(\tau)$ and*

$$|\bigcup_{i=1}^{k-1} \{u_i, v_i\} \cap \{u_k, v_k\}| = 1$$

Note that this definition makes clear that one node cannot be infected twice. Moreover, the condition for adding a new edge in the edge sequence ensures that at one edge point is infected and the other edge point is susceptible. Several sequences starting at the same vertex can describe the same tree. The set of those different sequences rooted in the same vertex v_0 that generate the tree τ is denoted $B_{(\tau,v_0)}$ and all sequences belonging to this set must satisfy the condition described in Definition 3.1.1.

Note that every node in τ can be the root v_0. In addition to having several sequences starting at the same vertex, there is at least one sequence starting at each vertex of the tree. The set of all possible edge sequences regardless of the root vertex that represents the tree τ, is called B_τ and can be described mathematically as:

$$B_\tau = \bigcup_{v \in V(\tau)} B_{(\tau,v)} \tag{3.1}$$

The probability of a node being the epidemic source is given by the sum across the sequences rooted on the node. Note that an edge sequence has a unique probability that can be calculated using the edge cut of the underlying graph at each edge. The Definition 2.3.2 makes clear that at each time step only one infection occurs in the epidemic, giving rise to a new edge whose probability is uniform in the edge cut. Thus, the probability of an edge sequence b is given by:

$$P(b) = \prod_{t=1}^{|V|-1} \frac{1}{C_t(b)} \tag{3.2}$$

where b is a rooted sequence which represents the spanning tree τ and $C_t(b)$ is the size of the edge cut in the underlying graph after the first t nodes have been infected.

Note that every infection event is independent of prior infections. Thus, the product of the probability of each infection determines the probability of the edge sequence.

The probability calculated in Equation 3.2 is the probability of a particular rooted edge sequence. To calculate the probability of the rooted tree, all possible edge sequences rooted in the same vertex $v_0 = r$ must be added. More formally,

$$P(\tau | v_0 = r) = \sum_{b \in B_{(\tau,v_0)}} P(b). \tag{3.3}$$

Example

Suppose $G = (V, E)$ is the complete graph with four vertices and six edges, and a star tree resulting from an epidemic, see Figure 3.2. We are interested in calculating the probability of a vertex x being the epidemic source. Table 3.2 shows all possible sequences starting on the vertex x, all sequences belonging to the set $B_{(\tau,x)}$.

Table 3.2: Sequence of infection of the nodes when starting in node x.

Sequence of infection of nodes
$(x,y),(x,z),(x,w)$
$(x,y),(x,w),(x,z)$
$(x,z),(x,y),(x,w)$
$(x,z),(x,w),(x,y)$
$(x,w),(x,y),(x,z)$
$(x,w),(x,z),(x,y)$

To calculate the probability of a sequence b, we need the edge cut at each new infection. Figure 3.2 presents the evolution of the edge cut and the different edges: the filled edges are those used in the spreading process, that is, the edges used for infection, and the dotted edges are those in the edge cut, that is, the edges that can be used for infecting.

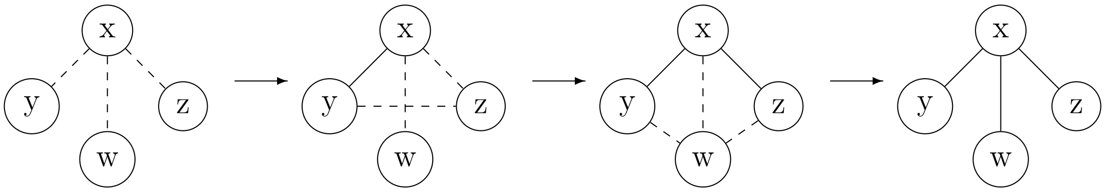


Figure 3.2: Evolution of the edge cut of an epidemic

As shown in Figure 3.2, the edge cut is equal to three at every infection. Note that other sequences will have this same probability and thus it depends on the number of sequences as follows:.

$$P(b) = \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$$

$$P(\tau | v_0 = x) = 6 \cdot \frac{1}{9} = \frac{2}{3}$$

The probability of interest is the probability of the tree τ rooted in vertex v_0. However, there is no information about which node v_0 might have started the epidemic. Thus, we assume a uniform prior for the source of the epidemic, and in particular:

$$P(v_0 = r) = \frac{1}{|V(\tau)|},\ r \in V(\tau) \tag{3.4}$$

As the probability of a spanning tree τ is equal to the sum of the probability under all possible rooted sequences across all possible root nodes, we can use the law of total probability to obtain:

$$P(\tau) = \sum_{r \in V(\tau)} P(\tau | v_0 = r) P(v_0 = r). \tag{3.5}$$

The probability of interest is the probability of a node being the root given τ. To calculate this probability we can use the Bayes rule and obtain

$$P(v_0 = r | \tau) = \frac{P(\tau | v_0 = r) P(v_0 = r)}{P(\tau)} \tag{3.6}$$

Analyzing Equations 3.4, 3.5 and 3.6 we can see that Equation 3.4 and Equation 3.5 are constant with respect to the root node. Thus, finding the most probable node is equal to maximizing the probability $P(v_0 = r|\tau)$ and, as Equation 3.5 and Equation 3.4 are constants, this is the same maximizing $P(\tau|v_0 = r)$. Thus,

$$\arg\max_{r \in V(\tau)} P(v_0 = r | \tau) = \arg\max_{r \in V(\tau)} P(\tau | v_0 = r). \tag{3.7}$$

Then, we can use the probability $P(\tau|v_0 = r)$ to identify the most probable source of the epidemic by computing

$$\arg\max_{r \in V(\tau)} \sum_{b \in B_{(\tau,r)}} P(b).$$

Note that finding the maximum above involves the rooted sequences and their probabilities. In general these are not trivial quantities as it depends on the underlying graph. However, in special cases this can be computed efficiently as in the case where the underlying graph is the complete graph, a scenario we discuss in the next chapter.

Chapter 4

Source Identification in Recursive Random Trees: a Special Case with Complete Graphs

When the underlying graph is complete, at every step of the epidemic, every susceptible vertex has the same probability to be infected than any other vertex. In particular, the size of the edge cut at a given time only depends on the number of infected nodes in epidemic and not on which nodes have been infected. This characteristic of the complete graph greatly simplifies the problem and will be detailed in this chapter.

In particular, we first show that the probability of a rooted sequence does not depend on the sequence because at any time the edge cut size depends only of the number of infected vertices at that time. Then, we show that the probability of a rooted tree depends on the number of rooted sequences that can generate the tree, and finally we show how to efficiently calculate the number of rooted sequences for a given tree. We also present a algorithm to calculate the root node that maximizes the probability of generate the tree given the root. We close this chapter by presenting a numerical evaluation of the algorithm.

4.1 Probability of an Edge Sequence

The probability that a node is the source considers the edge sequences that can generate the epidemic tree. As defined in Equation 3.2, this probability is proportional to the size of the edge cut induced by the epidemic. A interesting point is that in complete graphs, the size of the edge cut at any instant of the epidemic depends only on the number of infected nodes since all nodes are connected. In this case, the nodes infected do not matter, but only the number of infected nodes in the epidemic.

The probability of a specific edge sequence for an epidemic rooted in a fixed vertex r, can be expressed according to the formula given in Equation 4.1. This formula is the same that calculate the fraction of the number of susceptible nodes multiplied by the fraction of the number of infected nodes.

$$P(b) = \prod_{t=1}^{|V|-1} \frac{1}{C_t(b)} \, \mathbb{I}_{\{v_0=r\}} = \prod_{t=1}^{|V|-1} \frac{1}{(n-t)*t} \tag{4.1}$$

As explained before, at each unit of time, one and only one individual is infected. This way the time t is equal to the number of infected individuals. Note that knowing the population size and the number of infected individuals, we are able to calculate the size of the edge cut and we do not need any other information about the spreading process. The edge cut size is easily calculate because the full mixing assumption in the population. At each contamination in the population, we add in the cut the degree of the new infected minus the edges that can not used to contaminate susceptible individuals.

4.1.1 Numerical Example

Consider a complete graph with eight nodes as shown in Figure 4.1. To illustrate the behaviour of different rooted sequences generated from this graph. First we will calculate the probability of two different sequences representing the same tree and then we will calculate probability of a different tree.

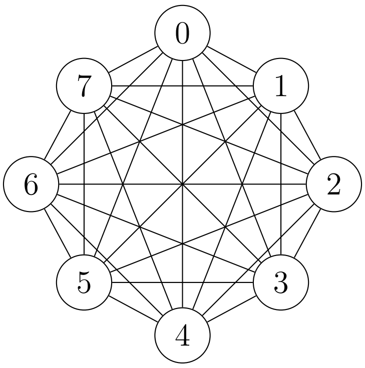


Figure 4.1: Complete graph with eight vertices.

If the epidemic tree that is a star tree, there are 5040 different rooted sequences that can generate the tree when the root is a fixed node v_0 (this number is equal to 7!). Next, we consider two different sequences b_1 and b_2 of a star tree rooted in the node $v_0 = 0$ to show that they are equally likely.

$$P(b_1) = \frac{1}{7 \cdot 12 \cdot 15 \cdot 16 \cdot 15 \cdot 12 \cdot 7} = 3.937 * 10^{-8}$$

Edge Sequences		
b_1	$=$	$\{(0,1);(0,2);(0,3);(0,4);(0,5);(0,6);(0,7)\}$
b_2	$=$	$\{(0,2);(0,4);(0,6);(0,1);(0,3);(0,5);(0,7)\}$

$$P(b_2) = \frac{1}{7 \cdot 12 \cdot 15 \cdot 16 \cdot 15 \cdot 12 \cdot 7} = 3.937 * 10{-}8$$

Consider a different sequence b_3 that generate a tree different from the star but also rooted at $v_0 = 0$. Note that the probability of b_3 is the same as b_1 and b_2.

