Statistical inference in complex networks

Bianca Madoka Shimizu Oe

Bianca Madoka Shimizu Oe

Statistical inference in complex networks

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
December 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Oe, Bianca Madoka Shimizu
O28s Inferência estatística em redes complexas /
Bianca Madoka Shimizu Oe; orientador Francisco
Aparecido Rodrigues. - São Carlos - SP, 2016.
79 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática Computacional)
- Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2016.

1. Complex networks. 2. Spreading processes.
3. Regression analysis. 4. Sampling. I. Rodrigues,
Francisco Aparecido, orient. II. Título.

Bianca Madoka Shimizu Oe

Inferência estatística em redes complexas

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestra em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
Dezembro de 2016

ACKNOWLEDGEMENTS

Firstly, I would like to thank my advisor Francisco Aparecido Rodrigues. Our quick meetings and brainstorms were fundamental for the development of this work. I would also like to thank prof. Yamir Moreno, who warmly received and oriented me during my short stay in BIFI, always giving me good feedbacks. Many thanks to my lab mates in BIFI for making me feel welcome and for the interesting discussions and cañas. Thank you also to Guilherme Arruda, who helped me in a lot of ways throughout this research and to Tomás Fonseca, who helped me to generate many of the images in this thesis. Finally, a very special thanks to my family, notably to Choucha and Catinea, for the comprehension, patience, love and support during this period.

This research was supported by grants #2014/12301-8 and #2015/23587-2, São Paulo Research Foundation (FAPESP) and was partially developed with computational resources of the Center for Mathematical Sciences Applied to Industry (CeMEAI) financed by the São Paulo Research Foundation (FAPESP).

RESUMO

OE, B. M. S. **Inferência estatística em redes complexas**. 2016. 79 p. Master dissertation (Master student Program in Computer Science and Computational Mathematics) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2016.

Vários fenômenos naturais e artificiais compostos de partes interconectadas vem sendo estudados pela teoria de redes complexas. Tal representação permite o estudo de processos dinâmicos que ocorrem em redes complexas, tais como propagação de epidemias e rumores. A evolução destes processos é influenciada pela organização das conexões da rede. O tamanho das redes do mundo real torna a análise da rede inteira computacionalmente proibitiva. Portanto, torna-se necessário representá-la com medidas topológicas ou amostrá-la para reduzir seu tamanho. Além disso, muitas redes são amostras de redes maiores cuja estrutura é difícil de ser capturada e deve ser inferida de amostras. Neste trabalho, ambos os problemas são estudados: a influência da estrutura da rede em processos de propagação e os efeitos da amostragem na estrutura da rede. Os resultados obtidos sugerem que é possível predizer o tamanho da epidemia ou do rumor com base em um modelo de regressão beta com dispersão variável, usando medidas topológicas como regressores. A medida mais influente em ambas as dinâmicas é a informação de busca média, que quantifica a facilidade com que se navega em uma rede. Também é mostrado que a estrutura de uma rede amostrada difere da original e que o tipo de mudança depende do método de amostragem utilizado. Por fim, quatro métodos de amostragem foram aplicados para estudar o comportamento do limiar epidêmico de uma rede quando amostrada com diferentes taxas de amostragem. Os resultados sugerem que a amostragem por busca em largura é a mais adequada para estimar o limiar epidêmico entre os métodos comparados.

Palavras-chave: Redes complexas, Processos de propagação, Análise de regressão, Amostragem.

ABSTRACT

OE, B. M. S. **Statistical inference in complex networks**. 2016. 79 p. Master dissertation (Master student Program in Computer Science and Computational Mathematics) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2016.

The complex network theory has been extensively used to understand various natural and artificial phenomena made of interconnected parts. This representation enables the study of dynamical processes running on complex systems, such as epidemics and rumor spreading. The evolution of these dynamical processes is influenced by the organization of the network. The size of some real world networks makes it prohibitive to analyse the whole network computationally. Thus it is necessary to represent it by a set of topological measures or to reduce its size by means of sampling. In addition, most networks are samples of a larger networks whose structure may not be captured and thus, need to be inferred from samples. In this work, we study both problems: the influence of the structure of the network in spreading processes and the effects of sampling in the structure of the network. Our results suggest that it is possible to predict the final fraction of infected individuals and the final fraction of individuals that came across a rumor by modeling them with a beta regression model and using topological measures as regressors. The most influential measure in both cases is the average search information, that quantifies the ease or difficulty to navigate through a network. We have also shown that the structure of a sampled network differs from the original network and that the type of change depends on the sampling method. Finally, we apply four sampling methods to study the behaviour of the epidemic threshold of a network when sampled with different sampling rates and found out that the breadth-first search sampling is most appropriate method to estimate the epidemic threshold among the studied ones.

Keywords: Complex networks, Spreading processes, Regression analysis, Sampling.

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BA	Barabási-Albert
BFSS	Breadth-first search sampling
ER	Erdös-Rényi
ESS	Incident subgraph sampling
ISR	Ignorant-Spreader-Stifler
NLBA	Nonlinear Barabási-Albert
RWS	Random walk sampling
SIR	Susceptible-Infected-Removed
SpatialSF	Spatial to scale-free
VSS	Induced subgraph sampling
WS	Watts-Strogatz

LIST OF SYMBOLS

$\mathcal{G}$ — Graph

$\mathcal{V}$ — Set of vertices

$\mathcal{E}$ — Set of edges

A — Adjacency matrix

$|\mathcal{S}|$ — Cardinality of set $\mathcal{S}$.

M_{ij} — Element in row i and column j of matrix M

k_u — Degree of vertex u

$P(k)$ — Degree distribution

H — Shannon entropy of the degree distribution

r — Degree assortativity

$\langle cc \rangle$ — Average of the clustering coefficient

C — Transitivity

$V(B)$ — Variance of the betweenness centrality

$\langle kc \rangle$ — Average coreness

d_{uv} — Distance between vertices u and v.

E — Efficiency of a network

S — Average search information

CONTENTS

CHAPTER

1

INTRODUCTION

1.1 Overview

Various systems that at first seem unrelated, such as the society, the Internet, protein chains and food webs, have something in common: they are all examples of complex systems. A complex system is characterized by having a large number of agents that interact with each other and with the environment and display organization without any external influence, making complex behaviours to emerge from simple rules (MITCHELL, 2009; AMARAL; OTTINO, 2004). They are often described by the rule: “The whole is more than the sum of its parts” (AMARAL; OTTINO, 2004).

Complex systems are naturally represented as networks, that are formed by a set of nodes (or vertices) connected by edges (or links) (AMARAL; OTTINO, 2004). The birth of network theory is universally attributed to Euler (AMARAL; OTTINO, 2004), when he provided the solution to the Königsberg bridge puzzle (EULER, 1741). The Königsberg bridge puzzle inquires whether it is possible to plan a walk through the town of Königsberg in such a way that every bridge will be crossed exactly once. In his solution, Euler represented the problem as a graph, with regions as nodes and bridges as edges. This approach has been used plentifully since then. For example, in the early 20th century, the graph representation was used to analyze social relationships and the international commerce (ALBERT; BARABÁSI, 2002; NEWMAN, 2010).

However, network theory as known today, was introduced in the end of the 1990s, when the topology of the Internet and the World Wide Web were mapped (BARABÁSI; ALBERT, 1999; FALOUTSOS; FALOUTSOS; FALOUTSOS, 1999). It was observed that the structure of these networks is highly irregular, with a great variability in the number of links of the nodes and modular organization (NEWMAN, 2010). The distribution of the number of links follows a power law, meaning that the majority of the nodes have few connections whereas a small fraction of nodes are densely connected (FALOUTSOS; FALOUTSOS; FALOUTSOS, 1999).

The advances in technology have enabled us to collect large amounts of data related to complex systems. For instance, online social networks provide data on social interactions among individuals and mobile phones provide data on communication networks. At the same time, new methods to quantify the interaction between proteins and molecules, have enabled the mapping of fundamental biological iterations (BARABÁSI, 2007).

With the network data, we can simulate dynamical systems, such as epidemic spreading and synchronization (NEWMAN, 2010), on the network structure. In this way, we can study how the network organization influence the emergence of endemic state or the collective behavior of coupled oscillators. The quantification of the network role for dynamical system evolution is important not only to understand the behavior of the process, but also to develop methods for predicting and control of the dynamics. For instance, if we can identify the main propagators in epidemic spreading, then we can propose very effective methods for vaccination (WANG *et al.*, 2016).

However, the quantification of this influence is not trivial (BOCCALETTI *et al.*, 2006), since the network organization generally presents very complex pattern of connections, such as degree heterogeneity and modular organization. Several aspects of the network organization affect dynamical processes and it is impossible to measure all influences. Thus, this study has uncertainty associated and it is natural to consider statistical concepts to determine which network properties play fundamental role on the dynamical process. Statistical tools, such as regression analysis, can be employed to quantify the influence of the topology of the network in dynamical processes, as performed in (ARRUDA *et al.*, 2013), where Bayesian regression analysis was considered to predict the degree of synchronization of Kuramoto oscillators.

Statistical methods are also important to make inference about the network structure. The large size of some networks, which may be formed by thousands or even millions of agents (BARABÁSI, 2007). For instance, more than 250TB of storage would be necessary to store only the topology of the Facebook online social network (GJOKA *et al.*, 2010; GJOKA *et al.*, 2011). Sometimes it is prohibitive to study the whole network, either because of computational limitations, as in the case of Border Gate Protocol simulations, that are restricted to a few thousand nodes (DIMITROPOULOS; RILEY, 2003), or because they exist in a decentralized form and their global structure is not fully visible to the public (MAIYA, 2011) as in case of online social networks (MISLOVE *et al.*, 2007). Thus, instead of considering the whole network, a smaller treatable subset of nodes and links of the original network can be used (MAIYA, 2011).

Since several methods are used to sample networks, such as protein interaction or social networks, it is fundamental to determine which method produce graphs whose structure are similar to the original network. In fact, the sampled network may have different topological properties from the original network. For instance, in (STUMPF; WIUF; MAY, 2005) it was shown that the distribution of the number of connections of a random uniform sample of a scale-free network is not strictly scale-free.

In a similar way, sampling can modify network properties related to dynamical processes or the outcome of the process itself. For instance the spectral properties. In this case, dynamical processes on networks are affected since they are related to critical parameters, such as the epidemic threshold (PASTOR-SATORRAS *et al.*, 2015) for disease spreading and the critical coupling to reach the synchronous state (BOCCALETTI *et al.*, 2006). Therefore, conclusions drawn based on a sample of a network should not be generalized to the original network.

This work focuses on studying structural and functional properties of complex networks with the aid of statistical tools. Structural properties are purely topological properties of a network, while functional properties are related to dynamical processes that occur over the network (MAIYA, 2011). For the latter, we concentrate on epidemic and rumor spreading.

The following sections present the working our hypotheses and goals, as well as the organization of the thesis.

1.2 Working hypotheses

The working hypotheses of this work are:

(i) The structure of the network influences the collective behaviour in spreading processes and such influence can be quantified.

   a) The structure of the network can be well represented by its topological measures;

   b) It is possible to quantify this influence by means of regression analysis.

(ii) Sampled networks have properties and characteristics that differ from the original network.

1.3 Objectives

This work has two main goals, related to the aforementioned hypotheses:

(i) Study the influence of the organization of links in spreading processes, specifically epidemic and rumor spreading.

   a) Determine which topological property most influences the final fraction of infected individuals in epidemic and rumor spreading processes by using regression analysis;

   b) Compare the regression analysis to determine the similarities between these processes in terms of network structure.

(ii) Study the structural and functional effects of sampling on complex networks. Here, we focus on sampling a subnetwork, rather than a set of vertices.

   a) Compare structural properties of networks with their samples;

b) Study the behaviour of the epidemic threshold and the infection curve on sampled networks.

c) Approximate the epidemic threshold and the infection curve of the original network based on sampled graphs.

1.4 Organization of the thesis

This thesis is organized as follows:

- Chapter 2 presents theoretical background on network theory, regression analysis and network sampling, as well as the related work;
- Chapter 3 presents the methods employed throughout this work;
- Chapter 4 presents our results;
- Chapter 5 presents our conclusions.

CHAPTER

2

BACKGROUND

This chapter provides the necessary background and notation for understanding the rest of the thesis. It is organized as follows: Section 2.1 introduces the basic concepts of network theory and some methods for generating synthetic networks. Section 2.2 describes the spreading processes that are the focus of this work. Section 2.3 introduces regression analysis and Section 2.4 describes network sampling methods. Finally, Section 2.5 presents some related works.

2.1 Network theory

2.1.1 Basic definitions

A network can be modeled by a graph, by representing the elements as vertices and their interactions as edges. A graph is an ordered pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges that connect pairs of vertices. Here, we are interested in simple graphs, so an edge is an unordered pair $e = \{u, v\}$ that connects, or is incident in, two distinct vertices $u, v \in \mathcal{V}$. Vertices connected by an edge are called adjacent or neighbors.

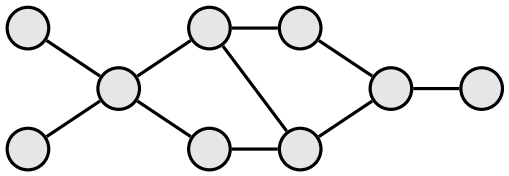
A graph is called simple if it is unweighted, undirected and has no multiple edges or self loops. In other words, the edges of the graph do not have weights and are bidirectional, there are no two edges connecting the same pair of vertices, and no edge connects a node to itself (GIBBONS, 1985). Simple and non-simple graphs are illustrated in Figure 1.

A simple graph can be represented by a boolean symmetric square matrix $A_{|\mathcal{V}|\times|\mathcal{V}|}$, called adjacency matrix. An element A_{uv} is one if there is an edge between vertices u and v and zero otherwise. Figure 2 illustrates a graph and its adjacency matrix.

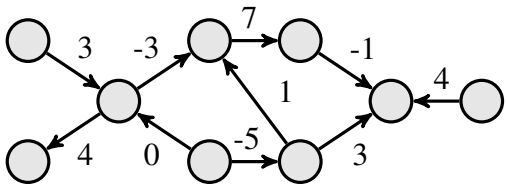
A walk is a sequence of edges connecting two vertices u and v $(\{u, w_1\}, \{w_1, w_2\}, \ldots, \{w_{l-1}, v\})$. The length of a walk is the number l of edges it contains. A walk is called a path if it passes

Figure 1 – Simple and non-simple graphs.

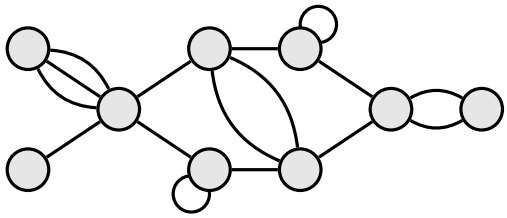
(a) Simple graph.



(b) Directed weighted graph.

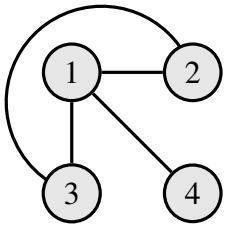


(c) Multigraph.



Source: Elaborated by the author.

Figure 2 – A simple graph and its adjacency matrix.



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

Source: Elaborated by the author.

through a vertex at most once and a cycle if it starts and ends in the same vertex. Two vertices are said connected if there is a path between them and a graph is connected if every pair of vertices is connected. An acyclic connected graph is called tree and a graph with edges between every pair of vertices is called clique. A vertex is called an articulation point if by removing it from the graph, the graph becomes disconnected.

A graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ is a subgraph of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if it is a graph and its set of vertices and edges are subsets of the vertices and edges of $\mathcal{G}$, that is $\mathcal{V}' \subset \mathcal{V}$ and $\mathcal{E}' \subset \mathcal{E}$. $\mathcal{G}$ is called supergraph of $\mathcal{G}'$. A connected component is a maximal connected subgraph, that is, every pair of vertices in $\mathcal{G}'$ is connected by a path and there are no additional vertices connected to them in $\mathcal{G}$.

2.1.2 Network measures

The degree k_u of a vertex u is the number of edges that are incident in it. The average degree of a network $\langle k \rangle$ is the average of k_u for all vertices of the network and indicates how dense it is. The degree distribution $P(k)$ is the probability that a randomly and uniformly drawn vertex has degree k.

From the degree distribution, it is possible to measure the heterogeneity of a network, which is closely related to its robustness to random failures (WANG *et al.*, 2006).

The second moment of the degree distribution $\langle k^2 \rangle$ indicates the variability in the degree distribution and is calculated as

$$\langle k^2 \rangle = \sum_k k^2 P(k). \tag{2.1}$$

For networks with the same average degree, the one with a larger variability in the distribution of edges will have a larger value of $\langle k^2 \rangle$. This measure also relates with other topological features, like the existence of a giant component (MOLLOY; REED, 1995).

Another way to measure the heterogeneity of a network is the Shannon entropy of the degree distribution S , which represents the diversity of the nodes in terms of the number of connections and is defined as

$$H = -\sum_k P(k) \log P(k). \tag{2.2}$$

The entropy is a non-negative value that is minimal, that is $H = 0$, when all the vertices have the same degree. Its maximum value is obtained when the degree distribution is uniform (WANG *et al.*, 2006).

Real world networks frequently exhibit correlation between the degrees of connected vertices. Social networks often show positive correlation, while technological and biological networks show negative correlation (NEWMAN, 2002).

The degree assortativity measures the correlation between the degree of connected vertices. A network is assortative if high-degree vertices are connected to other high-degree vertices and low-degree vertices are connected to other vertices with low-degree. When high-degree vertices are connected to low-degree vertices, the network is said disassortative. Finally, if there is no correlation between the degree of connected nodes, the network is said uncorrelated (NEWMAN, 2002).

One way to quantify the assortativity of a network is by calculating the Pearson correlation coefficient r of the degrees of connected vertices (NEWMAN, 2002). In terms of the adjacency matrix, the degree assortativity can be calculated as

$$r = \frac{\frac{1}{|\mathcal{E}|}\sum_{u<v} k_u k_v A_{uv} - \left[\frac{1}{|\mathcal{E}|}\sum_{u<v}\frac{1}{2}(k_u + k_v)A_{uv}\right]^2}{\frac{1}{|\mathcal{E}|}\sum_{u<v}\frac{1}{2}(k_u^2 + k_v^2)A_{uv} - \left[\frac{1}{|\mathcal{E}|}\sum_{u<v}\frac{1}{2}(k_u + k_v)A_{uv}\right]^2}. \tag{2.3}$$

If $r \approx 0$, the network is called uncorrelated. If $r > 0$ the network is assortative and if $r < 0$, the network is disassortative.

Many real world networks exhibit a larger number of cycles of size 3 – or triangles – than it is expected in random graphs (COSTA *et al.*, 2007). For example, in friendship networks, it shows the tendency of people having friends who are also friends with each other.

The clustering coefficient (WATTS; STROGATZ, 1998) is a local measure of the presence of triangles. The clustering coefficient of a node measures the probability that two of its neighbors drawn randomly and uniformly are connected and is equal to:

$$cc_u = \frac{2e_u}{k_u(k_u-1)}, \tag{2.4}$$

where e_u is the number of connections between neighbors of u.

A more global way to measure the presence of such cycles is the transitivity C (AMARAL; OTTINO, 2004), which is calculated as follows:

$$C = \frac{3N_\triangle}{N_3}, \tag{2.5}$$

where $N_\triangle$ is the number of triangles and N_3 is the number of connected triples. A connected triple is a sequence of three vertices connected by two edges, or a path of length two. In terms of the adjacency matrix,

$$N_\triangle = \sum_{u<v<w} A_{uv}A_{vw}A_{wu} \tag{2.6}$$

and

$$N_3 = \sum_{u<v<w} A_{uv}A_{vw} + A_{uw}A_{wv} + A_{vu}A_{uw}. \tag{2.7}$$

The transitivity assumes values in the range $[0,1]$. When $C = 0$, there are no triangles and when $C = 1$, the graph is composed of a set of cliques.

The extent to which a vertex is involved when passing information in a network, or its load, can be quantified by the betweenness centrality (FREEMAN, 1977). The betweenness of vertex u, B_u, is defined as the fraction of shortest paths of the network that pass through u, that is

$$B_u = \sum_{v<w} \frac{\sigma(v,u,w)}{\sigma(v,w)}, \tag{2.8}$$

where $\sigma(v,u,w)$ is the number of shortest paths between v and w that go through u and $\sigma(v,w)$ is the total number of shortest paths between v and w.

Vertices that are crossed by many shortest paths have a high value of the betweenness centrality and control the communication among other nodes in the network (FREEMAN, 1977) whereas low values of betweenness indicate peripheral vertices.

Another way to measure the importance of vertices is by using the coreness measure (SEIDMAN, 1983; BATAGELJ; ZAVERSNIK, 2003). The k-core of a graph is the maximal subgraph in which every node has degree at least k. The coreness of node u is kc_u, if it is part of a kc_u-core, but not of a (kc_u+1)-core. The overall organization of the network can be characterized by the average coreness $\langle kc \rangle$ taken over all of the vertices.

In transportation and communication networks, the distance between vertices is an important concept (BARTHÉLEMY, 2003). More specifically, the shortest path which connects

a pair of vertices is relevant for routing in computer networks and in navigation in road networks. In simple graphs, the distance d_{uv} between vertices u and v is the length of the shortest path connecting u and v.

The efficiency E of a network measures how efficiently it exchanges information between vertices. The efficiency in the communication between two vertices is calculated as the inverse of the distance between them. A global measure can be obtained by taking the average of all pairs of vertices (LATORA; MARCHIORI, 2001), that is

$$E = \frac{2}{N(N-1)} \sum_{u<v} \frac{1}{d_{uv}}. \tag{2.9}$$

The average search information S quantifies the difficulty to navigate or to search for information in a network (ROSVALL *et al.*, 2005), represented by the probability of following a shortest path to a specific vertex while doing a random walk, i.e., navigating through edges at random.

Formally, let $p(u,v)$ be a shortest path from u to v. The probability of following this path in a random walk is

$$P[p(u,v)] = \frac{1}{k_u} \prod_{\substack{w \in p(u,v) \\ w \notin \{u,v\}}} \frac{1}{k_w - 1}. \tag{2.10}$$

In this equation, the number of ways to leave vertex w is $k_w - 1$ because the edge used to get to w is not counted. Since u is the start vertex, it is possible to leave u through all of its edges, leading to k_u possibilities.

The search information $S(u,v)$, corresponding to the total information needed to identify one of the shortest paths connecting u and v, is given by

$$S(u,v) = -\log \sum_{\{p(u,v)\}} P[p(u,v)], \tag{2.11}$$

where $\{p(u,v)\}$ is the set of all shortest paths between u and v. The average search information is the average of the search information taken over all pairs of vertices (ROSVALL *et al.*, 2005)

$$S = \frac{1}{N^2} \sum_{u,v} S(u,v). \tag{2.12}$$

Other global measures of networks are characterized by the eigenvalues of the adjacency matrix (FARKAS *et al.*, 2001). For example, the shortest paths, number of cycles and connectivity properties of the network are related to the eigenvalues and eigenvectors of the network (COSTA *et al.*, 2007).

The largest eigenvalue of the adjacency matrix λ_1, also called principal eigenvalue or spectral radius, is related to the average degree of the graph and plays an important role in dynamical processes such as epidemic spreading (WANG *et al.*, 2003) and synchronization of coupled phase oscillators (RESTREPO; OTT; HUNT, 2005).

On simple graphs, it is unique, real and positive and is bounded from below by $\max(\langle k \rangle, \sqrt{k_{\max}})$ and from above by $k_{\max}$ (RESTREPO; OTT; HUNT, 2007; ZUMSTEIN, 2005). Besides, the λ_1 of a graph never increases when a vertex is removed (RESTREPO; OTT; HUNT, 2007). Therefore, subgraphs formed by deleting a set of vertices always have a spectral radius smaller than the original graph.

2.1.3 Network models

2.1.3.1 Random graphs (ER)

The Erdös and Rényi (ER) (ERDÖS; RÉNYI, 1959) model generates random graphs by connecting every pair of vertices with a constant probability p. The degree distribution is binomial and follows a Poisson distribution when the number of vertices $|\mathcal{V}|$ is large and p is small. Therefore, the probability that a vertex has k edges is given by:

$$P(k) = \frac{(|\mathcal{V}|p)^k e^{-|\mathcal{V}|p}}{k!}, \tag{2.13}$$

and $\langle k \rangle = \frac{(|\mathcal{V}|-1)p}{2}$.

The ER model generates graphs with a homogeneous degree distribution, low average clustering coefficient and small distances between vertices (COSTA *et al.*, 2007).

2.1.3.2 Small-world (WS)

The Watts-Strogatz small world model (WS) (WATTS; STROGATZ, 1998) generates graphs that exhibit the small world property, i.e., the average distance between vertices scales logarithmically with the number of vertices (WATTS; STROGATZ, 1998). They also have a larger clustering coefficient, characteristic present in real world networks.

The WS model starts with a regular lattice, where each vertex is connected to its $\frac{\langle k \rangle}{2}$ neighbors in each direction, totalizing $\langle k \rangle$ connections. Then, each edge is rewired with a constant probability p. When $p = 0$, the graph is a regular lattice, with large distances and high average clustering coefficient. As p increases, the distances are shortened and the average clustering coefficient decreases. Finally, when $p \to 1$, the graph becomes a random graph (COSTA *et al.*, 2007).

2.1.3.3 Barabási-Albert scale-free (BA)

The Barabási and Albert scale-free model (BA) (BARABÁSI; ALBERT, 1999) addresses the heterogeneous degree distribution observed in many real world networks, that exhibit some highly connected vertices and many vertices with few connections, with the absence of a characteristic degree (COSTA *et al.*, 2007). More specifically, the degree distribution follows a power law function of the form

$$P(k) \propto k^{-\gamma}, \tag{2.14}$$

where γ is called the degree exponent, that is usually between 2 and 3 in real world networks. Such networks are called scale-free (SF) (COSTA *et al.*, 2007).

The BA model generates scale-free networks by incorporating two concepts: growth and preferential attachment (BARABÁSI; ALBERT, 1999). The vertices are incrementally added to the graph, connecting to the existing vertices with a probability which depends on their degree. Formally, the graph starts with a set of m_0 connected vertices. Then, in each step, a vertex u is added to the graph and is connected to m existing vertices. The probability that u is connected to v is given by

$$P(\{u,v\}) = \frac{k_v}{\sum_i k_i}. \tag{2.15}$$

This process generates graphs with a power law degree distribution with $\gamma = 3$ (BARABÁSI; ALBERT, 1999) and a low clustering coefficient (BARTHÉLEMY, 2003).

In (ONODY; CASTRO, 2004), a variation of the BA model, called nonlinear Barabási-Albert (NLBA), was proposed. In the NLBA model, the network is generated by adding a new vertex in each step, as in the BA model, but the probability that the new vertex connects to an existing vertex v is proportional to ${k_v}^{\alpha}$, where α is a nonlinearity parameter.

2.1.3.4 Spatial random graphs (Waxman)

The Waxman model (WAXMAN, 1988) is a generalization of the ER model (ROUGHAN; TUKE; PARSONAGE, 2015) and introduces spatial information to the network by considering that longer links are more costly and, thus, less likely to be built. It considers that the vertices of the graph are embedded in space and that the probability of two vertices being connected depends on the euclidean distance between them.

Formally, vertices are randomly and uniformly distributed in a square and the probability that there is an edge between vertices u and v is

$$P(\{u,v\}) = \beta \exp \frac{-d_{uv}}{L\alpha}, \tag{2.16}$$

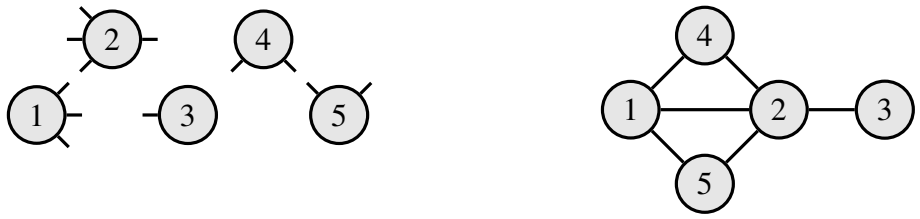
where d_{uv} is the euclidean distance between u and v, α and β are parameters in the range $(0,1]$ and L is the largest distance between any two points. The β parameter controls the density of edges, while α regulates the density of short edges relative to longer ones (WAXMAN, 1988).

2.1.3.5 Spatial to scale-free (SpatialSF)

The spatial to scale-free (SpatialSF) model (BARTHÉLEMY, 2003) incorporates not only the distance selection – short links are more likely to occur than long links, but also the preferential attachment, generating networks that lie between scale-free and spatial random graphs. It addresses a characteristic observed in many spatial real world networks, such as airlines: a long-range link usually connects to a hub (BARTHÉLEMY, 2003).

Figure 3 – Graph generated by the configuration model.

(a) Nodes with stubs. (b) Graph after connecting the stubs.



Source: Elaborated by the author.

The vertices are distributed uniformly in space and there is a set of m_0 initially connected nodes (BARTHÉLEMY, 2003). Like in the BA model, at each step, a new vertex is added to the network, connecting to m existing vertices. However, the probability of connecting the new vertex to an existing vertex depends not only on the degree of the existing vertex, but also on the euclidean distance between them.

Formally, the probability that a new node u will connect with an existing node v is

$$P(u,v) \propto \frac{Z(k_v)}{\Delta(d_{uv})}, \tag{2.17}$$

where Z and Δ are given functions (BARTHÉLEMY, 2003).

Depending on the functions chosen for Z and Δ, it is possible to have only preferential attachment, only distance selection or both (BARTHÉLEMY, 2003). For the last case, a possibility for a natural generalization of the Waxman model is $Z(k) = k+1$ and $\Delta(d) = \exp \frac{d}{\alpha}$ (BARTHÉLEMY, 2003), yielding the equation

$$P(u,v) \propto (k_v + 1) \exp \frac{-d_{uv}}{\alpha}. \tag{2.18}$$

2.1.3.6 Configuration model

The configuration model (NEWMAN; STROGATZ; WATTS, 2001) generates a random graph with a predefined degree distribution $P(k)$. Each vertex u has k_u stubs, where k_u is drawn from $P(k)$. Then, pairs of stubs are randomly and uniformly drawn and connected to form an edge until all stubs are used up. Figure 3 illustrates the process of generating a graph using the configuration model.

2.1.3.7 Assortative rewiring

A similar idea to that of creating small-world networks can be used to make a network more or less assortative (XULVI-BRUNET; SOKOLOV, 2004). To vary the assortativity, two edges are drawn randomly and their incident nodes are connected according to their degree: to make the network more assortative, the two nodes with largest degrees are connected, and to

make the network less assortative, the node with largest degree is connected to the node with smallest degree. This process changes the assortativity of the network, but maintains its degree distribution.

2.2 Spreading processes

Spreading processes consider the transmission of an information between pairs of subjects. Such information can be an infectious agent or some news, as in the case of rumors. As following, we describe the main models considered to study epidemic and rumor spreading.

2.2.1 Epidemic spreading

The spread of a disease can be regarded as the transition of individuals through compartments (or states) like susceptible, infected and recovered or removed (ANDERSON; MAY; ANDERSON, 1992). Here we are interested in the Susceptible-Infected-Removed (SIR) model (ANDERSON; MAY; ANDERSON, 1992), that simulates epidemic spreading in a network by changing the state of each individual according to the state of its neighbors (infection) or spontaneously (removal).

Infected individuals can spread the disease to their susceptible neighbors with probability δ, who may become infected. They can also recover from the disease with probability ν, becoming immune and playing no role on the process anymore. The SIR model imitates the behaviour of diseases as measles, a disease that is transmitted by close contact and has an infective period of approximately a week, after which the individual develops lifelong immunity (BJØRNSTAD; FINKENSTÄDT; GRENFELL, 2002).