Edge Sequence		
b_3	$=$	$\{(0,1);(0,2);(0,3);(1,4);(1,5);(2,6);(2,7)\}$

$$P(b_3) = \frac{1}{7 \cdot 12 \cdot 15 \cdot 16 \cdot 15 \cdot 12 \cdot 7} = 3.937 * 10{-}8$$

Note that the size of the edge cut first increases and then decreases. If starts with 7 and ends with 7 both cases representing the scenario with a single node infected (in the beginning) or a single node susceptible (at the end).

The edge cut size will grow until half, or half plus one, of the nodes are infected, giving the largest possible cut size. The Table below, shows the size of the edge cut considering the number of infected vertices in the epidemic process on a complete graph with 8 nodes. Clearly, the order in which the nodes are infected does not matter and the probability of infecting the vertices according to any rooted sequence is the same. This obsdervation will simplify significavely the analysis.

time of infection/ number of infected	**edge cut size (n-t)*t**
1	(8-1)·1 = 7
2	(8-2)·2 = 12
3	(8-3)·3 = 15
4	(8-4)·4 = 16
5	(8-5)·5 = 15
6	(8-6)·6 = 12
7	(8-7)·7 = 7

4.2 Probability of a Tree by Fixing a Root

Recall that more than one rooted sequence can describe the same tree, specially if we do not fix the root. Each of these sequences are independent of the others that represents the same tree. Thus, the probability of a tree must consider the probability of all possible sequences that can generate it. The probability of a tree is just the sum of the probabilities of each of these independent sequences, as described in Equation 3.3.

Recall that $B_{(\tau,v_0)}$ is the set of all possible rooted sequences describing the tree τ with root in v_0. For the complete graph, the probability of each sequence $b \in B_{(\tau,v_0)}$ to obtain the probability of the tree fixing a root in v_0 as follows:

$$\begin{aligned} P(\tau|v_0 = r) &= \sum_{b\,\in\,B_{(\tau,v_0)}} P(b) \\ &= \sum_{b\,\in\,B_{(\tau,v_0)}} \frac{1}{[(n-i)!]^2} \\ &= \frac{|B_{(\tau,r)}|}{[(n-1)!]^2}, \end{aligned} \tag{4.2}$$

Were $|B_{(\tau,v_0)}|$ is the size of the set $B_{(\tau,v_0)}$.

4.3 Counting the Number of Sequences in $B_{(\tau,v_0)}$

The probability of a spanning tree rooted in v_0, as present in Equation 4.2 of the previous Section, depends on the number of possible rooted sequences induced by the spanning tree. As the graph is complete we can use some characteristics as symmetry to count the number of different combinations that will allows to describe (τ, v_0).

These different combinations take into account the size of the sub-trees and the number of arrangements that can be made observing a vertex. To understand these calculations, we need to define some notations that will be used.

Notation	**Significance**
τ_{v_0}	Simplification of (τ, v_0);
$B(\tau_{v_0})$	Simplification of $B_{(\tau,v_0)}$;
$\tau_i^{v_0}$	Sub-tree of (τ, v_0) hung on node;
$\lvert\tau_i^{v_0}\rvert$	Size of the sub-tree of (τ, v_0) hung on node i;
$\mathscr{N}_i$	Set of neighbors of node i in the graph G.

Lemma 4.3.1. *Let $|B(\tau_k)|$ be the number of rooted sequences induced by a tree (τ, k)*

rooted in the vertex k.

$$|B(\tau_k)| \ = \ \prod_{j \in \mathscr{N}_k} |B(\tau_j^k)| \frac{(\sum_{i \in \mathscr{N}_k} |\tau_i^k|)!}{\prod_{i \in \mathscr{N}_k} |\tau_i^k|!} \tag{4.3}$$

Proof:

When we want to calculate the number of sequences of (τ, k), we need to consider the ways to construct the edge sequences. We can think of the construction of the edge sequence as the choice of edges of the sub-trees of the child nodes of k. In a simpler way, we partitioned the tree (τ, k) in d_k subsets, were d_k is the number of child nodes of k. This way we have d_k sub-trees that united compose (τ, k).

The simplest construction would be to build each one of these d_k sub-trees completely, that is, construct all d_k sub-tree until the leaves, and count the number of arrangements we can make with these sub-trees. This would be just the product of the number of sequences each of the d_k sub-trees has, which is the first part of the Equation 4.3. But we can construct (τ, k) with different combinations that combine together the d_k sub-trees.

The construction of (τ, k) can be done by alternating the edges of these d_k sub-trees. We can start one sub-tree, stop, start another sub-tree, return to a sub-tree already started and alternating until construct (τ, k). To count the number of ways to construct the edge sequence this way we need to calculate all the possible arrangements.

Calculating the number of possible arrangements is the same that having a box with *n-1* balls of d_k colors *(a, b, c,..., d_k)*, select one by one and note the color taken. At the end, when the box is empty we will have a sequence of color like *(a,a,b,c,b,b,a,c,...)*. With each ball representing one edge and each child node representing one color, in this sequence of size *n-1* with d_k colors note that we do not distinguish the edge itself but the sub-tree that this edge belongs. This occurs because there exists an order to put the edges and when the color is selected, that is, one sub-tree is chosen, we are choosing the sub-tree but will use the next possible edge in the selected sub-tree. This way the second term of the Equation 4.3 counts all the possible arrangements that we can make by selecting edges from different sub-trees to construct the edge sequence of (τ, k).

Proposition 4.3.1. *The number of rooted sequences induced by a tree $|B(\tau_k)|$ can be simplified as*

$$|B(\tau_k)| \ = \ \frac{(n-1)!}{\prod_{i \neq 0} |\tau_i^0|}. \tag{4.4}$$

Proof:

Let us assume, without less of generality, that the root 0 has k neighbors. Then, the Equation 4.4 can be expanded as follows.

$$|B(\tau_0)| = |B(\tau_1)| \cdots |B(\tau_k)| \, \frac{(|\tau_1| + \cdots + |\tau_k|)!}{|\tau_1|! \cdots |\tau_k|!}.$$

Similarly, for all the neighbors of the node zero we can apply and expand using the Equation 4.4 to calculate the arrangements of them. Considering that together the number of neighbors of them are l, we have that:

$$= |B(\tau_{k+1})| \cdots |B(\tau_{k+l})| \, \frac{(|\tau_1| + \cdots + |\tau_k|)!(|\tau_{k+1|+\cdots+|\tau_{k+l}})!}{|\tau_1|! \cdots |\tau_k|! \cdot |\tau_{k+1}|!|\tau_{k+l}|!}.$$

Repeating this step until there is no arrangement to expand and knowing that the arrangements of leaf nodes are equal to one and considering that leaf nodes will received labels from j to n,

$$= \frac{(|\tau_1| + \cdots + |\tau_k|)!(|\tau_{k+1|+\cdots+|\tau_{k+l}})! \cdots |\tau_j|! \cdots |\tau_n|!}{|\tau_1|! \cdots |\tau_k|! \cdot |\tau_{k+1}|!|\tau_{k+l}|! \cdots |\tau_j|! \cdots |\tau_n|!}$$

$$= \frac{(|\tau_1| + \cdots + |\tau_k|)!(|\tau_{k+1|+\cdots+|\tau_{k+l}})! \cdots |\tau_j|! \cdots |\tau_n|!}{\prod_{i \neq 0} |\tau_i|!}.$$

We know that the sum of the number of nodes in a sub-tree rooted in u, τ_u^k with neighbors v,x,y,z, will be the same that sum the sub-tree sizes $|\tau_v^k| + |\tau_x^k| + |\tau_y^k| + |\tau_z^k|$. This way, we can eliminate the factorial of these factors in the fraction above and use only the size of the tree rooted in u and this is valid for all sub-trees in this case.

So, using standard combinatorics, it is possible to show that the number of possible sequences which generate the same rooted spanning tree τ_0 obeys the following recursive formula:

$$\begin{aligned} |B(\tau_0)| &= \prod_{i \in \mathscr{N}_0} |B(\tau_i^0)| \frac{(|\tau_0| - 1)!}{\prod_{i \in \mathscr{N}_0} |\tau_i^0|!} \\ &= \prod_{i \in \mathscr{N}_0} \frac{|B(\tau_i^0)|}{|\tau_i^0|!} \cdot (|\tau_i^0| - 1)! \end{aligned}$$

So, recursively applying Equation 4.4 and using the fact that $|V| = n$ we obtain:

$$\begin{aligned} |B(\tau_0)| &= \frac{(|V| - 1)!}{\prod_{i \neq 0} |\tau_i^0|} \\ &= \frac{(n-1)!}{\prod_{i \neq 0} |\tau_i^0|}. \end{aligned} \tag{4.5}$$

Returning to Equation 4.2 in the previous Section and considering Lemma 4.3.1

and Proposition 4.3.1, we can make a simplification obtaining:

$$P(\tau|v_0 = r) = |B(\tau_r)|\frac{1}{((n-1)!)^2}$$

$$= \frac{(n-1)!}{\prod_{i\neq 0}|\tau_i^0|}\frac{1}{((n-1)!)^2}$$

$$= \frac{1}{(n-1)!}\frac{1}{\prod_{i\neq r}|\tau_i^r|} \tag{4.6}$$

4.3.1 Line and Star Example

We want to calculate $|B(\tau_0)|$ for two trees, the line tree and the star as illustrated in Figures 4.2 and 4.3.