During the spreading process, in a given time t, an individual may either be susceptible, infected or removed. The SIR model considers that the rate of infection and recovery is much faster than the lifespan of individuals, so births and natural deaths are unaccounted for, maintaining the size of the population constant throughout the process (MORENO; PASTOR-SATORRAS; VESPIGNANI, 2002). Thus, the densities of susceptible $s(t)$, infected $i(t)$ and removed $r(t)$ individuals are linked by the normalizing condition

$$s(t)+i(t)+r(t)=1. \tag{2.19}$$

For a homogeneous population, the SIR model (KERMACK; MCKENDRICK, 1927) is described by the following system of differential equations:

$$\begin{cases} \frac{ds(t)}{dt} = -\delta\langle k\rangle i(t)s(t), \\ \frac{di(t)}{dt} = -\nu i(t)+\delta\langle k\rangle i(t)s(t), \\ \frac{dr(t)}{dt} = \nu i(t), \end{cases} \tag{2.20}$$

with the initial conditions $s(0)=\frac{|\mathcal{V}|-1}{|\mathcal{V}|}$, $i(0)=\frac{1}{|\mathcal{V}|}$ and $r(0)=0$.

In this system, it assumed that the number of contacts per unit time $\langle k \rangle$ is constant for every individual, δ is the microscopic infection rate and ν is the recovery rate. The above equations state that the density of infected individuals increases at a rate proportional to the infection rate δ, the number of contacts $\langle k \rangle$ and the densities of infected and susceptible individuals, $i(t)$ and $s(t)$, while the increase in the density of removed individuals is proportional to the recovery rate ν and the density of infected individuals, $i(t)$, but is independent of the connectivity $\langle k \rangle$ (MORENO; PASTOR-SATORRAS; VESPIGNANI, 2002).

This model predicts the presence of epidemic threshold $\delta_c = \frac{1}{\langle k \rangle}$ (MURRAY, 2002), related to the basic reproductive number $R_0 \propto \delta \langle k \rangle$, that is the average number of secondary infections caused by an infected individual (MORENO; PASTOR-SATORRAS; VESPIGNANI, 2002). If the spreading rate is above δ_c, the disease infects a nonzero fraction of the population. If $\delta < \delta_c$, the number of infected individuals is infinitesimally small in thermodynamic limit, that is, when $|V| \to \infty$ (MARRO; DICKMAN, 2005).

The homogeneity assumption is inadequate to model real world networks, which present a heterogeneous topology, as the network of sexual partners, which is better described by a scale-free topology (LILJEROS *et al.*, 2003). In heterogeneous networks, the epidemic threshold decreases with an increasing standard deviation of the degree distribution (ANDERSON; MAY; ANDERSON, 1992). Scale-free networks with $2 < \gamma \leq 3$ are known to have a diverging $\langle k^2 \rangle$ on the thermodynamic limit (PASTOR-SATORRAS; VESPIGNANI, 2001; BOGUNÁ; PASTOR-SATORRAS; VESPIGNANI, 2003), meaning that the epidemic threshold vanishes for such networks and diseases become endemic even with a small infection rate (PASTOR-SATORRAS; VESPIGNANI, 2001; BOGUNÁ; PASTOR-SATORRAS; VESPIGNANI, 2003).

In general, under reasonable approximations, the epidemic threshold can be estimated by

$$\delta_c = \frac{1}{\lambda_1}, \tag{2.21}$$

where λ_1 is the largest eigenvalue of the adjacency matrix (WANG *et al.*, 2003).

2.2.2 Rumor spreading

The process of spreading a rumor resembles the one of a disease. The individuals are divided into three compartments: ignorants, which are individuals that do not know the rumor; spreaders, that are the ones who know the rumor and tell others; and stiflers, which are individuals who know the rumor but are not interested in it anymore. Hereafter this model will be addressed as the Ignorant-Spreader-Stifler (ISR) model.

In the classical Daley-Kendall model (DALEY; KENDALL, 1965), the rumor is transmitted by pairwise contacts between spreaders and other individuals. When spreaders contact ignorants, they get to know the rumor and turn into spreaders with probability δ. When the contacted individuals are spreaders or stiflers, the spreaders may lose interest in the rumor and

turn into stiflers with probability ν. In the Maki-Thompson variant (MAKI *et al.*, 1973), only the initial spreader may turn into a stifler.

There is a correspondence between the rumor spreading and the epidemic spreading states: ignorant to susceptible, spreader to infected and stiflers to removed. However, a spreader turns into a stifler only when contacting another spreader or stifler, while in the epidemic spreading, the recovery is always spontaneous. Also, stiflers are not removed from the rumor dynamics, as they still affect the process by causing spreaders to turn into stiflers.

In a homogeneous population, the model can be described in terms of the densities of ignorant $i(t)$, spreaders $s(t)$ and stiflers $r(t)$, linked by the normalizing condition

$$i(t) + s(t) + r(t) = 1. \tag{2.22}$$

The following system of equations describe the variation rate in the density of individuals in each compartment:

$$\begin{cases} \frac{di(t)}{dt} = -\delta\langle k\rangle s(t) i(t), \\ \frac{ds(t)}{dt} = -\nu\langle k\rangle s(t)[s(t) + r(t)] + \delta\langle k\rangle s(t) i(t), \\ \frac{dr(t)}{dt} = \nu\langle k\rangle s(t)[s(t) + r(t)], \end{cases} \tag{2.23}$$

with the initial conditions $i(0) = \frac{|\mathcal{V}|-1}{|\mathcal{V}|}$, $s(0) = \frac{1}{|\mathcal{V}|}$ and $r(0) = 0$.

In a similar way to the SIR model, the density of spreaders increases at a rate proportional to the density of spreaders $s(t)$ and ignorants $i(t)$, as well as the spreading rate δ and the number of contacts $\langle k\rangle$, which is constant for the whole population. However, in contrast to the SIR model, the increase in the density of stiflers depends not only on the loss of interest rate ν and on the density of spreaders $s(t)$, but also on the connectivity $\langle k\rangle$ and on the density of the individuals who once were aware of the rumor $[s(t) + r(t)]$.

An important consequence of the difference in the decay of the spreading process is that there is no "rumor threshold", opposed to the existence of the epidemic threshold (MORENO; NEKOVEE; PACHECO, 2004). In fact, there is not phase transition in the rumor spreading, while we have a second-order phase transition on epidemic spreading (PASTOR-SATORRAS *et al.*, 2015).

An interesting variation of the original model is the addition of the possibility of a spreader spontaneously forgetting the rumor and turning into a stifler. This change causes the emergence of a finite threshold, independent of ν, below which the rumor ceases to spread. And like in the SIR case, it vanishes for scale-free networks with γ between 2 and 3 (NEKOVEE *et al.*, 2007).

2.3 Regression analysis

2.3.1 Overview

Regression analysis is a statistical tool used to determine the relation between a dependent variable and one or more independent variables. Usually, the analysis is aimed at describing how the mean of a variable of interest varies with changing conditions (DRAPER; SMITH, 2014).

The variable of interest is called dependent variable or response, represented by Y, and the n explanatory variables are called independent variables, regressors or predictors, and are represented by an array $X = (X_1, X_2, \ldots, X_m)$.

There are several types of regression models. The most widely known is the linear regression, which assumes that there is a linear relation between X_t and the population mean $E(Y_t) = \mu_t$ of Y_t, i.e., μ_t can be written as

$$\mu_t = \beta_0 + \sum_{i=1}^{m} \beta_i X_{ti}. \tag{2.24}$$

β_0 is called the intercept and is the value of μ_t when $X_t = (0, \ldots, 0)$, and β_i is the increase or decrease rate in μ_t per unit change in X_{ti} when all the other values are held constant (DRAPER; SMITH, 2014). The subscript t indicates the observation index and the subscript i is the index of the explanatory variable.

It is assumed that the observations on X are measured without error, i.e., that they are a set of constants. The observation on Y_t is assumed to be a random observation from a distribution with mean μ_t and the deviation of Y_t from μ_t is accounted by a random error term ε_t (DRAPER; SMITH, 2014), yielding

$$Y_t = \beta_0 + \sum_{i=1}^{m} \beta_i X_{ti} + \varepsilon_t. \tag{2.25}$$

The errors ε_t are assumed to be normally and independently distributed with zero mean and a constant variance σ^2 (DRAPER; SMITH, 2014). This implies that the dependent variables Y_t are also normally and independently distributed with a constant variance σ^2.

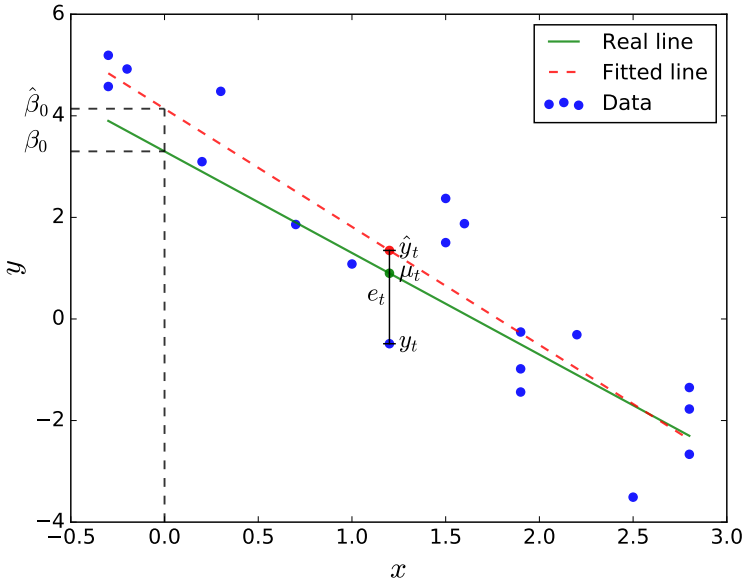
The coefficients are estimated from the data, according to some criteria. The least squares method uses the criterion that the solution minimizes the sum of the squared deviations of the observations of Y_t from their estimated mean, provided by the solution (DRAPER; SMITH, 2014). Formally, let $\hat{\beta}_i$ be the estimate of β_i and $\hat{Y}_t$ be the estimated mean of Y_t. $\hat{Y}_t$ is calculated by

$$\hat{Y}_t = \hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i X_{ti} \tag{2.26}$$

and the deviation of the t-th observation is calculated as

$$e_t = y_t - \hat{y}_t. \tag{2.27}$$

Figure 4 – Example of linear regression with one explanatory variable using the least squares method. β_1 is the slope of the real line and $\hat{\beta}_1$ is the slope of the fitted line.



Source: Elaborated by the author.

Therefore, $\hat{\beta}$ is chosen to minimize

$$\sum_t e_t^2 = \sum_t (y_t - \hat{y}_t)^2. \tag{2.28}$$

Figure 4 illustrates some of the presented concepts.

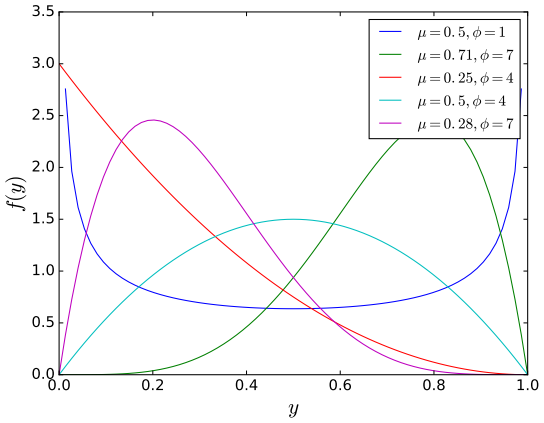
The coefficient of determination R^2 is the proportion of the variability in the data explained by the model and measures how well the regression line represents the data. It varies between zero and one, with zero meaning that the model explain none of the variability of the data around its mean, and one meaning that it explains all the variability of the data around its mean.

For a data set with observed values y and estimates $\hat{y}$, the square of the product moment correlation between y and $\hat{y}$ (DRAPER; SMITH, 2014) is calculated by

$$R^2 = \frac{\sum_t (\hat{y}_t - \bar{y}_t)^2}{\sum_t (y_t - \bar{y}_t)^2}. \tag{2.29}$$

The linear model makes some structural assumptions on the data: the random errors are unbiased; have constant variance; are uncorrelated; and are normally distributed. The departure from the underlying assumptions may cause problems, such as a biased estimation of $\hat{\beta}$ and over-estimation of σ^2 (SEBER; LEE, 2012). Thus, other regression models may be more appropriate depending on the application.

Figure 5 – Beta density for different parameters μ and ϕ.



Source: Elaborated by the author.

2.3.2 Beta regression

In many applications, the response is continuous and restricted to the interval $(0,1)$, as for rates and proportions. The linear regression model may yield fitted values outside of this interval, hence, it is not appropriate for these cases. A transformation in the response variable to assume values on the real line may solve this problem. However, the model parameters cannot be easily interpreted in terms of the original response (FERRARI; CRIBARI-NETO, 2004). Besides, measures of proportions are typically asymmetric, hence the normality assumption may be violated (FERRARI; CRIBARI-NETO, 2004).

The beta regression model (FERRARI; CRIBARI-NETO, 2004) assumes that the response is in the $(0,1)$ interval and is beta distributed. The beta density is given by:

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \tag{2.30}$$

in which $\Gamma(.)$ is the gamma function, $p > 0$ and $q > 0$ are parameters of the distribution and $y \in (0,1)$.

The distribution can be rewritten in terms of $\mu = \frac{p}{p+q}$ and $\phi = p+q$ (FERRARI; CRIBARI-NETO, 2004). The beta density function then becomes:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}. \tag{2.31}$$

The mean and the variance of the distribution according to the new parameters μ and ϕ are $E(y) = \mu$ and $V(y) = \frac{\mu(1-\mu)}{1+\phi}$, that is, μ is the mean of the response and ϕ can be seen as a precision parameter, as the variance of the response decreases with the increase of ϕ for a fixed μ (FERRARI; CRIBARI-NETO, 2004). Figure 5 illustrates the beta distribution for some values of μ and ϕ.

The model assumes that the observations on the dependent variable are a random sample $Y_1, \cdots, Y_n$ of independent beta distributed random variables (CRIBARI-NETO; ZEILEIS, 2009). The mean $E(Y_t) = \mu_t$ is, then, written as

$$g_1(\mu_t) = \sum_{i=1}^{m} X_{ti}\beta_i, \tag{2.32}$$

where $X_{t1}, \cdots, X_{tm}$ are observations of $m < n$ regressors that are considered fixed and known and $\beta_1, \cdots, \beta_m$ are the unknown regression parameters. Finally, $g_1(.)$ is a strictly monotonic and twice differentiable link function, that maps $(0,1)$ into $\mathbb{R}$ (FERRARI; CRIBARI-NETO, 2004).

The beta regression model assumes that the precision parameter ϕ is constant across observations. The variable dispersion beta regression model is a variant of the beta regression model that allows variation in the precision parameter (SIMAS; BARRETO-SOUZA; ROCHA, 2010). In this model, the precision parameter ϕ_t of observation t is written as

$$g_2(\phi_t) = \sum_{i=1}^{p} Z_{ti}\gamma_i, \tag{2.33}$$

where $Z_{t1}, \cdots, Z_{tp}$ are observations of $p < n$ regressors, $\gamma_1, \cdots, \gamma_p$ are the unknown regression parameters and $g_2(.)$ is a strictly monotonic and twice differentiable link function, that maps $(0,1)$ into $\mathbb{R}$ (SIMAS; BARRETO-SOUZA; ROCHA, 2010).

The regression parameters are estimated by the maximum likelihood estimation method (FERRARI; CRIBARI-NETO, 2004), that chooses the parameters that will maximize the probability of appearance of the observed data among all the parameter space.

The suitability of the model can be assessed by diagnostic plots of residuals and predictions, such as the half-normal plot of residuals with simulated envelope and the predicted versus observed values plot, as well as by the pseudo-R^2 coefficient (CRIBARI-NETO; ZEILEIS, 2009).

The half-normal plot of residuals with simulated envelope plots the absolute value of the ordered residuals against their normal quantile and a confidence interval created by simulating samples of the response variable. If the model suits the data, the points should fall inside the simulated envelope and if the residuals are normally distributed, the points should fall in a straight line (ATKINSON, 1981). The plot of predicted versus observed values shows the values predicted for each data point against their true value. Finally, the pseudo-R^2 is a measure of explained variation, that ranges between zero with one representing a perfect agreement between values predicted by the model and the observed values (FERRARI; CRIBARI-NETO, 2004).