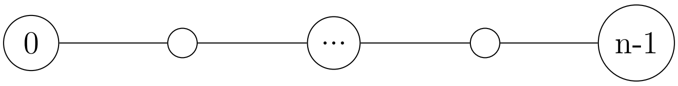


Figure 4.2: Line tree - τ^{Line}

Using Equation 4.5, considering the line tree in example and fixing the root in the node zero, the orientation to use in the calculations will be starting in the node zero to $n-1$ - see Figure 4.2. So, the size of each sub-tree will be different of the previous sub-tree by a unit ($|\tau_{n-1}^0| = 1$, $|\tau_{n-2}^0| = 2, \ldots |\tau_{n-k}^0| = k$). This way, we have that:

$$|B(\tau_0)| = \frac{(n-1)!}{\prod_{i\neq 0}|\tau_i^0|}$$

$$= \frac{(n-1)!}{\prod_{i=1}^{n-1} i}$$

$$= \frac{(n-1)!}{(n-1)!} = 1$$

This means that starting in a node at the end of the tree, the number of ways to construct this tree is equal to one. The line tree is the example of tree based in a complete graph with smaller probability because has a smaller number of different sequences that forms the same tree starting in a fixed node, that is one.

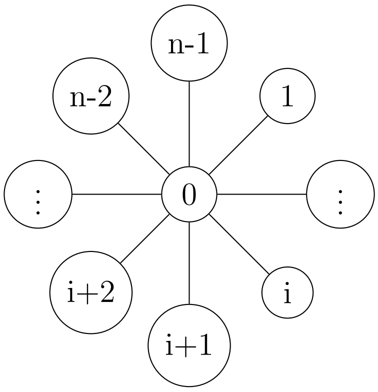


Figure 4.3: Star tree - τ^{Star}

Using Equation 4.5 and considering the star tree in the example, we fix the root in the node zero. The size of each sub-tree hung in every neighbour of 0 will be equal to one because is a star tree. So, we have that:

$$\begin{aligned} |B(\tau_0)| &= \frac{(n-1)!}{\prod_{i\neq 0} |\tau_i^0|} \\ &= (n-1)! \end{aligned}$$

This way, by comparing the probability 4.2 in both trees, we have that:

$$\begin{aligned} P(\tau^{Line}|v_0 = 0) &= \frac{1}{[(n-1)!]^2} \\ P(\tau^{Star}|v_0 = 0) &= \frac{1}{(n-1)!} \end{aligned}$$

We can use those two examples as a lower and upper bound respectively to complete graphs. Note that, despite the probability of any sequence in the complete graph be equal, the number of ways to construct each one is very different as we can see in the probabilities above, because the number of ways to construct the tree is very different.

4.4 The Most Likely Source

As we can see from Equation 4.6, the first part is constant depending only on the size of the network and the second part depends on the sub-trees size. To maximize this probability we must minimize the denominator of this ratio so that smaller the value of the denominator greater the value of the ratio since the numerator is constant equal to one. This way, we need minimize the product of the sub-tree sizes and find the node that makes this product as small as possible. Overall, we reduce to

$$\arg\min_{r} \prod_{i \neq r} |\tau_i^r|.$$

In order to find a way to calculate this arg min more quickly, we will use the next three lemmas which will be of extreme importance for understanding the algorithm.

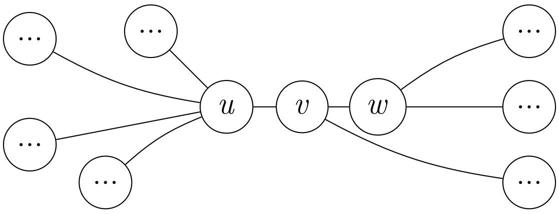
Lemma 4.4.1. *Given u and v, neighbors in the tree τ,*

$$|\tau_w^v| = |\tau_w^u| \quad \forall\, w \in V(\tau)\backslash u, v;$$

Proof:

Since the size of any tree is the number of nodes it has, the size of τ_w^k is the number of nodes below w in the sub-tree of τ^k hung in node w. Given u and v, neighbors in the tree τ and a node w, $w \neq u, v$, there are two possibilities: either $w \in \tau_v^u$ or $w \in \tau_u^v$. As there is only one path that connects the nodes u to w, and this path passes through v, the size of the sub-tree hung in the node w will be the same if the tree has root in the node u or v.

The tree in Figure below. will leave the idea clearer to the reader.



Lemma 4.4.2. *Given u and v neighbors in the tree τ*

$$\frac{P(\tau|v_0 = u)}{P(\tau|v_0 = v)} = \frac{|\tau_v^u|}{|\tau_u^v|}.$$

Proof:

From Equation 4.6 we have:

$$\frac{P(\tau|v_0 = u)}{P(\tau|v_0 = v)} = \frac{\frac{1}{(n-1)!}\frac{1}{\prod_{i \neq u}|\tau_i^u|}}{\frac{1}{(n-1)!}\frac{1}{\prod_{i \neq v}|\tau_i^v|}}.$$

As the first part of both probabilities are constant and equal, the ratio of these two probabilities will be equal to:

$$\frac{P(\tau|v_0 = u)}{P(\tau|v_0 = v)} = \frac{\frac{1}{\prod_{i \neq u}|\tau_i^u|}}{\frac{1}{\prod_{i \neq v}|\tau_i^v|}}$$

$$= \frac{\prod_{i \neq v}|\tau_i^v|}{\prod_{i \neq u}|\tau_i^u|}.$$

From Lemma 4.4.1 we can write the sizes of the sub-trees differing by only one term as below

$$\begin{aligned}\frac{P(\tau|v_0=u)}{P(\tau|v_0=v)} &= \frac{\prod_{i\neq u.v}|\tau_i^v|\cdot|\tau_u^v|}{\prod_{i\neq u.v}|\tau_i^u|\cdot|\tau_v^u|}\\ &= \frac{|\tau_u^v|}{|\tau_v^u|}.\end{aligned}$$

Note that this Lemma serves to choose the node with larger probability when comparing two nodes. But, when the tree has a even number of nodes, two nodes can have the same sub-tree size that is half the number of tree nodes, so there may be two neighbors u and v belonging to τ such that $|\tau_v^u| = |\tau_v^v| = {}^{|\tau|}/_2$ this implies that

$$P(\tau|v_0=u) = P(\tau|v_0=v).$$

Lemma 4.4.3. *Let τ be a tree and r^* be such that $|\tau_{r^*}^w| \geq \frac{|\tau|}{2}$, $\forall\, w \in V(\tau)$. Then,*

$$\prod_{i\neq r^*}|\tau_i^{r^*}| \leq \prod_{i\neq r}|\tau_i^r|\;\forall\, r \in V(\tau)$$

Proof:

Let us assume, towards a contradiction, that $\exists\; r \in V(\tau)$ such that,

$$\prod_{i\neq r}|\tau_i^r| < \prod_{i\neq r^*}|\tau_i^{r^*}| \tag{4.7}$$

Let p denote the unique path in τ connecting r and r^*; we shall denote by $V(p)$ and $E(p)$, the vertices and edges of p.

Fact 1: $\forall\; w \in V(\tau)\backslash V(p)$ it holds that $|\tau_w^r| = |\tau_w^{r^*}|$ thus, Equation 4.7 reduces to

$$\prod_{\substack{i\in V(p)\\ i\neq r}}|\tau_i^r| < \prod_{\substack{i\in V(p)\\ i\neq r^*}}|\tau_i^{r^*}|$$

The last equation can be written as

$$\left(\prod_{i\neq r,r^*}|\tau_i^r|\right)|\tau_{r^*}^r| < \left(\prod_{i\neq r,r^*}|\tau_i^{r^*}|\right)|\tau_r^{r^*}| \tag{4.8}$$

If $|\tau_{r^*}^r| < |\tau_r^{r^*}|$, then, to Equation 4.8 be true, must exist at least one i so that

$$\prod_{i\neq r,r^*}|\tau_i^r| < \prod_{i\neq r,r^*}|\tau_i^{r^*}|$$

Fact 2: Note that $|\tau_r^{r^*}| + |\tau_{r^*}^r| \leq |\tau|$ and due to the hypothesis that $|\tau_{r^*}^w| \geq \frac{|\tau|}{2}$ $\forall w$, we have that

$$|\tau_r^{r^*}| \leq |\tau| - |\tau_{r^*}^r| \leq \frac{|\tau|}{2} \leq |\tau_{r^*}^r|$$

Thus, if 4.8 holds $\implies \exists i \in V(p)\ i \neq r, r^*$ such that

$$|\tau_i^r| < |\tau_i^{r^*}|$$

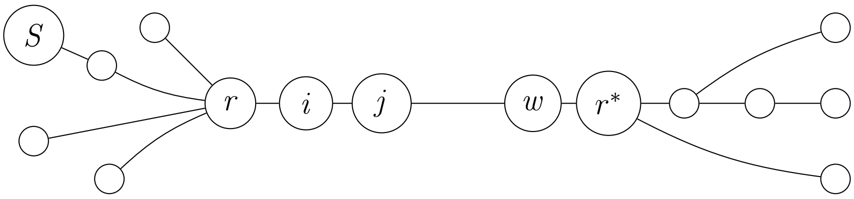
whenever this is the case, we should have that

$$|\tau_j^r| < |\tau_j^{r^*}|$$

$\forall\ j \in p$, were p is the path connecting i to r^*, thus in particular we must have that

$$|\tau_{w^*}^r| < |\tau_{w^*}^{r^*}| \tag{4.9}$$

were $w^* \in p$ and $w^*, r^* \in E(p)$ (w^* is a neighbor of r^* in p).



However, if w^* is a neighbor of r^* in p, it holds that

1) $|\tau_{w^*}^{r^*}| = |\tau| - |\tau_{r^*}^{w^*}| \leq \frac{|\tau|}{2}$

2) $|\tau_{w^*}^r| \geq \frac{|\tau|}{2}$

Thus, if 4.9 holds it must also be the case that

$$\frac{|\tau|}{2} \leq |\tau_{r^*}^r| < |\tau_{w^*}^r| < |\tau_{w^*}^{r^*}| \leq \frac{|\tau|}{2}$$

a contradiction!