2.4 Sampling in complex networks

In this section, N denotes the desired sample size, $G = (V, E)$ stands for the original graph, $G' = (V', E')$ is a graph sampled from G, in which V' is the set of sampled vertices and E'

is the set of sampled edges. The graph G will henceforth be called original graph and the graph G' will be called sampled graph or subgraph.

2.4.1 Induced subgraph sampling (VSS)

The induced subgraph sampling method (VSS, where V stands for vertex) consists in drawing vertices from the original graph in a random uniform way and constructing the subgraph that is formed by those vertices and all the edges in the original graph that connect them (KOLASCYK, 2013), called induced subgraph. Formally, V' is composed of N vertices of V drawn randomly uniformly without replacement and $E' = \{\{u,v\} \in E | (u,v) \in V' \times V'\}$. The algorithm of the method is shown in Algorithm 1 and an example of a sample can be seen in Figure 6.

This method is used by social network researchers to build contact networks. First a sample of individuals (vertices) is selected and then the individuals are interviewed to discover the links between them (edges) (KOLASCYK, 2013).

The probability of inclusion of a vertex follows directly from the definition of the VSS and is equal to:

$$\pi_u = \frac{N}{|\mathcal{V}|}. \tag{2.34}$$

The inclusion of an edge depends on the inclusion of both of its incident vertices. Hence, the probability of inclusion of an edge is equal to

$$\pi_{\{u,v\}} = \frac{N(N-1)}{|\mathcal{V}|(|\mathcal{V}|-1)}. \tag{2.35}$$

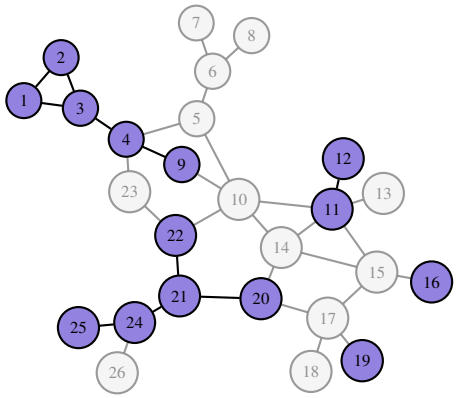
The VSS method is interesting when trying to estimate a global measure that is obtained by the sum or the average of locally measured variables, such as the average degree of the original network or the average age of individuals in a network. An unbiased estimate of the global measure is obtained by the sum or average of the values observed in the sampled nodes or edges.

However, the VSS method operates under the assumption of existence of a vertex list, i.e., that it is possible to draw vertices uniformly from the original graph, which is not always true. Also, the samples it produces may be disconnected and contain isolated vertices.

2.4.2 Incident subgraph sampling (ESS)

The incident subgraph sampling method (ESS, where E stands for edge) is similar to the induced subgraph sampling method. In this case, the sample is based on the edges of the original graph: instead of drawing vertices uniformly, the method draws edges with a uniform probability. The set E' is composed of N edges drawn randomly and uniformly from E and $V' = \{u | e = \{u,v\} \in E'\}$, that is, the V' is composed of all the vertices that are incident on an

Figure 6 – Example of a graph sampled by the VSS method.



Source: Elaborated by the author.

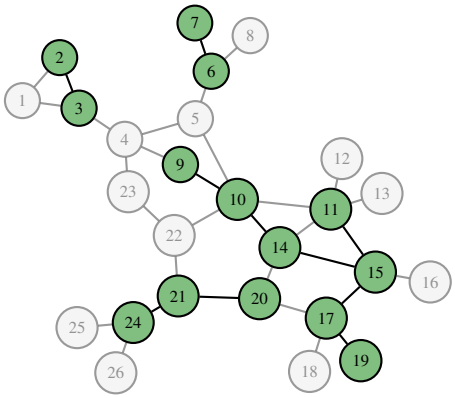
Algorithm 1 – Induced subgraph sampling.

procedure VSS(G,N)
input: Graph $G=(V,E)$; Sample size N.
output: Sampled graph G'.
    let V' be the set of sampled vertices
    let E' be the set of sampled edges
    let C be the set of vertices that are not in the sample
    initialize V' with the empty set
    initialize E' with the empty set
    initialize C with V
    while $|V'|$ is less than N **do**
        let u be the sampled vertex in this step
        draw u from C randomly and uniformly
        remove u from C
        add u to V'
    end while
    for all edge $e=\{u,v\}$ in E **do**
        if u is in V' and v is in V' **then**
            add e to E'
        end if
    end for
    return $G'=(V',E')$
end procedure

edge in E' (KOLASCYK, 2013). The ESS algorithm is shown in Algorithm 2 and an example of sample is shown in Figure 7.

It is straightforward to see that $|V'|$ is usually different from N, varying between approximately $\sqrt{2N}$, when G' is a clique and $2N$, when each edge is a connected component of G'.

Figure 7 – Example of a graph sampled by the ESS method.



Source: Elaborated by the author.

The probability of inclusion of an edge is uniform and equal to

$$\pi_{\{u,v\}} = \frac{N}{|\mathcal{E}|}, \tag{2.36}$$

and the probability of inclusion of a vertex is the complement of the probability that none of its incident edges are sampled (KOLASCYK, 2013) and given by:

$$\pi_u = \begin{cases} 1 - \frac{\binom{|\mathcal{E}|-k_u}{N}}{\binom{|\mathcal{E}|}{N}} & \text{if } N \leq |\mathcal{E}| - k_u, \\ 1 & \text{otherwise.} \end{cases} \tag{2.37}$$

The ESS is implicit when constructing the network of telephone calls. First the telephone calls are sampled and then the phone numbers of the involved parties are observed (KOLASCYK, 2013).

The ESS has the same drawbacks as the VSS: the sampled graph can be disconnected and it needs of a list of edges.

2.4.3 Breadth-first search sampling (BFSS)

The breadth-first search sampling method is a breadth-first search based sampling algorithm. The breadth-first search is a graph traversal algorithm that starts from a seed and explores the vertices according to their distance from the seed. More specifically, at each step, it explores an unexplored neighbor of the earliest explored vertex (CORMEN *et al.*, 2001).

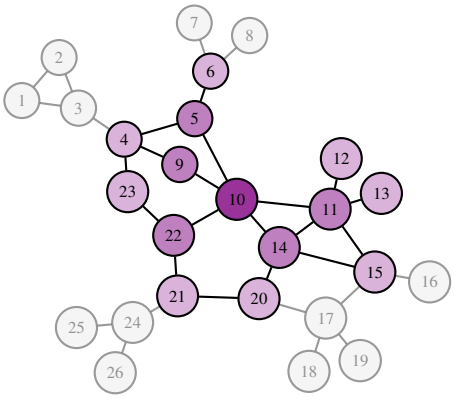
Likewise, the BFSS adds vertices to the sample gradually, starting from the ones closer to the seed until the desired sample size is reached. The sampled graph can be obtained by the explored vertices and the edges that reached them, however, this would always create a tree, so we construct the sampled graph as the subgraph induced from the sampled vertices.

Although it has unknown statistical properties and is empirically known to be biased toward high degree nodes (KURANT; MARKOPOULOU; THIRAN, 2011), it is widely used to

Algorithm 2 – Incident subgraph sampling.

procedure ESS(G,N)
input: Graph $G = (V,E)$; Sample size N.
output: Sampled graph G'.
    let E' be the set of sampled edges
    let V' be the set of sampled vertices
    let C be the set of edges that are not in the sample
    initialize E' with the empty set
    initialize V' with the empty set
    initialize C with E
    while $|E'|$ is less than N **do**
        let e be the sampled edge in this step
        draw e from C randomly and uniformly
        remove e from C
        add e to E'
    end while
    for all edge $e = u,v$ in E' **do**
        add u to V'
        add v to V'
    end for
    return $G' = (V',E')$
end procedure

Figure 8 – Example of a graph sampled by the BFSS method.



Source: Elaborated by the author.

sample social networks (MISLOVE *et al.*, 2007; AHN *et al.*, 2007; WILSON *et al.*, 2009) as its sample is a plausible graph on its own (KURANT; MARKOPOULOU; THIRAN, 2011).

The algorithm is shown in Algorithm 3 and an example of a sample can be seen in Figure 8.

Algorithm 3 – Breadth-first search sampling.

```
1: procedure BFSS(G,N)
input: Graph G; Sample size N; Seed vertex n_0.
output: Sampled graph G'.
2:     let V' be the set of sampled vertices
3:     let Q be the queue of explored vertices
4:     initialize V' with the empty set
5:     initialize Q with the empty queue
6:     enqueue n_0 in Q
7:     while Q is not empty and |V'| is less than N do
8:         let u be the foremost element of Q
9:         dequeue u from Q
10:        let N_u be randomly ordered set of neighbors of u in G
11:        for all vertex v in N_u such that v is not in V' do
12:            if |V'| is less than SSize then
13:                add v to V'
14:                enqueue v in Q
15:            end if
16:        end for
17:    end while
18:    let G' be the subgraph of G induced from V'
19:    return G'
20: end procedure
```

2.4.4 Random walk sampling (RWS)

The random walk sampling method (RWS) is a random walk based sampling method. A random walk in a graph is a special case of a Markov Chain, in which the states are vertices and the next vertex is drawn randomly and uniformly from the neighbors of the current vertex. The transition matrix $P(u,v)$ that describes the probability of transitioning to vertex v, given that the walk is currently in vertex u is

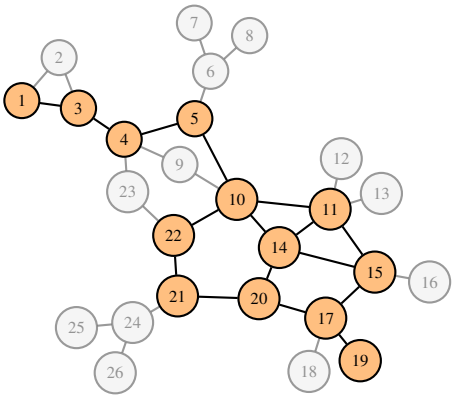
$$P(u,v) = \begin{cases} \frac{1}{k_u} & \text{if } u \text{ and } v \text{ are neighbors,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.38}$$

In the RWS method, the set of sampled vertices is obtained by navigating through the edges of the graph. As in the random walk, the next vertex to be visited is chosen randomly and uniformly among the neighbors of the current active vertex (RIBEIRO; TOWSLEY, 2010). Each visited vertex is added once in the sample although it may be visited more than once. The sampled graph is the subgraph induced from the sampled vertices.

The RWS method is biased toward higher degree vertices as the stationary probability of a vertex being visited in a random walk is proportional to its degree (STUTZBACH *et al.*, 2009) and equal to

$$\pi_u = \frac{k_u}{2|\mathcal{E}|}. \tag{2.39}$$

Figure 9 – Example of a graph sampled by the RWS method.



Source: Elaborated by the author.

It is possible to remove the bias by changing the transition probabilities in order to obtain a uniform sample of the nodes as in the Metropolis-Hastings random walk method, that is based on the Metropolis-Hastings algorithm.

The Metropolis-Hasting algorithm is a Markov Chain Monte Carlo method that constructs a Markov Chain with the desired probability distribution as its stationary distribution (HASTINGS, 1970). To obtain a stationary distribution μ_u, it creates a modified transition matrix $Q(u,v)$, given by

$$Q(u,v) = \begin{cases} P(u,v)\min\left(\frac{\mu_v P(v,u)}{\mu_u P(u,v)}, 1\right) & \text{if } u \neq v, \\ 1 - \sum_{u \neq v} Q(u,v) & \text{otherwise.} \end{cases} \tag{2.40}$$

In the uniform case, $\frac{\mu_v}{\mu_u} = 1$, leading to

$$Q(u,v) = \begin{cases} \frac{1}{k_u}\min\left(\frac{k_u}{k_v}, 1\right) & \text{if } u \text{ and } v \text{ are neighbors,} \\ 1 - \sum_{u \neq v} Q(u,v) & \text{if } v = u, \\ 0 & \text{otherwise.} \end{cases} \tag{2.41}$$

The algorithm for the RWS method is provided in Algorithm 4 and an example of sample can be seen in Figure 9.

2.5 Related work

Various works have studied the influence of structure on spreading processes processes. For instance, Newman (NEWMAN, 2002) showed that epidemic outbreaks occur more easily in assortative networks, but spread less. In (KITSAK *et al.*, 2010), it was shown that the coreness represents better the influence of a node in an epidemic spreading than its degree, while in (BORGE-HOLTHOEFER; MORENO, 2012), it was verified that there are no influential spreaders in a rumor spreading. Watts and Strogatz (WATTS; STROGATZ, 1998) studied the effect of the rewiring probability in small-world networks in the time needed for a disease

Algorithm 4 – Random walk sampling.

```
1: procedure RWS(G, N)
input: Graph G; Sample size N; Seed vertex n_0.
output: Sampled graph G'.
2:     let V' be the set of sampled vertices
3:     let u be the currently active vertex
4:     initialize V' with the empty set
5:     add n_0 to V'
6:     u ← n_0
7:     while |V'| is less than N do
8:         let N_u be set of neighbors of u in G
9:         draw a vertex v from N_u randomly and uniformly
10:        if v is not in V' then
11:            add v to V'
12:            u ← v
13:        end if
14:    end while
15:    let G' be the subgraph of G induced from V'
16:    return G'
17: end procedure
```

to reach the entire population and discovered that heterogeneous networks spread the disease quicker. In (HÉBERT-DUFRESNE; ALTHOUSE, 2015), the effect of clustering on epidemic spreading was studied for the contact process, in which every node contacts one neighbor per time step. It was shown that the disease tends to be kept within the clusters, hindering the spreading. Arruda (ARRUDA *et al.*, 2013) applied bayesian inference to study the influence of topology over the synchronization of Kuramoto oscillators and showed that the average shortest path length is the most influential topological measure.

Sampling in complex networks was addressed in various works with different goals. For example, Stumpf (STUMPF; WIUF; MAY, 2005) showed that samples of scale-free network generated by random sampling are not strictly scale-free. In (LESKOVEC; FALOUTSOS, 2006), sampling methods were compared with two different goals: back-in-time sampling and scale-down sampling. Back-in-time sampling means generating a sample that is similar to the original network in a point in time in the past, while scale-down sampling means obtaining a sample that is similar to the original network, but in a smaller scale. Maiya (MAIYA, 2011) compared sampling methods when trying to obtain hierarchical information of the network and representative samples of comunities. In the same work, the methods were compared when trying to approximate the set of best spreaders of information. The bias of samples in plant-pollinator networks was studied by using bootstrap confidence intervals in (LIN, 2015). In (KURANT; MARKOPOULOU; THIRAN, 2011), the degree bias of the BFSS method was analytically quantified.

CHAPTER

3

METHODS

This chapter describes in details the methods and parameters employed throughout this work.

3.1 Regression Analysis

3.1.1 Data collection

3.1.1.1 Independent and dependent variables

To characterize the structure of the network, we chose the measures described in Section 2.1.2: (i) average search information (S), (ii) number of articulation points (A), (iii) second moment ($\langle k^2 \rangle$) and (iv) shannon entropy (H) of the degree distribution, (v) average efficiency (E), (vi) assortativity coefficient (r), (vii) average clustering coefficient ($\langle cc \rangle$), (viii) average coreness ($\langle kc \rangle$), (ix) transitivity (C) and (x) variance of the betweenness centrality ($V(B)$).

Because the measures have different scales, the regressors were obtained by normalizing the measures with the z-normalization: for a list of observations x with average $\langle x \rangle$ and standard deviation σ_x, an observation x_t is normalized as $z_t = \frac{x_t - \langle x \rangle}{\sigma_x}$. The probability distribution of each normalized measure has zero mean and unit variance.

The response variable was calculated by simulating the spreading process 100 times starting from each node of the network and averaging the result of the runs. For the epidemic spreading, we set the spreading probability to 0.6 and the recovery probability to 1 and for the rumor spreading, the spreading probability was set to 1 and the loss of interest probability to 0.6.

The rumor model was simulated by using the truncated process, in which a spreader contacts all its neighbors in a random order. If it turns into a stifler during a contact, it immediately stops contacting further nodes (BORGE-HOLTHOEFER; MORENO, 2012), and the epidemic spreading was simulated by the reactive process, in which an infected node contacts all of its

neighbors in every time step.

3.1.1.2 Networks

The data set contains 500 networks generated by the models presented in Section 2.1.3: ER, BA, NLBA, WS, Waxman, Spatial SF, as well as SF networks generated by the configuration model and rewired to become more or less assortative. The models were chosen as they generate networks with different topological properties (COSTA *et al.*, 2007), yielding variability in the independent variables. The networks were generated with $|\mathcal{V}| = 1000$ and $\langle k \rangle = 4$, as we are interested in the influence of the organization of the edges in the dynamical processes.

Table 1 shows the parameters used for each model. Concealed parameters were calculated in a way that the average degree remained constant for all the networks.

Table 1 – Model variations for generating networks.

Model	Parameters
ER	-
BA	-
NLBA	$\alpha = 1.3, 1.5$
WS	$p = 0.1, 0.3$
Waxman	$\alpha = 0.3$
Spatial SF	$\alpha = 0.3$
Assortative SF	$\gamma = 3.0$, iterations $= 2000$
Disassortative SF	$\gamma = 3.0$, iterations $= 2000$

For each model variation, we generated 50 networks, yielding the total of 500 networks.

3.1.2 Analysis

The data was analysed with the variable dispersion beta regression model because the dependent variable is a proportion, restricted to the interval $(0,1)$ and other regression models may not be appropriate for such situations, since they may yield fitted values that lay outside of this range (FERRARI; CRIBARI-NETO, 2004).

In our model, the mean μ_t depends on regressors (i) to (ix) and the precision parameter ϕ_t depends on regressors (iv), (vii) and (x). As link functions, we adopted the log-log link for the mean and the log link for the precision as suggested in (FERRARI; CRIBARI-NETO, 2004; CRIBARI-NETO; ZEILEIS, 2009). The regression equations are:

$$g_1(\mu) = \beta_0 + \beta_1 S + \beta_2 A + \beta_3 \langle k^2 \rangle + \beta_4 E + \beta_5 H + \beta_6 r + \beta_7 \langle cc \rangle + \beta_8 \langle kc \rangle + \beta_9 C, \tag{3.1}$$

and

$$g_2(\phi) = \gamma_0 + \gamma_1 V(B) + \gamma_2 \langle cc \rangle + \gamma_3 H + \gamma_4 V(B) \langle cc \rangle + \gamma_5 V(B) H + \gamma_6 \langle cc \rangle H + \gamma_7 V(B) \langle cc \rangle H. \tag{3.2}$$

Interaction terms were not added in the regression model for the mean because of the large number of regressors and because it complicates the interpretation of the effects. Also, it was later verified that they are not needed.

3.2 Network Sampling

3.2.1 Data collection

3.2.1.1 Structural properties

To compare the structure of the original versus sampled networks, we analyse the degree distribution and measures that account for other important properties of a network, such as connectivity and clustering: average degree, transitivity and the spectral radius.

The average degree is calculated over the sampled graph, that is, we take the average of the degrees of the nodes in the sampled graph, instead of the degrees in the original graph.

3.2.1.2 Functional properties

The behaviour of the SIR dynamic was represented by the epidemic threshold and the curve of removed individuals per infection rate (henceforth called SIR or evolution curve).

The epidemic threshold was approximated by the inverse of the largest eigenvalue of the network and the SIR curve was obtained by simulating the SIR dynamic 30 times per network (original and sampled) and averaging the result of the runs.

3.2.1.3 Networks

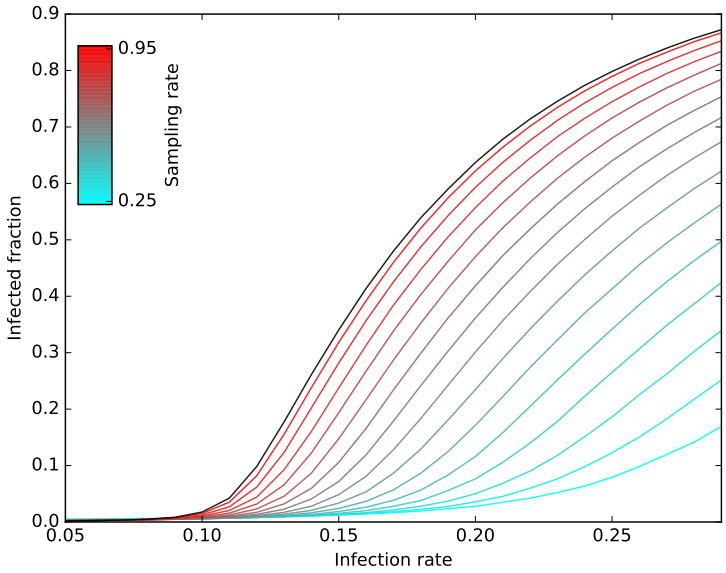
The data set was extracted from two real world networks and two sets of synthetic networks, one composed of ER networks and one composed of uncorrelated scale-free networks generated by the configuration model with $\gamma = 2.2$ (Uncorrelated SF). The ER and Uncorrelated SF networks have $\mathcal{V} = 1000$ and $\langle k \rangle \approx 10$. The real world networks are a subset of the Facebook network (NIPS) (MCAULEY; LESKOVEC, 2012) and the e-mail communication network of the University Rovira i Virgili (E-mail) (GUIMERA *et al.*, 2003).

3.2.2 Sampling methods

We considered four sampling methods: induced subgraph sampling (VSS), incident subgraph sampling (ESS), breadth-first-search sampling (BFSS) and random walk sampling (RWS).

Each method was used to sample networks with various sampling rates, ranging from 25% to 95% of the network. For each sampling rate, the real world networks were sampled 50

Figure 10 – SIR curves of subgraphs of an ER network for different sampling rates.



Source: Elaborated by the author.

times and each synthetic network was sampled 10 times, generating 50 sampled networks per real world network and 500 sampled networks per synthetic network set.

3.2.3 Approximation of the SIR curve

The SIR curve of the original network seems to be related to the one of sampled networks, as shown in Figure 10. In order to approximate the original curve, we chose to rescale the x-axis (infection rate) by multiplying it by a constant. The constant was chosen to be the ratio between the spectral radius of the original and sampled networks, so that the epidemic thresholds match. As a best case scenario, we use the rescaling factor that yields the closest curve considering the Fréchet distance (see Appendix A) between the original and rescaled SIR curves.

CHAPTER

4

RESULTS

This chapter presents the results of this work. Section 4.1 focuses on the influence of the topology of the network in spreading processes and Section 4.2 presents our findings on network sampling.

4.1 Regression analysis

Tables 2 and 3 present the coefficients of the regression model for the mean in the epidemics and rumor models, and the coefficients of the regression model for the precision parameter are shown in Tables 4 and 5.

	Mean	Std. error	p-value	Measure
β_0	1.913	0.001	≈ 0	–
β_1	-0.2	0.008	≈ 0	S
β_2	-0.123	0.004	≈ 0	A
β_3	-0.097	0.011	≈ 0	$\langle k^2 \rangle$
β_4	0.108	0.013	≈ 0	E
β_5	0.063	0.008	≈ 0	H
β_6	-0.134	0.002	≈ 0	r
β_7	0.002	0.012	0.843	$\langle cc \rangle$
β_8	0.066	0.008	≈ 0	$\langle kc \rangle$
β_9	0.119	0.012	≈ 0	C

Table 2 – SIR – Estimates of the β parameters.

The measure that is statistically significant (has a p-value smaller than 0.05) and presents the largest coefficient in absolute value is the average search information (S) in both cases, i.e. disease and rumor spreading, indicating that it is the measure that most influences in both spreading processes. In addition, the coefficients related to the average search information are

	Mean	Std. error	p-value	Measure
β_0	1.43	0.001	≈ 0	–
β_1	-0.29	0.01	≈ 0	S
β_2	-0.142	0.005	≈ 0	A
β_3	0.196	0.017	≈ 0	$\langle k^2 \rangle$
β_4	-0.231	0.018	≈ 0	E
β_5	0.043	0.008	≈ 0	H
β_6	≈ 0	0.002	0.813	r
β_7	-0.188	0.02	≈ 0	$\langle cc \rangle$
β_8	0.127	0.008	≈ 0	$\langle kc \rangle$
β_9	0.072	0.017	≈ 0	C

Table 3 – ISR – Estimates of the β parameters.

negative, so harder navigation implies in fewer contacted nodes during an epidemic or rumor spreading.

The number of articulation points (A) also influences negatively the outcome of both dynamics, because if a node that is an articulation point does not become a spreader, a part of the network has no chance of being reached. Conversely, the transitivity (C) and the avearage coreness ($\langle kc \rangle$) influence positively because triangles and cores create redundant paths through which the spread may continue if a node is not contacted.

The assortativity (r) is not influential in the rumor spreading, and affects negatively the epidemic spreading. For the epidemic spreading, this means that the fraction of individuals that are infected is larger in networks that are disassortative, as discussed in (NEWMAN, 2002) and also shown for the percolation process.

The average clustering coefficient ($\langle cc \rangle$) is not influential in the epidemic spreading and influences negatively the rumor spreading and the efficiency (E) affects positively in the SIR dynamic, but negatively in the ISR dynamic. This is due to differences in the removal process, that is spontaneous in the epidemic spreading and by contact in the rumor spreading. A high average clustering coefficient means a high average probability of having connected neighbors, increasing the chance of contact between spreaders. In a similar way, a higher efficiency favours the spread of a disease, but may hinder the spread of a rumor as it facilitates the contacts of type spreader-spreader and spreader-stifler.

The diversity of degrees in the network, represented by the Shannon entropy of the degree distribution (H) has a positive impact in both epidemic and rumor spreading. In contrast, the second moment of the degree distribution ($\langle k^2 \rangle$) has a positive coefficient in the ISR dynamics while for the SIR dynamics, the coefficient is negative. A large $\langle k^2 \rangle$ for a fixed average degree means that there are some few very large hubs. If the average degree is low, as in this case, the edges will mostly connect a low-degree node to one of the few hubs, making the hubs responsible for the spreading.

In the case of epidemic spreading, having smaller hubs and more triangles can be more beneficial, as the spreading will still continue after the hubs are removed. In the case of rumor spreading, the same is true, however, having a few large hubs means that when the hub tries to spread the rumor to its neighbors, there will be only a few neighbors that are spreaders or stiflers, meaning that it will spread the rumor to a larger fraction of the network before turning into a stifler.

This can be illustrated for a star-like network with two large hubs that connect to the remaining nodes, which can be obtained by the NLBA model with a non-linearity parameter of 3.0. For this network, the average fraction of removed nodes is 0.84 and the average fraction of nodes that got to know about the rumor is 0.75. Both values are high, but for the epidemic spreading, the removed fraction in the training data lies in the range from approximately 0.83 to 0.91, while in the rumor spreading, the stifler fraction is in the range between 0.47 to 0.89. Because of that, even though the removed fraction is higher than the stifler fraction, the coefficient associated to $\langle k^2 \rangle$ in the epidemic spreading is negative, while in the rumor spreading is positive.

	Mean	Std. error	p-value	Measure
γ_0	9.611	0.151	≈ 0	–
γ_1	0.065	0.158	0.681	$V(B)$
γ_2	-0.562	0.304	0.064	$\langle cc \rangle$
γ_3	-0.388	0.201	0.053	H
γ_4	-0.483	0.346	0.163	$V(B) : \langle cc \rangle$
γ_5	-0.191	0.333	0.566	$V(B) : H$
γ_6	-0.01	0.263	0.968	$\langle cc \rangle : H$
γ_7	-0.223	0.183	0.224	$V(B) : \langle cc \rangle : H$

Table 4 – SIR – Estimates of the γ parameters.

	Mean	Std. error	p-value	Measure
γ_0	8.828	0.151	≈ 0	–
γ_1	-1.863	0.158	≈ 0	$V(B)$
γ_2	-1.007	0.303	0.001	$\langle cc \rangle$
γ_3	-1.194	0.201	≈ 0	H
γ_4	0.134	0.346	0.698	$V(B) : \langle cc \rangle$
γ_5	-0.875	0.333	0.009	$V(B) : H$
γ_6	0.338	0.263	0.198	$\langle cc \rangle : H$
γ_7	-0.021	0.183	0.911	$V(B) : \langle cc \rangle : H$

Table 5 – ISR – Estimates of the γ parameters.

The precision parameter may be considered constant across observations in the SIR model, while it depends on the variance of the betweenness centrality ($V(B)$), the average clustering coefficient ($\langle cc \rangle$) and the Shannon entropy of the degree distribution (H), as well as a first order interaction between $V(B)$ and H in the ISR model.

For the rumor dynamics, a large average clustering coefficient implies that a spreader is likely to have neighbors that also are spreaders, as neighbors of your neighbors are likely to be your neighbors. Hence, depending on the order of contact of the neighbors, a node may spread the rumor more or less, thus the negative regression parameter for $\langle cc \rangle$. The variance of the betweenness centrality and the Shannon entropy of the degree distribution can be seen as measures of diversity of the nodes, in terms of the number of links and importance in communication. The negative values of the regression parameters for $V(B)$, H, and for the first order interaction indicate that heterogeneous networks show larger dispersion in the outcome of the rumor spreading.

The pseudo R^2 of the models are 0.98 (SIR) and 0.99 (ISR), suggesting that most of the variability in the data is explained by the proposed models. Figure 11 presents the half-normal plot and predicted versus observed values plot for both dynamics.

Most of the data points lie inside of the 95% confidence interval delimited by the black lines in both plots, which indicates that the regression model is adequate for both the SIR and ISR dynamics. This can also be seen in the predicted versus observed values plots. In both cases, the correlation between the predicted and measured values of the variable of interest is larger than 0.95.

4.2 Network sampling

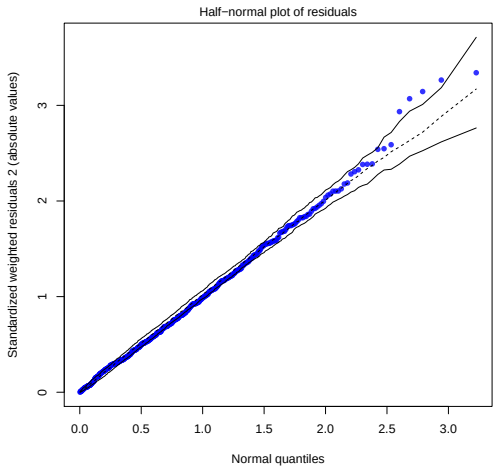
4.2.1 Structural properties

In general, the sampled networks do not represent well the original network considering the analysed properties. The 95% confidence interval for the average degree and transitivity are presented on Figures 12 and 15, respectively. The colored range represents the 95% confidence interval of the sample and the red dotted line is the value of the measure in the original network.

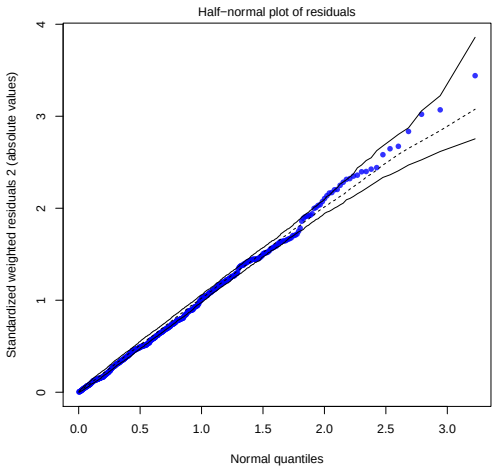
The average degree is not maintained by any of the sampling methods. The BFSS and RWS methods overestimate the average degree for the real world networks because they are biased to sample hubs, that are not present in ER networks and are small and numerous in the uncorrelated SF network. Nevertheless, the bias by itself does not explain why hubs in the original network are still hubs in the sample.