Lemma 4.4.4. *Given a tree τ it always exists at least one vertex r^* satisfying $|\tau_{r^*}^w| \geq \frac{|\tau|}{2}$ $\forall w \in V(\tau)$.*

Proof:

Given a tree, a vertex is randomly selected as root to give orientation to this tree and then we are able to calculate the sub-tree sizes starting on the leaf nodes. As explained before, the size of a sub-tree is the number of nodes that this sub-tree has, this way leaf nodes has sub-tree size equal one.

We start to calculate the number of nodes in each sub-tree by the leaves and iteratively "leveling up" in the tree. For each calculated sub-tree, its respective number of nodes is added to the node that is hanging, i.e., if the sub-tree hung on a has four nodes and the sub-tree hung on b has five nodes, the sub-tree hung on c has nine nodes if the child nodes of c was a and b. Using the information of the sub-tree size of child nodes we avoid recalculate several times how many nodes has at the levels below. This way we calculate the sub-tree sizes by level in the tree and as the tree has size equal to $|\tau|$ always exist a node that has sub-tree size bigger than $|\tau|/2$ satisfying this condition of existence of a node with size $|\tau|/2$.

Observing the graph in Lemma 4.4.3, let s be the initial vertex were the tree begins and r^* the identified source. We know that $\tau^s_{r^*} \geq \frac{|\tau|}{2}$ and the size of all sub-trees hung in vertices in $\tau^{r^*}_s$ are smaller than $\frac{|\tau|}{2}$. To analyse the correctness proof we want to look for these two cases presented below.

Case 1: $v \in V(\tau)\backslash\tau^s_{r^*}$

If $\tau^s_{r^*} \geq \frac{|\tau|}{2}$ we know that for all $v \in V\backslash\tau^s_{r^*}$ the sub-tree size $|\tau^{r^*}_v| < \frac{|\tau|}{2}$, than v cannot be the vertex which maximizes Equation 4.6.

Case 2: $v \in V(\tau^s_{r^*})$ and neighbor of r^*

If $\tau^s_{r^*} \geq \frac{|\tau|}{2}$ than, between s and r^* has less than $\frac{|\tau|}{2}$ vertices. So, for a tree with root in r^* hung in a vertex in the neighborhood of r^*, $|\tau^{r^*}_v| < |\tau^s_{r^*}| \leq \frac{|\tau|}{2}$, then $|\tau^{r^*}_v| < \frac{|\tau|}{2}$. This is valid for all neighbors of v and neighbors of neighbors, so, v cannot be the vertex which maximizes Equation 4.6 and r^* satisfy Lemma 4.4.3.

4.5 Source Identification Algorithm

The algorithm takes as input the spanning tree that represents the epidemic without temporal labels, and it solves the optimization problem presented in Lemma 4.4.3. To do this, we choose a node uniformly at random from all nodes of the spanning tree and use a BFS to induce orientation to this spanning tree. The information that we use from the BFS is the number of child of each node and its parent. With these two information we are able to perform a post-order and calculate the size of the sub-trees.

It is worth remembering that the size of each sub-tree is the number of vertices that it has and this calculations are recursively made until all nodes are calculated or satisfy the stopping criterion. To calculate the number of nodes of each sub-tree, we need to set the leaf nodes, that is, the nodes with no child nodes. Selecting a node from the leaf list, we add a unit to the size of its sub-tree and its size to the sub-tree size of it respective parent. After this, we decrease a unit in the number of child nodes of its parent and, if the number of child is zero, this parent is added to the leaf list, if not, we select a new node from the leaf list until the list be empty or, at some point, the stopping criterion is satisfied. The stopping criterion is to find node whose sub-tree size is at least half the vertices. This allows us identifying the node as the node whose probability is larger than every others in the tree.

The pseudo-code used in this thesis will be presented below and the symbols used in the pseudo-code are described in the table below:

Variable	**Significance**
$\|\tau\|$	Number of vertices in the tree τ;
τ_r	Spanning tree rooted in the node r;
τ_i^r	Sub-tree of τ_r hung on node i;
$Parent[i]$	Unique node neighbor of i in the tree and at one level higher than i
$Child[i]$	All i neighbors that are one level below the level of i;

The pseudo-code follows some steps to find the epidemic source. The first step consists in choosing a node u uniformly at random in τ as root - see line 2 of Algorithm 1. In the second step, a BFS algorithm is used to induce orientation in the spanning tree received as input. This BFS will return two lists, one with the Parents and another with the number of Children of each node as represented in line 3 of Algorithm 1. After creating these lists of parents and child, all nodes with no child will be added in a list of leaves, lines 4-7 in Algorithm 1. In the third step, the size of the sub-trees are calculated from the leaves to the root.

During the calculations of a sub-tree size hung on a node i, we count the number of nodes hung on i, add in the sub-tree size of the parent of i this number and decrements the number of children that the parent of i have. This way, we know when all the children nodes have its sizes of sub-tree calculated. When this occurs, the node with no children is added to the leaves list. When there is no node to calculate the sub-tree size, the *While* condition is satisfied.

Now, if during the calculations of sub-tree sizes, Lemma 4.4.3 is satisfied, that is, when the tree whose size is half the number of nodes (or this value added/decreased one unit) is found, the If within this While is satisfied. When the condition of this If is satisfied, the leaves list is emptied so that it also satisfies the While condition and completes the loop and the node that satisfied the If condition is declared as an epidemic source (r).

Algorithm 1: Root Identification of Recursive Random Spanning Trees

```
Data: Epidemic tree ($\tau$).
begin
    u = any node in $\tau$;
    $Parent, Child$ = bfs($\tau$, u);
    $Leaves = \emptyset$;
    $Sizes[\cdot] = 0$;
    for $i \in V(\tau)$ do
        if $Child[i] == 0$ then
            $Leaves.push(i)$;
    end
    while $Leaves \neq \emptyset$ do
        $i = Leaves.pop()$;
        $Sizes[i]$ += 1;
        $Sizes[Parent[i]]$ += $Sizes[i]$;
        $Child[Parent[i]]$ -= 1;
        if $Sizes[i] \geq \tau/2$ then
            $root = Parent[i]$;
            Break;
        if $Child[Parent[i]] == 0$ then
            $Leaves.push(Parent[i])$;
    end
    return root;
end
```

The proof of the algorithm correctness is exactly the proof of Lemma 4.4.3 and 4.4.4.

Chapter 5

Numerical evaluation

The numerical evaluation used the source identification algorithm 1 taking as input the spanning tree which represents the spread of an epidemics through a complete graph. This spanning tree was obtained using the classical recursive random trees algorithm described below. Note that the node label corresponds to the order the node was added to the tree. Thus, node 0 is the epidemic root and node i is the i-th node to be infected or equivalently join the tree.

Algorithm 2: Recursive Random Spanning Trees

```
  Data: Number of nodes of the spanning tree ($n$).
1 begin
2     G = Graph();
3     G.add.node(0);
4     for i=0 to n-1 do
5         node = random.uniform(0, i+1);
6         G.add.node(i+1);
7         G.add.edge(node, i+1);
8     end
9 end
```

Considering that the graph representing the population is a complete graph, each new individual added to the tree has equal probability of being connected to any other node already belonging to the tree. This means that, for each new contamination, the individual has equal probability of being infected by any of the individuals already infected in the epidemic. Thus, line 5 of the algorithm chooses this node uniformly at random.

5.1 Numerical evaluation of relative frequency of identified sources

We simulate an epidemics through a population of 10 individuals. Figure 5.1 shows the relative frequency in which each one of the individuals was identified by Algorithm 1 as the epidemic source. That is, the number of times that each individual was returned as the epidemic source divided by the 100000 rounds of independent simulations.

The individuals are labeled from zero to nine representing the time in which they were infected by the epidemic. Note that, as expected, the nodes labeled as zero and one have relative frequency almost identical because these two nodes belong to the first edge in the epidemic and as the epidemics are represented by edges, these two individuals have the same probability. In fact, the structure of the tree τ is identical from the point of view of each of these nodes, statistically speaking. Differences in their relative frequencies are due to simulations.

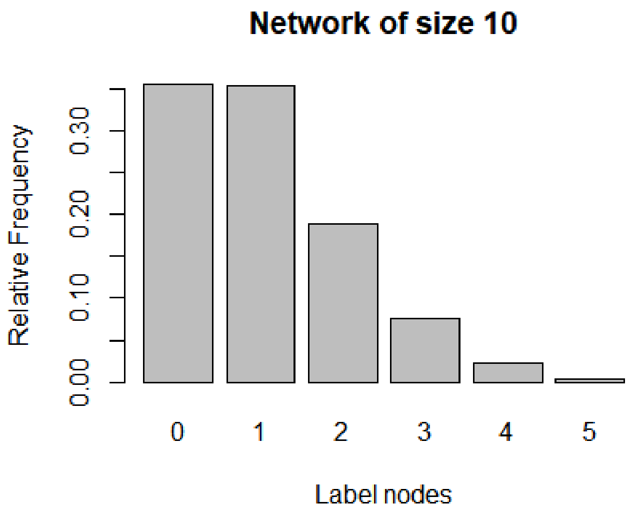


Figure 5.1: Relative frequency with which a node is identified as the epidemic source.

The first bar shows the relative frequency at which the algorithm actually found the true epidemic source, i.e., the node with label zero. This shows that almost 35.4% of the time the algorithm finds the true source of the epidemic, and also 35.4% choose as source the first node to be infected in the epidemic. Thus, if we add the relative frequency values we can see that the method identifies the nodes incident to the first infected edge about 70% of the time showing a good rate of accuracy in the identification of the real epidemic source. We can also observe the monotonic behavior of the chart showing that as the node "enters the epidemic later it is less likely to be identified as the source" since its labels reflect the moment of entry into the epidemic tree. Note that nodes with label larger than 5 (which entered after half the tree was generated) are never identified as the source.