In the BFSS method, this happens by construction, because after exploring a hub, the neighbors of the hub will be enqueued to be explored. Therefore, they are likely to be in the sample, except for limitations in the sample size. In contrast, in the RWS method, there is no guarantee that many neighbors of a hub will be visited. However, the random walk is biased toward hubs and the process is memoryless, so a hub is likely to be visited many times, and every visit is a chance of visiting a new neighbor. This is illustrated in Figure 13, that shows the distribution of the excess visits according to the degree compared to the degree distribution for

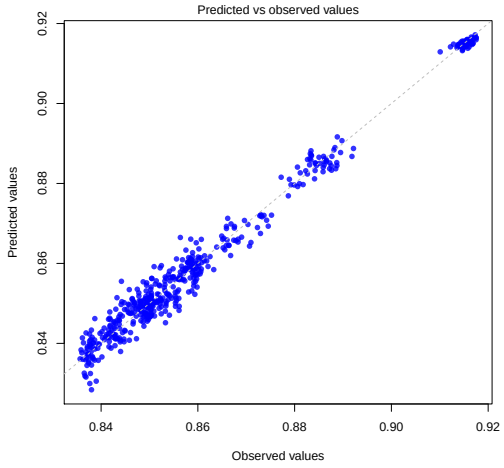
Figure 11 – Half-normal plot of residuals and predicted versus observed values plots.



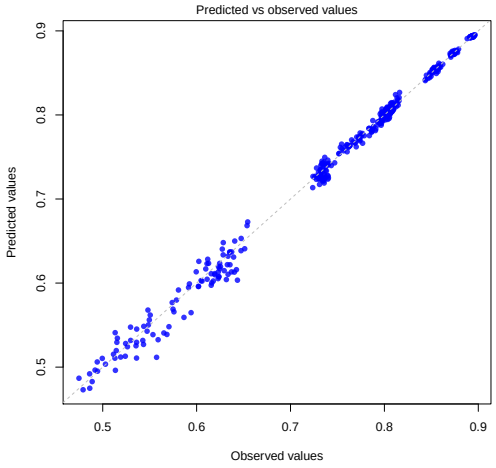
(a) Half normal plot of residuals for the SIR model.



(b) Half normal plot of residuals for the ISR model.



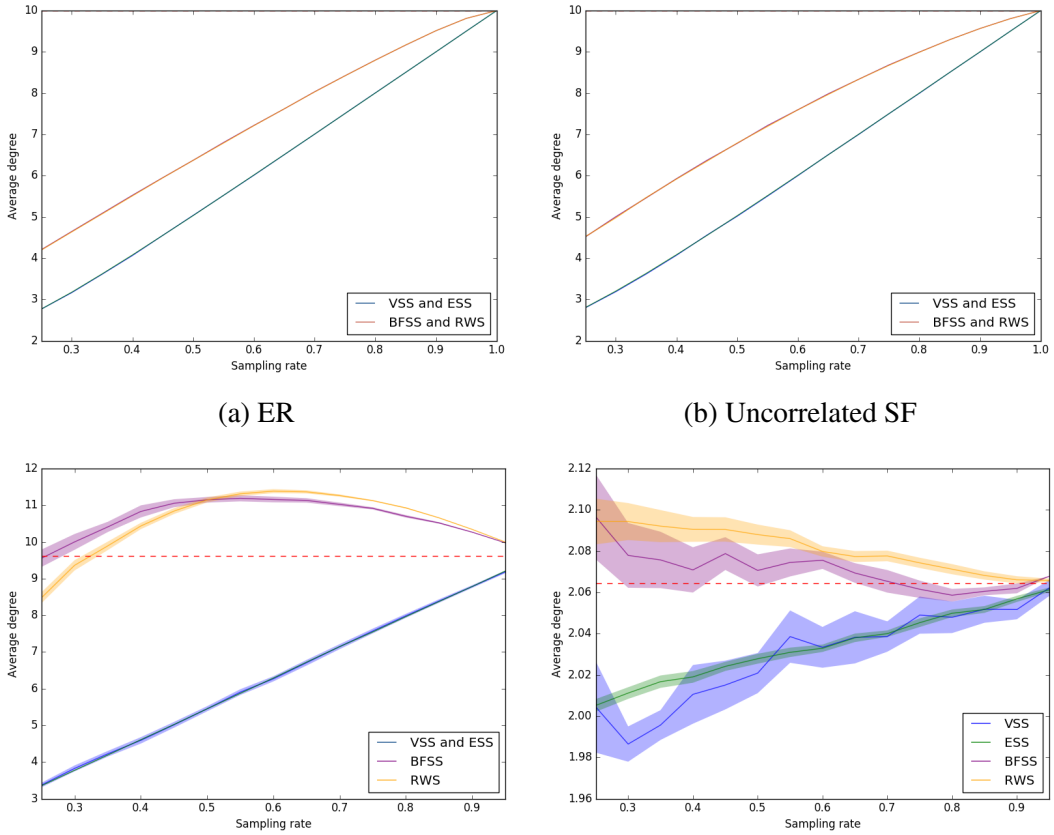
(c) Predicted vs Observed values for the SIR model.



(d) Predicted vs Observed values for the ISR model.

Source: Elaborated by the author.

Figure 12 – Average degree of samples of different networks.



(a) ER (b) Uncorrelated SF

(c) E-mail (d) NIPS

Source: Elaborated by the author.

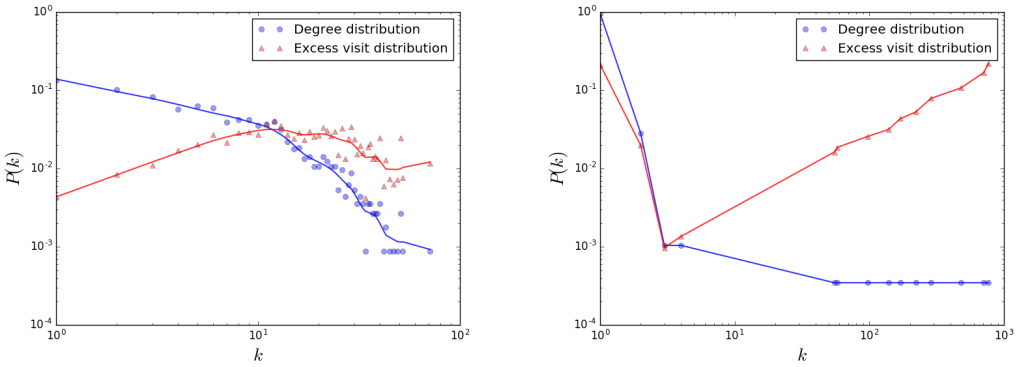
the E-mail and NIPS networks. The excess visit of a node is the number of times a node was visited aside from the first visit.

The ESS method is also biased toward hubs, but the procedure of getting all the vertices that are incident in sampled edges causes the sampled vertices set to be very large, while maintaining the edge set small. For instance, samples of an ER network with 25% of the edges (approximately 1250 edges) contain, in average, approximately 900 vertices (90% of the vertices). The VSS samples the vertices with a uniform probability, so hubs are unlikely to be chosen and the subgraphs have a small average degree.

The degree distribution is shown for samples of size 50% in Figure 14. The degree distribution of the original network differs significantly from that of the samples in all the cases.

It is interesting to notice that, on top of producing samples with the same average degree, the samples of the ESS and VSS methods have a close degree distribution, with a higher probability for intermediate values than for high degrees. In contrast, although the average degree is similar for the RWS and BFSS methods, the degree distribution of the samples have different properties, with the BFSS method presenting a higher probability in the extremes of the

Figure 13 – Excess visit distribution for samples with sampling rate of 25%.



(a) E-mail (b) NIPS

Source: Elaborated by the author.

distribution than the RWS method.

The VSS method is suitable to estimate the transitivity of the network, as the true transitivity lies inside the confidence interval for most sampling rates. The samples obtained via the ESS method present a smaller transitivity than the original network. This may occur because sampling a triangle requires that the three edges that compose it are sampled. In contrast, for crawling methods, it is only necessary that a vertex and its connected neighbors are sampled.

The analysis for the spectral radius is presented in the following subsection.

4.2.2 Epidemic spreading

4.2.2.1 Epidemic threshold

The epidemic threshold was approximated with the inverse of the principal eigenvalue of the adjacency matrix (λ_1^{-1}). Figure 16 shows the growth of λ_1^{-1} according to the sampling rate.

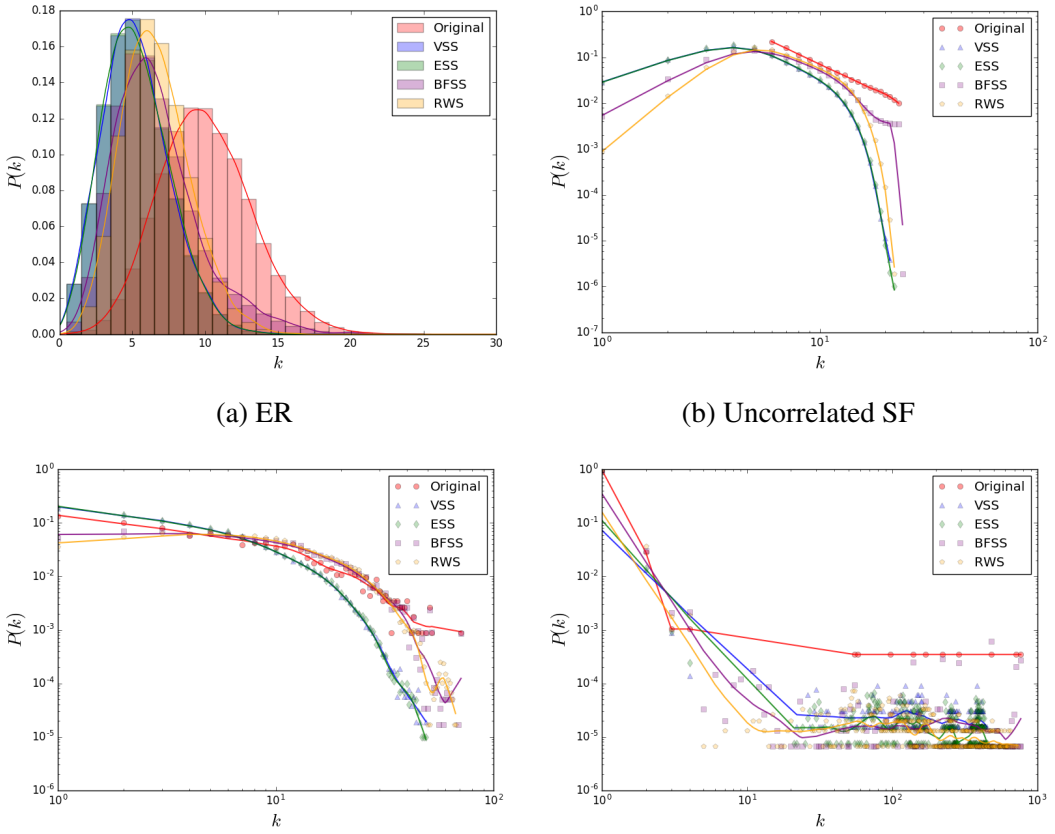
The sampling methods have a consistent performance across the networks. The BFSS is the method that best approximates λ_1^{-1}. Besides, the obtained value is closer for real world networks than for the synthetic networks.

Since the spectral radius of a graph is never smaller than the spectral radius of a subgraph obtained by removing any number of nodes, we can obtain a better estimate of λ_1^{-1} by choosing the minimum among the obtained values in a sample of subgraphs of a graph. In Figure 17, the lines represent the minimum value of the sample.

As expected, none of the curves underestimate the true value. The BFSS method shows the best results in all the networks, and for the real world networks, obtains the true value for a samples with 65% of the network.

There are cases, however, in which we are unable to have many samples of the network.

Figure 14 – Degree distribution of samples with sampling rate of 50%.



(a) ER (b) Uncorrelated SF

(c) E-mail (d) NIPS

Source: Elaborated by the author.

In these cases, it is desired to minimize λ_1^{-1} while maintaining the sample size. For this analysis, we will consider only the BFSS method, as it showed better results than the other methods.

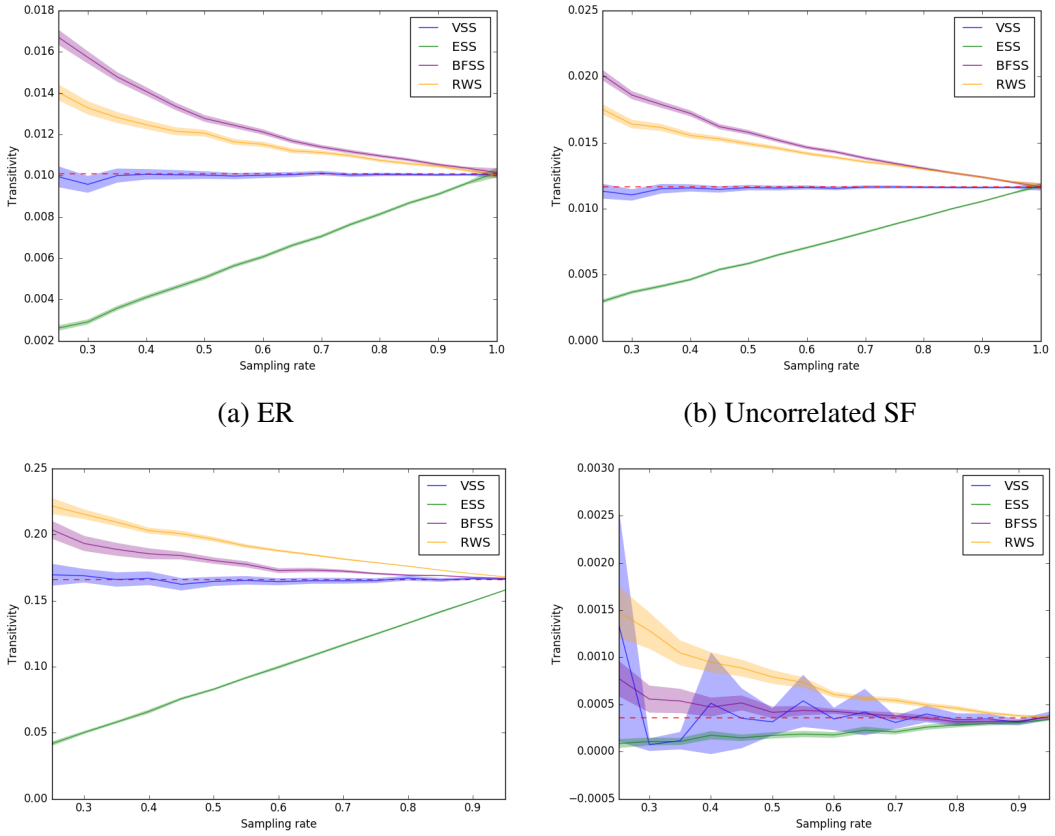
Our results suggest that for a constant sampling rate, more heterogeneous samples present lower values of λ_1^{-1}, as shown in Figure 18. The NIPS network is not shown, as the obtained value of λ_1^{-1} is already close to the true value for samples of 25% of the network.

4.2.2.2 Approximation of the SIR curve

All sampling methods underestimate the final fraction of removed individuals for the ER and uncorrelated SF networks, while for the real world networks, the results vary, as shown in Figure 19 and 20 for samples of 30% and 50% of the network. When the sampling rate is increased, the curves get closer to the original curve. However, they are still far from being an accurate representation of it.

The ratio of the spectral radius is a good rescaling coefficient for the ER and uncorrelated SF networks, as well as the E-mail network, when considering only the VSS and ESS method. However, in the other cases, it yields curves that are further away than the non-rescaled curves.

Figure 15 – Transitivity of samples compared to the true value.