The histogram in Figures 5.2 and 5.3 show the relative frequency for larger trees with sizes equal to 100, 300, 900, 2000, 5000 and 10000 nodes. In these plots, the x-axis represents the label of the nodes, that is, their order of infection in the epidemic tree, while the y-axis represents the relative frequency with which these nodes were identified as the epidemic source by the proposed algorithm.

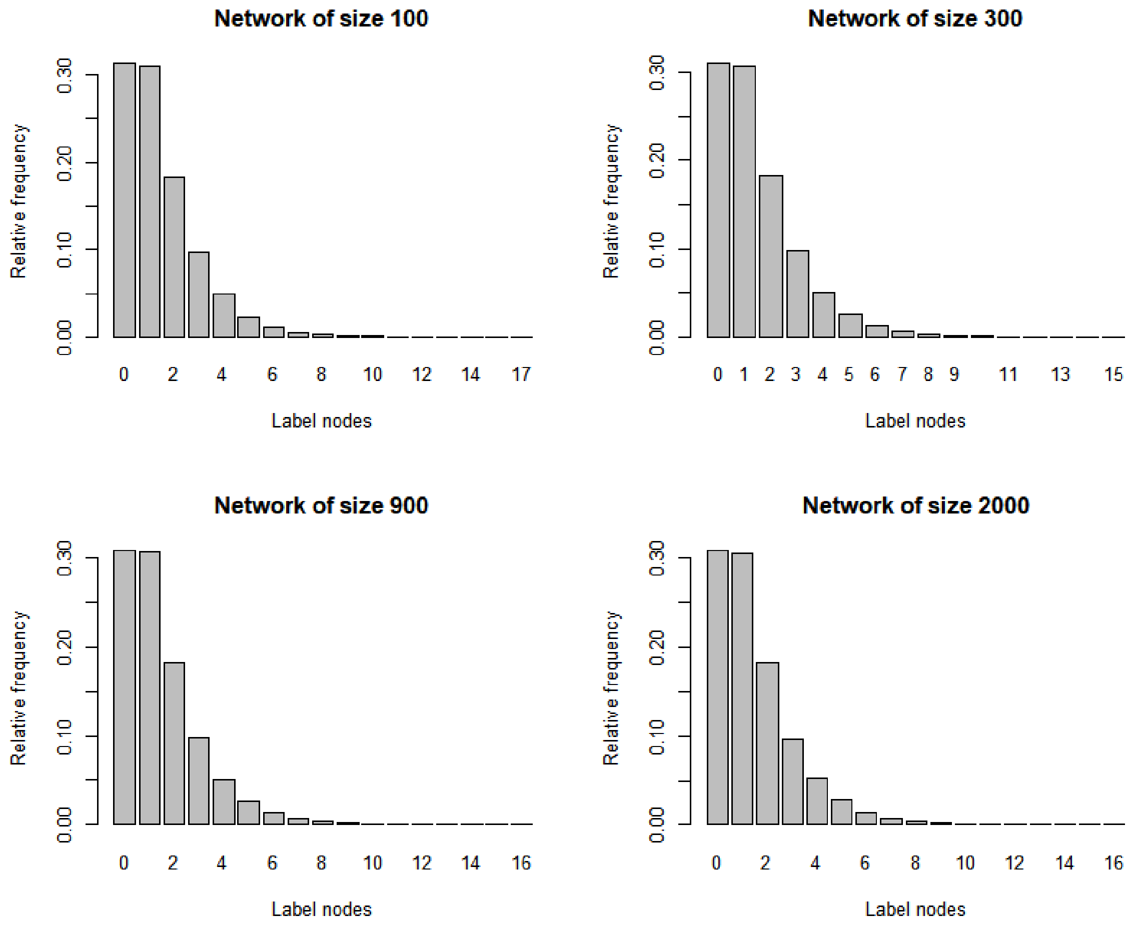


Figure 5.2: Relative frequency with which a node is identified as the epidemic source by the algorithm (average over 100000 rounds).

From these plots it is clear that the nodes with the high relative frequencies are the ones with the smallest labels, that is, the ones that entered the tree first. In all scenarios, we observe the same monotonic behavior and an exponential decay of the relative frequency, showing that, in fact, the algorithm returns much more often the first nodes to be infected, and nodes that were infected after half of the tree has been constructed are never returned; note that in our simulations nodes with label higher than 20 are never returned.

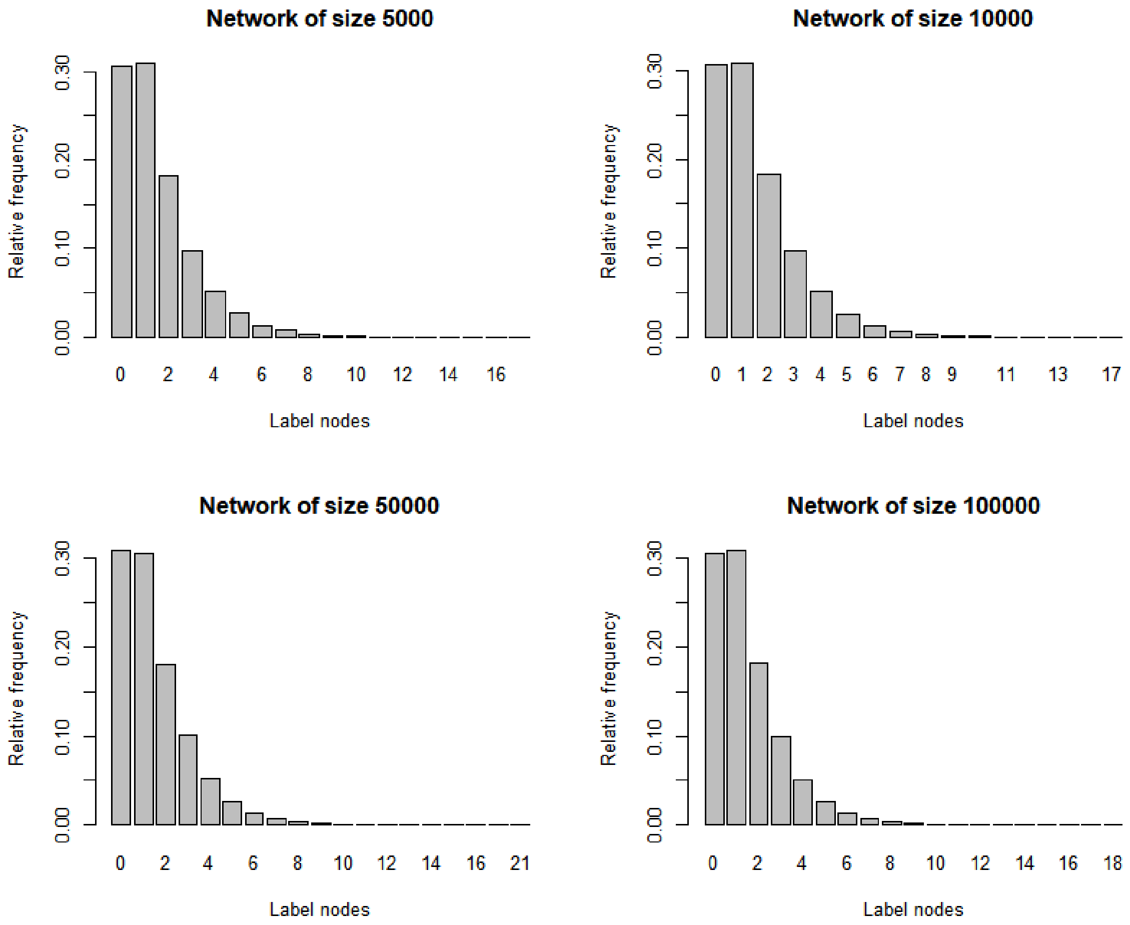


Figure 5.3: Relative frequency with which a node is identified as the epidemic source by the algorithm (average over 100000 rounds).

Figure 5.4 shows the relative frequency of correct identification of the proposed algorithm with 100000 iterations. Note that this relative frequency is around 30% regardless of the network size. This is important because it indicates that the performance of the proposed algorithm does not decay as the tree grows. However, the performance also does not improve as the tree grows in size. The values of relative frequency for the first 5 infected nodes of the realized simulations are presented in Table 5.1. Note that for all networks the nodes with label zero and one, as expected, have the same probability and as the label increases, the frequency drops in half in each step.

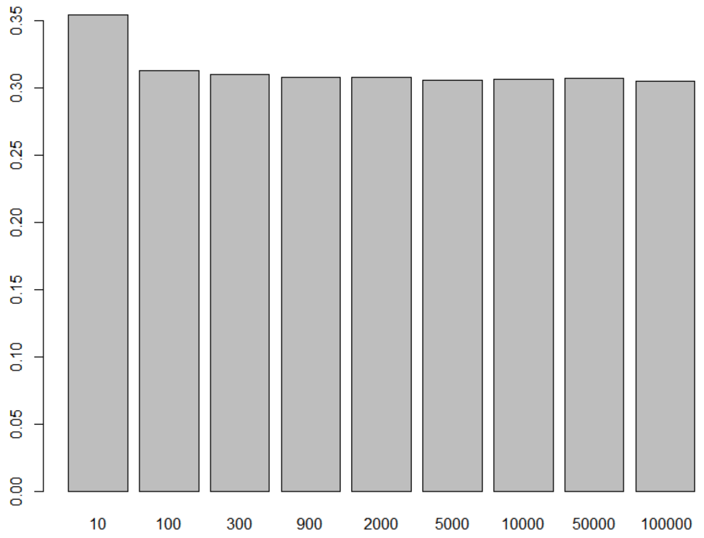


Figure 5.4: Frequency with which node zero (the actual source) is identified as epidemic source for different network sizes.

Is important emphasize that as the algorithm works using edges the probability of returning the epidemic source and the the first node to be infected is the same since they share the first edge of the contamination process. Note that the probability of returning the source of the epidemic or the first infected node is about 60%, independently of the network size and with probability 97% the algorithm returns the first 5 infected nodes in the epidemic, shown in the *Accumulated* column in Table 5.1.

Table 5.1: Frequency with which the nodes are returned as source in networks with different sizes.

Nodes	**0**	**1**	**2**	**3**	**4**	**5**	**Acumulated**
size 10	0.354	0.353	0.187	0.076	0.022	0.004	0.996
size 100	0.313	0.310	0.183	0.097	0.049	0.023	0.975
size 300	0.310	0.307	0.183	0.097	0.050	0.025	0.972
size 900	0.307	0.306	0.182	0.097	0.050	0.027	0.969
size 2000	0.308	0.305	0.182	0.096	0.052	0.027	0.970
size 5000	0.305	0.309	0.181	0.097	0.051	0.026	0.969
size 10000	0.306	0.308	0.182	0.097	0.051	0.026	0.970
size 50000	0.307	0.304	0.180	0.101	0.051	0.026	0.969
size 100000	0.305	0.308	0.182	0.099	0.050	0.026	0.970

As the algorithm returns nodes with labels different from zero and one, it is worth understanding the distance that these nodes have from the real epidemic source (node zero). This distance is calculated by the number of nodes in the path between the identified source and node zero and can be interpreted as the error when identifying the epidemic source. This measure is interesting because it gives us information on the distance between the source identified by the algorithm and the real epidemic source.

5.2 Distance distribution on source identification

The purpose of analysing the distance distribution is to understand "how close" the source returned by the proposed algorithm is to the real epidemic source.

Since the epidemic is represented by a spanning tree, we know by definition that a path connecting the real epidemic source and the source identified by the algorithm is unique. This distance $d(0, v_0)$ is calculated by running a BFS in node 0 (the true source) and checking the level of the node identified by the algorithm on this tree. Algorithm 3 describes this procedure.

For each one of the networks with sizes 100, 300, 900, 2000, 5000 and 10000 and for each one of the 100000 epidemics simulated for each network size, we have an associated distance. With this values we calculate the distribution of these distances. This empirical distribution allows us to estimate the probability that the node identified by the algorithm is at distance k from node zero, the real epidemic source.

Algorithm 3: Distance between the 0 and the source identified.

```
Data: Epidemic tree ($\tau$);
      $v_0$ = identified source.
begin
    if $v_0 = 0$ then
        return 0;
    Parent = bfs(0);
    Distance = 1;
    $i = v_0$;
    while $Parent[i]! = 0$ do
        Distance += 1;
        $i = Parent[i]$;
    end
    return Distance
end
```

Figure 5.5 shows the distance on the tree between the node identified as source and node zero, the actual epidemic source.

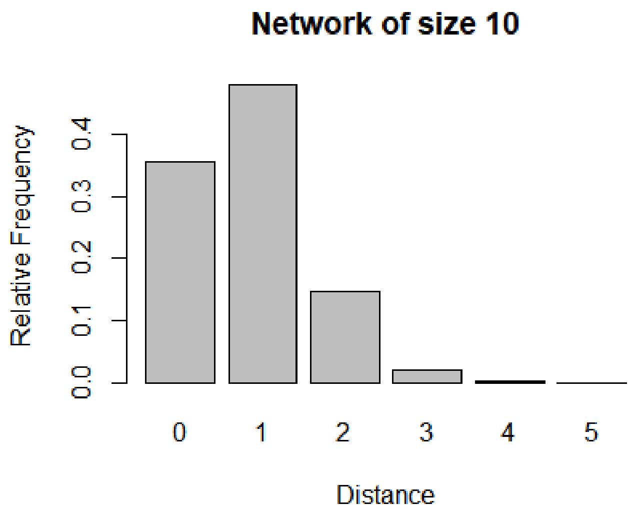


Figure 5.5: Relative frequency of the distance of the identified epidemic source to the real epidemic source for a tree with size 10.

Figure 5.5 shows that the nodes with distance one from the real source have a larger relative frequency than the true source (distance zero). This occurs because this frequency accumulates the frequency with which the node with label one is returned as source plus the cases in which a neighbor of the true source is returned. As the node with label one has the same probability to be identified as the epidemic source as node zero, it is expected that the relative frequency of nodes with distance one (the neighbors of the real epidemic source) is larger. The Table 12 present the values of the nodes returned by the algorithm and distances from the true source.

Table 5.2: Identification and distances of nodes in a network of size 10.

	0	**1**	**2**	**3**	**4**	**5**
Identification	0.354	0.354	0.188	0.077	0.023	0.004
Distance	0.354	0.479	0.145	0.021	0.001	0.00004

Figure 5.6 presents the relative frequency with which the identified source is at distance k from node zero. The relative frequency is calculated by the number of distances equal to k divided by the number of iterations made in the simulations for each network size.

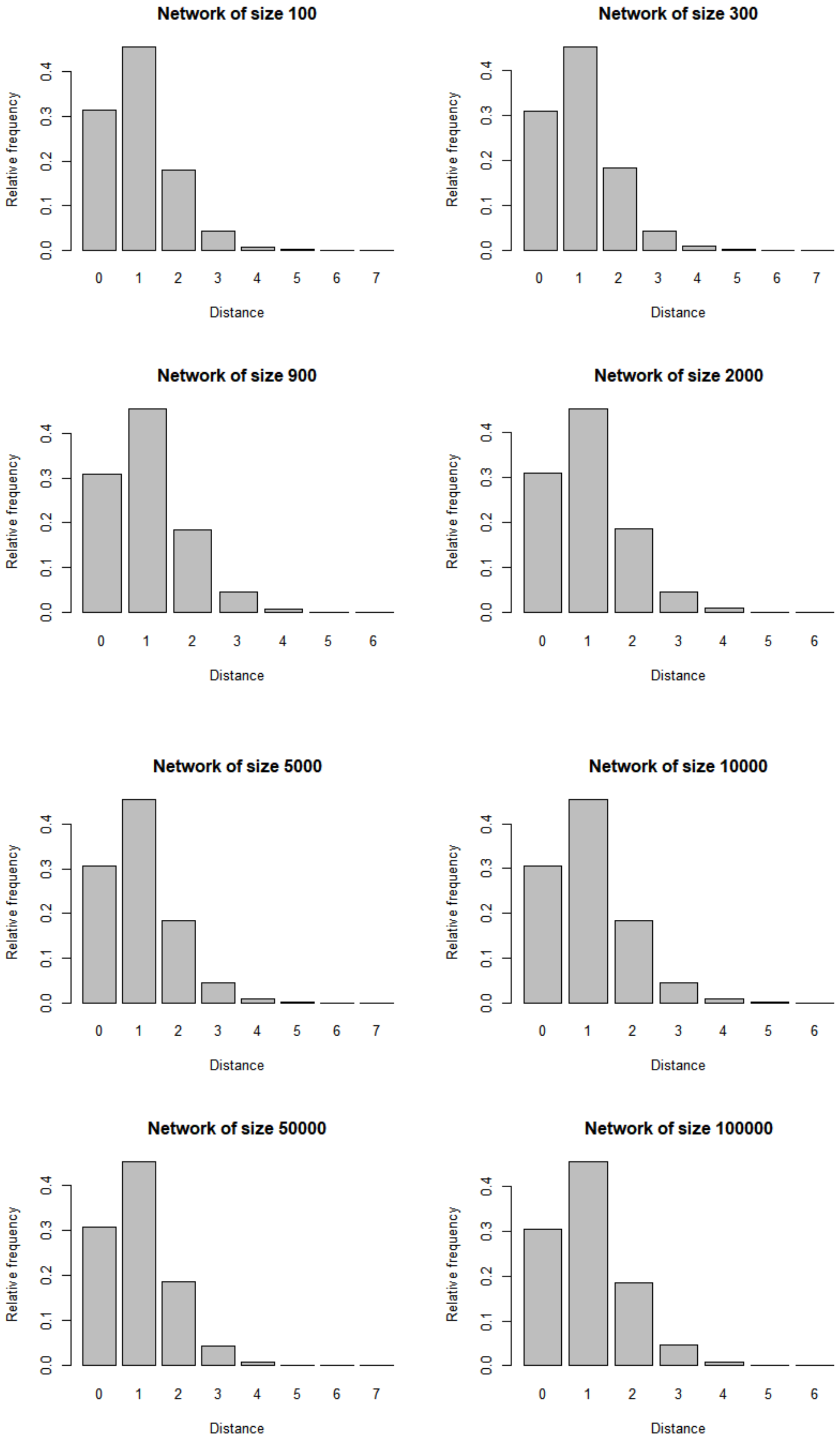


Figure 5.6: Relative frequency of the distance of the node identified as the source and the real epidemic source.

As expected, the nodes with distance one of the real epidemic source have a higher relative frequency because the first node infected in the epidemic has the same probability to be chosen as the epidemic source. This behavior is observed in all networks regardless of their size. Table 5.3 presents the mean and standard deviation of the relative frequency of distance 5 from the real epidemic source (node zero). Note that the probability decays very fast (exponentially) with the distance for all network sizes.

Table 5.3: Mean and SD of the distance distribution.

	0	**1**	**2**	**3**	**4**	**5**
size = 10	0.35 (0.28)	0.48 (0.35)	0.15 (0.13)	0.02 (0.02)	1*10-3 (1*10-3)	4*10-5 (4*10-5)
size = 100	0.31 (0.26)	0.46 (0.34)	0.18 (0.16)	0.04 (0.04)	0.01 (0.01)	0.001 (0.001)
size = 300	0.31 (0.26)	0.45 (0.34)	0.18 (0.17)	0.04 (0.04)	0.01 (0.01)	0.001 (0.001)
size = 900	0.31 (0.26)	0.45 (0.34)	0.18 (0.17)	0.04 (0.04)	0.01 (0.01)	0.001 (0.001)
size = 2000	0.31 (0.26)	0.45 (0.33)	0.19 (0.17)	0.05 (0.04)	0.01 (0.01)	0.001 (0.001)
size = 5000	0.31 (0.25)	0.45 (0.34)	0.18 (0.17)	0.05 (0.04)	0.01 (0.01)	0.001 (0.001)
size = 10000	0.31 (0.26)	0.45 (0.34)	0.18 (0.17)	0.05 (0.04)	0.01 (0.01)	0.001 (0.001)

Note that, similar to (BUBECK, 2017), we can calculate a set of possible sources. Here, this set accumulating $1-\epsilon$ of probability, estimates that the identified epidemic source will be within k steps of the actual epidemic source. For example, if we want a set of node with probability 0.99 to contain the real epidemic source, the identified epidemic source will be until 3 hops from the real epidemic source for networks of size equal to 10000 nodes.