(a) ER

(b) Uncorrelated SF

(c) E-mail

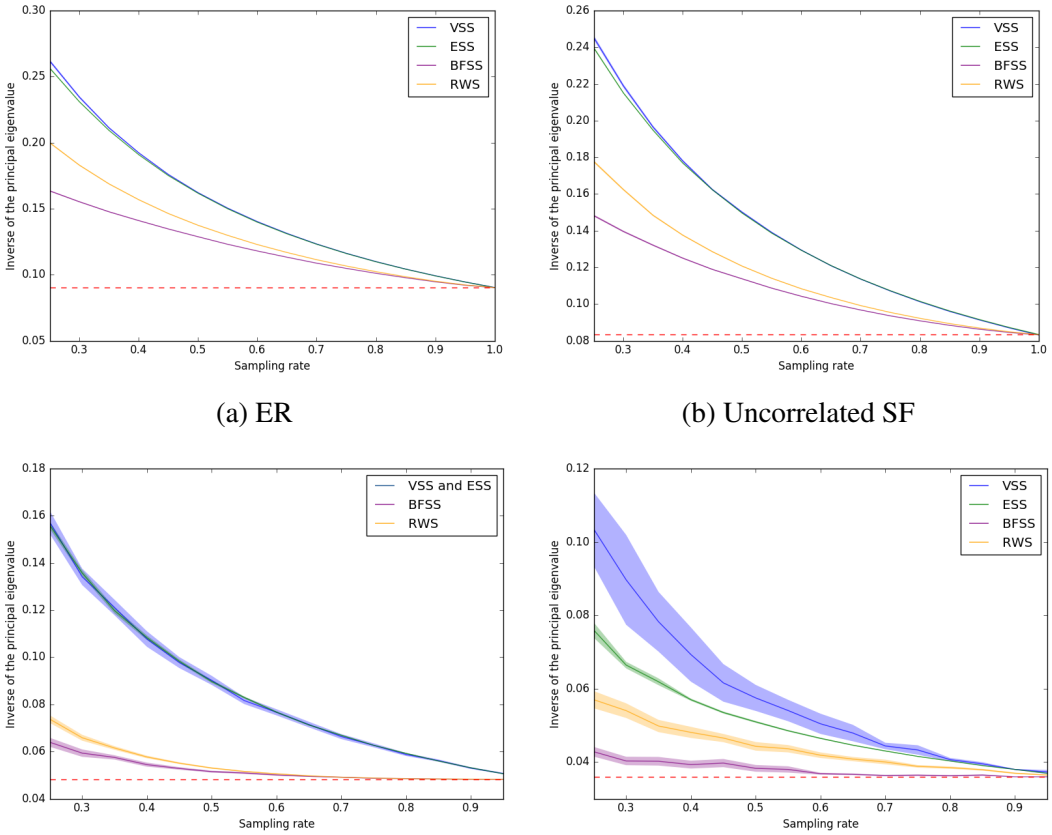
(d) NIPS

Source: Elaborated by the author.

This always happens in cases in which the sampled curve overestimates the true curve, as the ratio of the eigenvalues is always smaller than one because of the property that the largest eigenvalue of a sample never surpasses that of the original network.

The Fréchet-rescaled curves are close to the original curves, which tells us that it is possible to rescale the SIR curve so that the sampled curve has a similar shape to that of the original curve. Nevertheless, finding the Fréchet coefficient is not a trivial task and, for now, requires knowledge of the structure of the original network. We leave this step for further analysis.

Figure 16 – Inverse of the principal eigenvalue of samples compared to the true value.



(a) ER

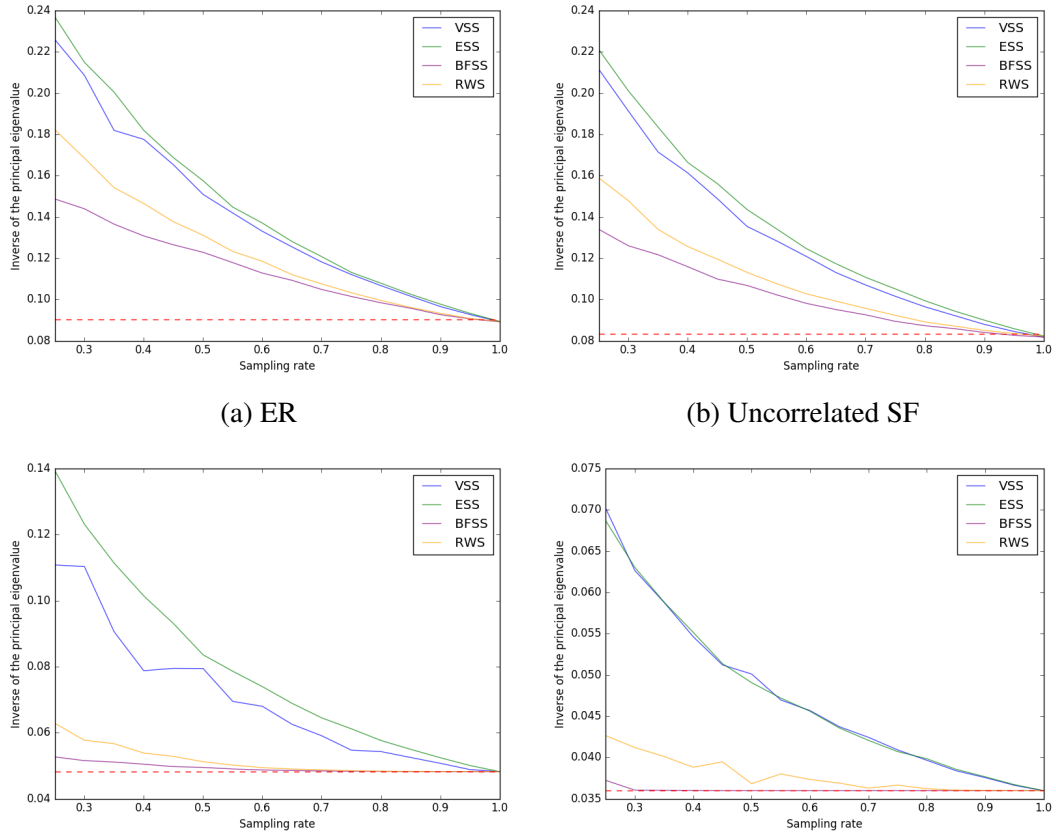
(b) Uncorrelated SF

(c) E-mail

(d) NIPS

Source: Elaborated by the author.

Figure 17 – Minimum inverse of the principal eigenvalue of samples compared to the true value.



(a) ER

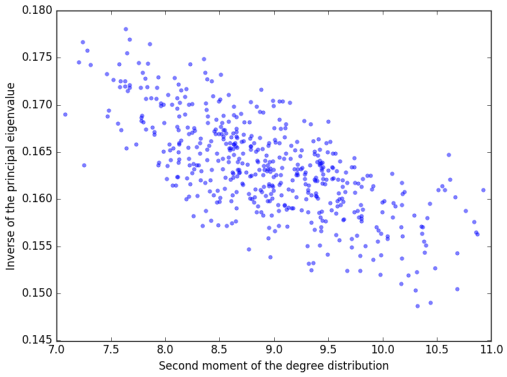
(b) Uncorrelated SF

(c) E-mail

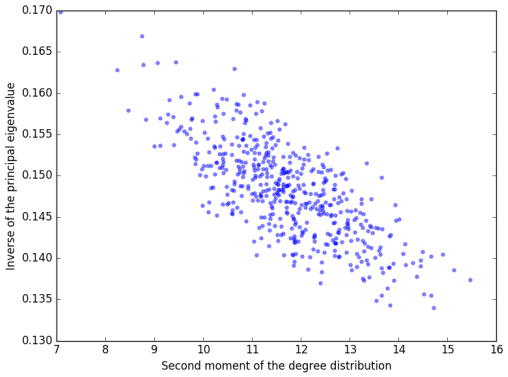
(d) NIPS

Source: Elaborated by the author.

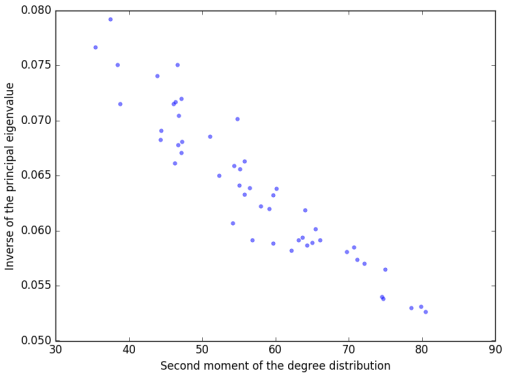
Figure 18 – Inverse of the principal eigenvalue versus second moment of the degree distribution for samples of 25% of the network.



(a) ER



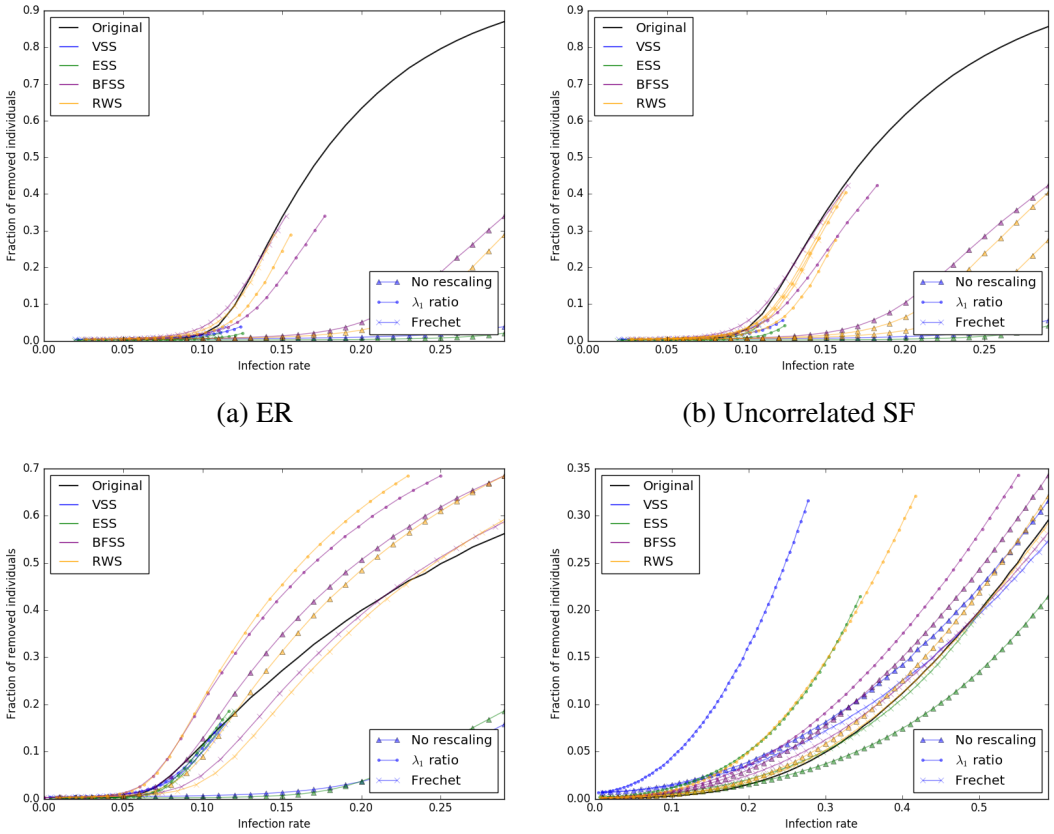
(b) Uncorrelated SF



(c) E-mail

Source: Elaborated by the author.

Figure 19 – Approximation of the SIR curve sampling 35% of the network.



(a) ER

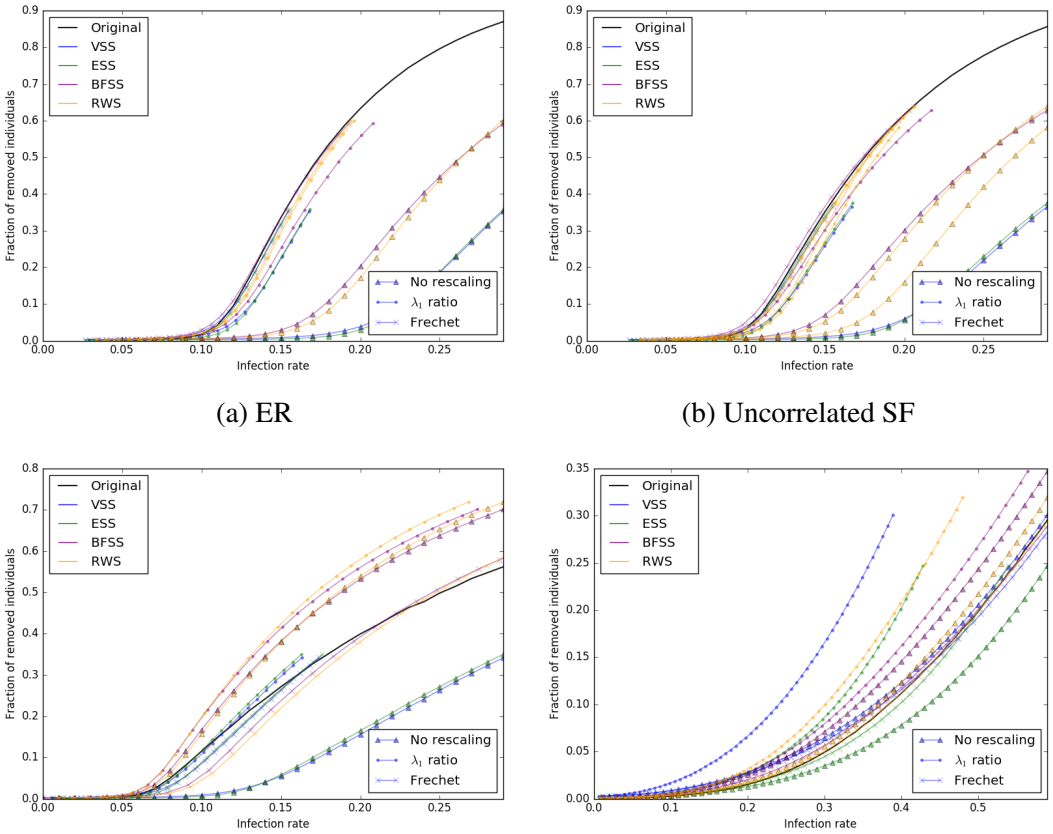
(b) Uncorrelated SF

(c) E-mail

(d) NIPS

Source: Elaborated by the author.

Figure 20 – Approximation of the SIR curve sampling 50% of the network.



(a) ER

(b) Uncorrelated SF

(c) E-mail

(d) NIPS

Source: Elaborated by the author.

CHAPTER

5

CONCLUSION

Many real world phenomena, such as the outbreak of a disease and a rumor, can be modeled and analysed as spreading processes in complex networks. As the size of the networks and availability of the data grows, it is necessary to study them by means of topological measures, such as the number of vertices per node and the degree correlation between nodes that are incident in an edge. Complementarily, it is possible to reduce the size of the network by sampling a subgraph of it.

In this work, we have shown that the organization of the edges in a network influences the final fraction of removed individuals in an epidemic spreading and the final fraction of individuals that got to know about a rumor and that the most influential topological measure in both dynamics is the average search information, that quantifies the ease or difficulty of navigating through the network. Although the most influential measure is the same for both networks, there are differences, such as the need of a variable precision parameter in the rumor dynamic.

We also studied the behaviour of structural measures, such as the average degree, when a network is sampled by four sampling methods. Our results have shown that the structure of the sampled network differs significantly from the original network. Therefore, conclusions made over sampled network data should not be carelessly generalized to the original network.

As an application of sampling for dynamical processes, we estimated the epidemic threshold based on sampled networks generated by four sampling methods. Our results indicate that the most appropriate method, in this case, is the breadth-first-search sampling method, which produces the closest estimations for lower sampling rates and that more heterogeneous samples yield closer estimates.

Finally, we tried to approximate the evolution curve of the SIR dynamic by rescaling it using the ratio between the epidemic threshold of the sample and the original network. The rescaling coefficient was proved to be inadequate, although it matches the epidemic threshold of

the two networks, there is a limitation that curves that overestimate the real curve will be rescaled to overestimate it even more. However, our results also suggest that the rescaling is possible, as the curve generated by the Fréchet coefficient, determined empirically and with knowledge of the original network, is close to the real one.

BIBLIOGRAPHY

AHN, Y.-Y.; HAN, S.; KWAK, H.; MOON, S.; JEONG, H. Analysis of topological characteristics of huge online social networking services. In: ACM. **Proceedings of the 16th international conference on World Wide Web**. [S.l.], 2007. p. 835–844. Citation on page 49.