5.3 Runtime

This Section present the runtime of the simulations including just the time to run the algorithm. In Table 5.4 we can see the mean and standard deviation of runtime in seconds for all the sizes of network returned from the simulations in 100000 rounds (just the time to run the source identification algorithm). For each network size, these are the average execution times for the 100000 iterations performed. Note that the runtime increases almost linearly as the network size increases.

Figure 5.8 presents the mean of the runtime over 100000 iterations for all tested network sizes. In the x-axis we have the size of these networks and in the y-axis we have the average runtime in seconds.

Table 5.4: Mean and standard deviation of runtime of source identification algorithm in the simulation.

	Mean	**SD**
size = 10	4.525*10-05	1.934*10-05
size = 100	3.033*10-04	1.247*10-04
size = 300	9.673*10-04	2.884*10-04
size = 900	2.869*10-03	6.726*10-04
size = 2000	6.589*10-03	2.848*10-03
size = 5000	1.668*10-02	6.632*10-03
size = 10000	4.654*10-02	1.317*10-02
size = 50000	5.422*10-01	70.905
size = 10000	9.508*10-01	72.884

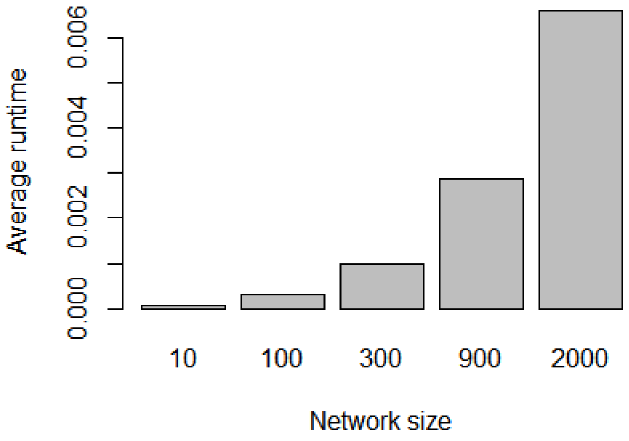


Figure 5.7: Barplot of the average runtime.

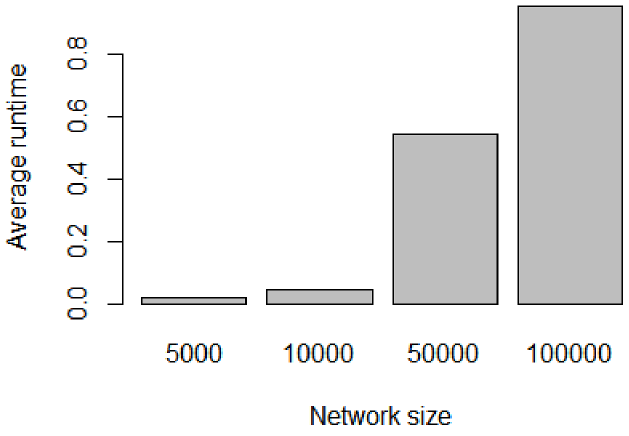


Figure 5.8: Barplot of the average runtime.

Chapter 6

Related Work

This chapter presents a brief review of articles published in the last 20 years concerning epidemic models on networks and rumors source identification. As in this thesis, the susceptible-infected epidemic model, is the focus of this brief review.

The seminal of work of Vespignani, Pastor-Satorras, et. al. (2004) models the spread of information in networks using epidemic spreading models, such as the susceptible-infected (SI) and susceptible-infected-recovered (SIR) models. Their work showed the importance of the network structure on the epidemic process and in particular the degree distribution specially when it follows a power law. This highly influential paper opened the doors for many subsequent works.

The work of Shah and Zaman (2010) was among the first to model and analyse mathematically the rumor source identification. Their seminal work considers SI epidemics on infinite trees and the observation of infected nodes after a long time period (as well as knowledge of the tree).

Below, we provide a summary of some of the main articles addressing rumor source identification in network epidemics and briefly discuss the main differences with the model studied in this thesis.

The work of Shah and Zaman (2010) presents the first steps towards modeling the question of identifying the rumor source in infinite networks based on infected nodes of an SI epidemic. The authors proposed a node centrality metric called rumor centrality to estimate the rumor source nodes showing that the node with maximal rumor centrality is the maximum likelihood estimator (MLE) of the source of the epidemic. Another main contribution is a linear time message-passing algorithm to compute the rumor centrality for every node in the network (tree).

Xu, Peng, et. al. (2006) present a modified susceptible-infected-susceptible (SIS) model where the observation of an infection suffers a random delay. Considering different network topologies and both uniform and degree-dependent delays, the contagion process enhanced and there are more prevalent infectious in the network. They focus on network topology reconstruction and not epidemic source identification.

The subsequent work of Shah and Zaman (2012) shows that as the size of infected graph increases, the probability of source identification remains bounded away from 0 and $\frac{1}{2}$ depending on the graph topology. This result assumes that the epidemic model is the Susceptible-Infected (SI) with exponential infection times. Simulations are used to compute this value for specifies cases.

Dong, Zang and Tan (2013) approach the study of rumor source identification with a set of suspects (as opposed to a single node) conditioning on an observed subset of infected nodes in the network. The goal is to identify the source based on the network structure and the subset of infected nodes observed. The main result is the calculation of the probability of correct detection using a set of suspects of the population varying its size and comparing results.

Zheng and Tan (2015) consider a probabilistic approach to rumor source identification like Shah and Zaman (2010) e Shah and Zaman (2012). Differently from Shah and Zaman, they characterize the boundary of the rumor center given the graph, and the infected nodes. The goal is to infer the rumor source using the network topology and the boundary of the rumor graph. A main contribution is the probabilistic analysis of the rumor boundary in terms of the graph connectivity properties and the observation of infection time, as well as the formulation of the maximum likelihood estimation problem and the proposal of a distributed message-passing algorithm to solve it. Their model is based on differential equations with several states that represent the epidemic process.

Lugosi, Devroye and Bubeck (2017) investigate algorithms to find the epidemic source of large trees generated by either the uniform attachment or preferential attachment model. The algorithm outputs a set of K nodes, such that, with probability at least *1*-ϵ_k, the epidemic source belongs to this set.

Antulov-Fantulin, Lančić, et. al. (2015) use exact analytical calculations and Monte Carlo simulations to demonstrate the limits for correctly identifying the rumor source in the Susceptible-Infected-Recovered (SIR) model. They also demonstrate their approach in the simulation of a sexually transmitted infection spreading over a temporal network of sexual interactions.

In a more recent paper, Shah and Zaman (2016) overcome the limitations of their previous articles and establish the effectiveness of rumor centrality for source identification for generic random trees and an SI model with generic infection time distribution. The main result is an interesting connection between a continuous time branching process and the effectiveness of rumor centrality, as well as an estimation of the probability for correctly identifying the epidemic source.

Yu, Tan and Fu (2017) study network boundary effects and the message-passing algorithm in arbitrary graphs, solving the constrained maximum likelihood estimation problem using a generalized rumor centrality metric. They propose a message-passing algorithm that is near-optimal for graphs with more complex boundaries consisting of multiple end vertices, i.e., the susceptible nodes with only a single neighbour.

Melnyk and Styopochkina (2019) investigate the problem of malicious information source detection among the users of online social networks. They analyse the advantages and disadvantages of existing algorithms of rumor source detection in rest data of information spread, using as baseline the message-passing of Shah and Zaman (2010).

Lugosi and Pereira (2019) assume that the epidemic starts with a small graph (as opposed to a single node), and consider the problem of finding the source tree in large observed tree. They determine when is possible to identify this source and the role of the initial tree structure on the difficulty of identification problem. They consider three types of initial trees: paths, stars, and small random uniform recursive trees.

Chapter 7

Conclusion

This thesis addressed the problem of identifying the source of random epidemics in finite networks. The environment through which the epidemic spread, that is, the population, is represented by a graph where the nodes represent the individuals of this population and the edges their interactions. We focus on the most common representation of the population, i.e., complete graphs, which encodes the assumption that the population in homogeneously mixed.

The epidemic model used is the SI model where individuals can only be in one state at time, susceptible or infected, and once infected the individual remains infected ever since. This model was chosen because it is able to capture the process of information diffusion in networks; once one becomes aware of the information, it is no longer lost. It is important to emphasize that in this model the entire population will be infected after some time, since the network is connected. Thus, when an epidemic occurs in a graph, as a individual cannot be infected twice, the result of this epidemic can be represented by a spanning tree because a rooted spanning tree must have all the nodes in the graph and each node has a parent which is the node that infect it.

To model the spread of an epidemic, in this thesis, we use the recursive random spanning tree model and its algorithm to simulate the epidemic. This algorithm uses a complete graph as underlying graph and returns a labeled spanning tree that represents the epidemic occurred in the population, where the labels encode the infection times of each individual (node).

Removing the labels of the resulting spanning tree that allow us to know where the epidemic starts and the exact path the epidemic through the population, we have only a unlabeled tree with the edges of the contamination process. Our aim was, using only these edges as input information, identify where the epidemic started. To do this, we represent the spanning tree by rooted edge sequences and calculate all the possible combinations that can generate the corresponding tree, and compare these numbers. This edge sequence is made by fixing one node as root and finding the possible edge sequences that respect the structure of the spanning tree; after that, the node with the highest number of combinations is the most likely node to be the source.

The calculation considers all possible combinations of edges to construct the epidemic tree must respect the order of contamination, that is, an edge connecting a leaf node cannot be added in the sequence before the edges connecting all nodes in the path between the fixed root and this leaf node. Mathematically, this calculation is based on the the size of the population and the product of the sub-tree sizes. As the size of the population is constant, the product of the sub-tree sizes is the part of the equation that we can compare to infer which node will be the most likely epidemic source.

This thesis proposes an efficient (linear time) algorithm that uses a stopping criteria to ensure that the returned node is the one that maximizes the probability of being the source. This stopping criterion returns the first node that has as sub-tree size at least half of the nodes of the epidemic tree. In Chapter 4 we prove that this stopping criterion, in fact, returns the most probable node to be the epidemic source, independent of the orientation of the tree.

As it turns out, the numerical evaluation of the algorithm shows that the probability of identifying the real epidemic source converges to 30%. Since the tree description is given by edge sequences, the algorithm identifies with equal probability the real epidemic source and the first infected node. This way, the probability of identifying the real epidemic source or the first infected converges to 60%. Beside that, we can observe that at each unit of time in the contamination process, the frequency with which the node is identified as epidemic source falls by half comparing to the previous time, indicating ab exponential decay.

As some nodes with higher labels where identified as epidemic source by our algorithm, we analyse the distance (in the tree) that the node identified has from the actual epidemic source (node with label zero). The higher frequency was in the nodes at distance one, that is, the neighbors of the actual source in the epidemic tree, showing that the algorithm returns nodes that aare close (in the tree) to the actual source.

7.1 Future Work

A natural extension of the work presented in this thesis that would be interesting to address is the identification of the most probable epidemic source in other underlying graph, which are more realistic models of human and social interaction.

One possible extension are regular graphs, as all nodes have the same degree. Differently from complete graphs, in regular graphs (and more general graphs too) the sequences representing the same tree have different probabilities and the order of contamination makes difference, which makes the analysis hard. In what follows we discuss two kinds of regular graphs.

The hypercube graph Q_n has 2^n vertices, $2^{n-1}n$ edges, and is a regular graph with n edges touching each vertex. For instance, the hypercube graph Q_3 is the graph formed by the 8 vertices and 12 edges of a three-dimensional cube.

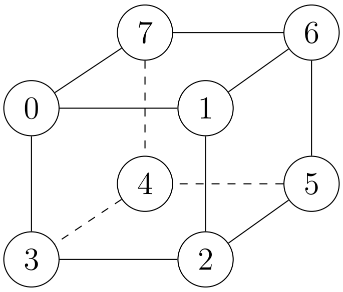


Figure 7.1: Hypercube Q_3

Now, suppose a line spanning tree is the result of the epidemic on the hypercube Q_3. The sequences b_1 and b_2, shown below, that represent epidemic processes occurred in the graph of Figure 7.1, both rooted in zero, give rise the same line tree but have different probabilities as we will show next.

Edge Sequence		
b_1	=	$\{(0,1);(1,2);(2,3);(3,4);(4,5);(5,6);(6,7)\}$
b_2	=	$\{(0,1);(1,6);(6,7);(7,4);(4,3);(3,2);(2,5)\}$

$$P(b_1) = \frac{1}{3 \cdot 4 \cdot 5 \cdot 5 \cdot 5 \cdot 4 \cdot 3} = 5.555 * 10{-}5$$

$$P(b_2) = \frac{1}{3 \cdot 4 \cdot 5 \cdot 4 \cdot 5 \cdot 4 \cdot 3} = 6.944 * 10{-}5$$

This occurs because, the order of contamination changes the size of the edge cut. More specifically, the number of edges added to the edge cut at each new contamination depends on the number of neighbours already infected of the new infected node. Thus, how the edge cut changes at a given infection does not just depend on which node is infected but also on whom has been infected until that time.

In regular graphs, different from complete graphs we need more information to know the size of the edge cut, because one node can increase the size of the edge cut in different ways depending on the number of neighbors already infected. So, when the underlying graph is a regular graph, the order in which vertices are added to the tree, that is, the moment of contamination of an individual matters.

A lattice graph, or grid graph, is a graph whose drawing, embedded in some Euclidean space R^n, forms a regular tiling, with wrap around. Figure 7.2 depicts a lattice with 9 nodes and 18 edges.

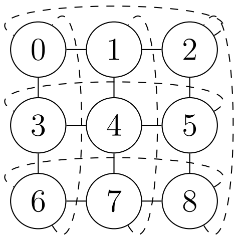


Figure 7.2: 2D lattice with 9 nodes.

As in the previous example, let us consider the case in which the spanning tree representing the epidemic process is a line tree. Below we have two different edge sequences rooted in zero representing a line tree; as it turns out, these sequences have different probabilities.

Edge Sequence		
b_1	$=$	$\{(0,1);(1,2);(2,5);(5,4);(4,3);(3,6);(6,7);(7,8)\}$
b_2	$=$	$\{(0,6);(6,8);(8,2);(2,1);(1,7);(7,4);(4,3);(3,5)\}$

$$P(b_1) = \frac{1}{4 \cdot 6 \cdot 6 \cdot 8 \cdot 6 \cdot 6 \cdot 5 \cdot 4} = 1.205 * 10{-}6$$

$$P(b_2) = \frac{1}{4 \cdot 6 \cdot 8 \cdot 8 \cdot 8 \cdot 6 \cdot 4 \cdot 4} = 8.477 * 10{-}7$$

As expected, the order of addition of the vertices in the infection sequence of the epidemic process matters, making it clear that the approach used for complete graphs cannot directly be applied to arbitrasry regular graphs.

Bibliography

[1] ANTULOV-FANTULIN, N., L. A. T. -H. . M., 2015, “Identification of Patient Zero in Static and Temporal Networks: Robustness and Limitations”, *Physical Review Letters*, v. 114 (6).

[2] BARABÁSI, A.-L., 2017, *Network Science*. 1 ed. United Kingdon, Cambridge University.

[3] BARRY, J. M., 2004, “The site of origin of the 1918 influenza pandemic and its public health implications”, *Journal of Translational Medicine*, v. 2 (1).

[4] BARTHÉLEMY, M., B. A. P.-S. R., VESPIGNANI, A., 2004, “Velocity and hierarchical spread of epidemic outbreaks in scale-free networks”, *Physical Review Letters*, v. 92 (4).

[5] BERNOULLI, D., 1766, “Essai d’une nouvelle analyse de la mortalite causee par la petite verole”, *Histoire de l’académie royale des sciences*, pp. 1–45.

[6] BUBECK, S., D. L. L.-G., 2017, “Finding Adam in random growing trees”, *Wiley Periodicals, Random Structures and Algorithms*, v. 50 (3), pp. 158–172.

[7] BUBECK, S., D. L. L.-G., 2019, “Finding the seed of uniform attachment trees”, *Electronic Journal of Probability*, p. 15.

[8] DALEY, D. J., KENDALL, D. G., 1964, “Epidemics and Rumors”, *Nature Publishing Group*, v. 204 (12), pp. 1118.

[9] DONG W., ZHANG W., T.-C. W., 2013, “Rooting out the Rumor Culprit from Suspects”, *IEEE International Symposium on Information Theory*, (10), pp. 2671–2675.

[10] HARRIS, B., WILSON, A. G., 1978, “Equilibrium Values and Dynamics of Attractiveness Terms in Production-Constrained Spatial-Interaction Models”, *Environment and Planning A: Economy and Space*, v. 10 (4), pp. 371–388.

[11] KEPHART, J. O., WHITE, S. R., 1991, “Directed-Graph Epidemiological Models of Computer Viruses”, *IEEE Computer Society Symposium on Research in Security and Privacy*, v. 1 (5), pp. 343.

[12] MELNYK, V. V., STYOPOCHKINA, I. V., 2019, “Malicious Information Source Detection in Social Networks”, *Social engineering and methods of counteracting destructive effects on consciousness in cyberspace.*

[13] NEWMAN, M. E. J., 2012, *Networks: An Introduction.* New York, Oxford University Press.

[14] PASTOR-SATORRAS, R., VESPIGNANI, A., 2001, “Epidemic Spreading in Scale-Free Networks”, *Physical Review Letters*, v. 86 (4), pp. 3200–3203.

[15] SHAH, D., ZAMAN, T., 2010, “Detecting sources of computer viruses in networks: Theory and experiment”, *Proceedings ACM Sigmetrics*, v. 15 (6), pp. 203–214.

[16] SHAH, D., ZAMAN, T., 2012, “Rumor centrality: A universal source detector”, *Sigmetrics Performance Evaluation Review - SIGMETRICS*, v. 40 (6), pp. 199–210.

[17] SHAH, D., ZAMAN, T., 2016, “Finding Rumor Sources on Random Trees”, *Institute for Operations Research and the Management Sciences (INFORMS)*, v. 64 (2).

[18] XU, XIN-JIAN; PENG, H.-O. W. X.-M., WANG, Y.-H., 2006, “Epidemic spreading with time delay in complex networks”, *Physica A: Statistical Mechanics and its Applications*, v. 367 (7), pp. 525–530.

[19] YU, PEI-DUO; TAN, C. W., FU, H.-L., 2017, “Rumor Source Detection in Finite Graphs with Boundary Effects by Message-passing Algorithms”, *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2017)*, (7), pp. 86–90.

[20] ZHENG, L., TAN, C. W., 2015, “A Probabilistic Characterization of the Rumor Graph Boundary in Rumor Source Detection”, *IEEE International Conference on Digital Signal Processing (DSP)*, (9), pp. 765–769.