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of modern physics**, APS, v. 74, n. 1, p. 47, 2002. Citation on page 27.

ALT, H.; GODAU, M. Computing the fréchet distance between two polygonal curves. **International Journal of Computational Geometry & Applications**, World Scientific, v. 5, n. 01n02, p. 75–91, 1995. Citation on page 79.

AMARAL, L. A.; OTTINO, J. M. Complex networks. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 38, n. 2, p. 147–162, 2004. Citations on pages 27 and 34.

ANDERSON, R. M.; MAY, R. M.; ANDERSON, B. **Infectious diseases of humans: dynamics and control**. [S.l.]: Wiley Online Library, 1992. Citations on pages 39 and 40.

ARRUDA, G. F. de; PERON, T. K. D.; ANDRADE, M. G. de; ACHCAR, J. A.; RODRIGUES, F. A. The influence of network properties on the synchronization of kuramoto oscillators quantified by a bayesian regression analysis. **Journal of Statistical Physics**, Springer, v. 152, n. 3, p. 519–533, 2013. Citations on pages 28 and 52.

ATKINSON, A. Two graphical displays for outlying and influential observations in regression. **Biometrika**, Biometrika Trust, v. 68, n. 1, p. 13–20, 1981. Citation on page 45.

BARABÁSI, A.-L. The architecture of complexity. **IEEE control systems**, IEEE, v. 27, n. 4, p. 33–42, 2007. Citation on page 28.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citations on pages 27, 36, and 37.

BARTHÉLEMY, M. Crossover from scale-free to spatial networks. **EPL (Europhysics Letters)**, IOP Publishing, v. 63, n. 6, p. 915, 2003. Citations on pages 34, 37, and 38.

BATAGELJ, V.; ZAVERSNIK, M. An o (m) algorithm for cores decomposition of networks. **arXiv preprint cs/0310049**, 2003. Citation on page 34.

BJØRNSTAD, O. N.; FINKENSTÄDT, B. F.; GRENFELL, B. T. Dynamics of measles epidemics: estimating scaling of transmission rates using a time series sir model. **Ecological Monographs**, Wiley Online Library, v. 72, n. 2, p. 169–184, 2002. Citation on page 39.

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks: Structure and dynamics. **Physics reports**, Elsevier, v. 424, n. 4, p. 175–308, 2006. Citations on pages 28 and 29.

BOGUNÁ, M.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Absence of epidemic threshold in scale-free networks with degree correlations. **Physical review letters**, APS, v. 90, n. 2, p. 028701, 2003. Citation on page 40.

BORGE-HOLTHOEFER, J.; MORENO, Y. Absence of influential spreaders in rumor dynamics. **Physical Review E**, APS, v. 85, n. 2, p. 026116, 2012. Citations on pages 51 and 53.

CORMEN, T. H.; LEISERSON, C. E.; RIVEST, R. L.; STEIN, C. **Introduction to algorithms**. [S.l.]: MIT press Cambridge, 2001. Citation on page 48.

COSTA, L. d. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. **Advances in physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007. Citations on pages 33, 35, 36, 37, and 54.

CRIBARI-NETO, F.; ZEILEIS, A. Beta regression in r. Department of Statistics and Mathematics x, WU Vienna University of Economics and Business, 2009. Citations on pages 45 and 54.

DALEY, D.; KENDALL, D. G. Stochastic rumours. **IMA Journal of Applied Mathematics**, IMA, v. 1, n. 1, p. 42–55, 1965. Citation on page 40.

DIMITROPOULOS, X. A.; RILEY, G. F. Creating realistic bgp models. In: IEEE. **Modeling, Analysis and Simulation of Computer Telecommunications Systems, 2003. MASCOTS 2003. 11th IEEE/ACM International Symposium on**. [S.l.], 2003. p. 64–70. Citation on page 28.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. [S.l.]: John Wiley & Sons, 2014. Citations on pages 42 and 43.

ERDÖS, P.; RÉNYI, A. On random graphs, i. **Publicationes Mathematicae (Debrecen)**, v. 6, p. 290–297, 1959. Citation on page 36.

EULER, L. Solutio problematis ad geometriam situs pertinentis. **Commentarii academiae scientiarum Petropolitanae**, v. 8, p. 128–140, 1741. Citation on page 27.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. In: ACM. **ACM SIGCOMM computer communication review**. [S.l.], 1999. v. 29, n. 4, p. 251–262. Citation on page 27.

FARKAS, I. J.; DERÉNYI, I.; BARABÁSI, A.-L.; VICSEK, T. Spectra of "real-world" graphs: Beyond the semicircle law. **Physical Review E**, APS, v. 64, n. 2, p. 026704, 2001. Citation on page 35.

FERRARI, S.; CRIBARI-NETO, F. Beta regression for modelling rates and proportions. **Journal of Applied Statistics**, Taylor & Francis, v. 31, n. 7, p. 799–815, 2004. Citations on pages 44, 45, and 54.

FRÉCHET, M. M. Sur quelques points du calcul fonctionnel. **Rendiconti del Circolo Matematico di Palermo (1884-1940)**, Springer, v. 22, n. 1, p. 1–72, 1906. Citation on page 79.

FREEMAN, L. C. A set of measures of centrality based on betweenness. **Sociometry**, JSTOR, p. 35–41, 1977. Citation on page 34.

GIBBONS, A. **Algorithmic graph theory**. [S.l.]: Cambridge University Press, 1985. Citation on page 31.

GJOKA, M.; KURANT, M.; BUTTS, C. T.; MARKOPOULOU, A. Walking in facebook: a case study of unbiased sampling of osns. In: IEEE. **Infocom, 2010 Proceedings IEEE**. [S.l.], 2010. p. 1–9. Citation on page 28.

____. Practical recommendations on crawling online social networks. **IEEE Journal on Selected Areas in Communications**, IEEE, v. 29, n. 9, p. 1872–1892, 2011. Citation on page 28.

GUIMERA, R.; DANON, L.; DIAZ-GUILERA, A.; GIRALT, F.; ARENAS, A. Self-similar community structure in a network of human interactions. **Physical review E**, APS, v. 68, n. 6, p. 065103, 2003. Citation on page 55.

HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. **Biometrika**, Biometrika Trust, v. 57, n. 1, p. 97–109, 1970. Citation on page 51.

HÉBERT-DUFRESNE, L.; ALTHOUSE, B. M. Complex dynamics of synergistic coinfections on realistically clustered networks. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 112, n. 33, p. 10551–10556, 2015. Citation on page 52.

KERMACK, W. O.; MCKENDRICK, A. G. A contribution to the mathematical theory of epidemics. In: THE ROYAL SOCIETY. **Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences**. [S.l.], 1927. v. 115, n. 772, p. 700–721. Citation on page 39.

KITSAK, M.; GALLOS, L. K.; HAVLIN, S.; LILJEROS, F.; MUCHNIK, L.; STANLEY, H. E.; MAKSE, H. A. Identification of influential spreaders in complex networks. **Nature physics**, Nature Publishing Group, v. 6, n. 11, p. 888–893, 2010. Citation on page 51.

KOLASCYK, E. Statistical analysis of network data. **SAMSI program on Complex networks. Boston university**, 2013. Citations on pages 46, 47, and 48.

KURANT, M.; MARKOPOULOU, A.; THIRAN, P. Towards unbiased bfs sampling. **IEEE Journal on Selected Areas in Communications**, IEEE, v. 29, n. 9, p. 1799–1809, 2011. Citations on pages 48, 49, and 52.

LATORA, V.; MARCHIORI, M. Efficient behavior of small-world networks. **Physical review letters**, APS, v. 87, n. 19, p. 198701, 2001. Citation on page 35.

LESKOVEC, J.; FALOUTSOS, C. Sampling from large graphs. In: ACM. **Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.], 2006. p. 631–636. Citation on page 52.

LILJEROS, F.; EDLING, C. R.; STANLEY, H. E.; ÅBERG, Y.; AMARAL, L. A. Social networks (communication arising): Sexual contacts and epidemic thresholds. **Nature**, Nature Publishing Group, v. 423, n. 6940, p. 606–606, 2003. Citation on page 40.

LIN, Y. **On the Estimation of Network Metrics in the Presence of Sampling Effects**. Phd Thesis (PhD Thesis), 2015. Citation on page 52.

MAIYA, A. S. **Sampling and inference in complex networks**. Phd Thesis (PhD Thesis) — Stanford University, 2011. Citations on pages 28, 29, and 52.

MAKI, D. P. T. *et al.* **Mathematical models and applications: with emphasis on the social life, and management sciences**. [S.l.], 1973. Citation on page 41.

MARRO, J.; DICKMAN, R. **Nonequilibrium phase transitions in lattice models**. [S.l.]: Cambridge University Press, 2005. Citation on page 40.

MCAULEY, J. J.; LESKOVEC, J. Learning to discover social circles in ego networks. In: **NIPS**. [S.l.: s.n.], 2012. v. 2012, p. 548–56. Citation on page 55.

MISLOVE, A.; MARCON, M.; GUMMADI, K. P.; DRUSCHEL, P.; BHATTACHARJEE, B. Measurement and analysis of online social networks. In: ACM. **Proceedings of the 7th ACM SIGCOMM conference on Internet measurement**. [S.l.], 2007. p. 29–42. Citations on pages 28 and 49.

MITCHELL, M. **Complexity: A guided tour**. [S.l.]: Oxford University Press, 2009. Citation on page 27.

MOLLOY, M.; REED, B. A critical point for random graphs with a given degree sequence. **Random structures & algorithms**, Wiley Online Library, v. 6, n. 2-3, p. 161–180, 1995. Citation on page 33.

MORENO, Y.; NEKOVEE, M.; PACHECO, A. F. Dynamics of rumor spreading in complex networks. **Physical Review E**, APS, v. 69, n. 6, p. 066130, 2004. Citation on page 41.

MORENO, Y.; PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic outbreaks in complex heterogeneous networks. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 26, n. 4, p. 521–529, 2002. Citations on pages 39 and 40.

MURRAY, J. D. **Mathematical biology I: an introduction, Vol. 17 of interdisciplinary applied mathematics**. [S.l.]: Springer, New York, NY, USA,, 2002. Citation on page 40.

NEKOVEE, M.; MORENO, Y.; BIANCONI, G.; MARSILI, M. Theory of rumour spreading in complex social networks. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 374, n. 1, p. 457–470, 2007. Citation on page 41.

NEWMAN, M. **Networks: an introduction**. [S.l.]: Oxford university press, 2010. Citations on pages 27 and 28.

NEWMAN, M. E. Assortative mixing in networks. **Physical review letters**, APS, v. 89, n. 20, p. 208701, 2002. Citations on pages 33, 51, and 58.

NEWMAN, M. E.; STROGATZ, S. H.; WATTS, D. J. Random graphs with arbitrary degree distributions and their applications. **Physical review E**, APS, v. 64, n. 2, p. 026118, 2001. Citation on page 38.

ONODY, R. N.; CASTRO, P. A. de. Nonlinear barabási–albert network. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 336, n. 3, p. 491–502, 2004. Citation on page 37.

PASTOR-SATORRAS, R.; CASTELLANO, C.; MIEGHEM, P. V.; VESPIGNANI, A. Epidemic processes in complex networks. **Reviews of Modern Physics**, APS, v. 87, n. 3, p. 925, 2015. Citations on pages 29 and 41.

PASTOR-SATORRAS, R.; VESPIGNANI, A. Epidemic spreading in scale-free networks. **Physical review letters**, APS, v. 86, n. 14, p. 3200, 2001. Citation on page 40.

RESTREPO, J. G.; OTT, E.; HUNT, B. R. Onset of synchronization in large networks of coupled oscillators. **Physical Review E**, APS, v. 71, n. 3, p. 036151, 2005. Citation on page 35.

____. Approximating the largest eigenvalue of network adjacency matrices. **Physical Review E**, APS, v. 76, n. 5, p. 056119, 2007. Citation on page 36.

RIBEIRO, B.; TOWSLEY, D. Estimating and sampling graphs with multidimensional random walks. In: ACM. **Proceedings of the 10th ACM SIGCOMM conference on Internet measurement**. [S.l.], 2010. p. 390–403. Citation on page 50.

ROSVALL, M.; TRUSINA, A.; MINNHAGEN, P.; SNEPPEN, K. Networks and cities: An information perspective. **Physical Review Letters**, APS, v. 94, n. 2, p. 028701, 2005. Citation on page 35.

ROUGHAN, M.; TUKE, J.; PARSONAGE, E. Estimating the parameters of the waxman random graph. **arXiv preprint arXiv:1506.07974**, 2015. Citation on page 37.

SEBER, G. A.; LEE, A. J. **Linear regression analysis**. [S.l.]: John Wiley & Sons, 2012. Citation on page 43.

SEIDMAN, S. B. Network structure and minimum degree. **Social networks**, Elsevier, v. 5, n. 3, p. 269–287, 1983. Citation on page 34.

SIMAS, A. B.; BARRETO-SOUZA, W.; ROCHA, A. V. Improved estimators for a general class of beta regression models. **Computational Statistics & Data Analysis**, Elsevier, v. 54, n. 2, p. 348–366, 2010. Citation on page 45.

STUMPF, M. P.; WIUF, C.; MAY, R. M. Subnets of scale-free networks are not scale-free: sampling properties of networks. **Proceedings of the National Academy of Sciences of the United States of America**, National Acad Sciences, v. 102, n. 12, p. 4221–4224, 2005. Citations on pages 28 and 52.

STUTZBACH, D.; REJAIE, R.; DUFFIELD, N.; SEN, S.; WILLINGER, W. On unbiased sampling for unstructured peer-to-peer networks. **IEEE/ACM Transactions on Networking (TON)**, IEEE Press, v. 17, n. 2, p. 377–390, 2009. Citation on page 50.

WANG, B.; TANG, H.; GUO, C.; XIU, Z. Entropy optimization of scale-free networks' robustness to random failures. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 363, n. 2, p. 591–596, 2006. Citation on page 33.

WANG, Y.; CHAKRABARTI, D.; WANG, C.; FALOUTSOS, C. Epidemic spreading in real networks: An eigenvalue viewpoint. In: IEEE. **Reliable Distributed Systems, 2003. Proceedings. 22nd International Symposium on**. [S.l.], 2003. p. 25–34. Citations on pages 35 and 40.

WANG, Z.; BAUCH, C. T.; BHATTACHARYYA, S.; D'ONOFRIO, A.; MANFREDI, P.; PERC, M.; PERRA, N.; SALATHÉ, M.; ZHAO, D. Statistical physics of vaccination. **Physics Reports**, Elsevier, 2016. Citation on page 28.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citations on pages 34, 36, and 51.

WAXMAN, B. M. Routing of multipoint connections. **IEEE journal on selected areas in communications**, IEEE, v. 6, n. 9, p. 1617–1622, 1988. Citation on page 37.

WILSON, C.; BOE, B.; SALA, A.; PUTTASWAMY, K. P.; ZHAO, B. Y. User interactions in social networks and their implications. In: ACM. **Proceedings of the 4th ACM European conference on Computer systems**. [S.l.], 2009. p. 205–218. Citation on page 49.

XULVI-BRUNET, R.; SOKOLOV, I. Reshuffling scale-free networks: From random to assortative. **Physical Review E**, APS, v. 70, n. 6, p. 066102, 2004. Citation on page 38.

ZUMSTEIN, P. Comparison of spectral methods through the adjacency matrix and the laplacian of a graph. **TH Diploma, ETH Zürich**, 2005. Citation on page 36.

APPENDIX

A

DISCRETE FRÉCHET DISTANCE

The discrete Fréchet distance is a measure of similarity between two polygonal curves. A polygonal curve is a continuous piecewise linear curve composed of line segments.

The discrete Fréchet distance takes into account not only the location of the points, but also their ordering, and is informally defined as the minimum length of a leash required to connect a dog and its owner, as they walk on two separate paths, given that they may vary their speed, but not backtrack along their path (ALT; GODAU, 1995).

Formally, let $P : [0,N]$ be a polygonal curve with N line segments. P can be parametrized with a parameter $a \in [0,N]$ such that $P(a)$ indicates a position in the curve, with $P(0)$ as the first point of the curve and $P(N)$ as the last one. The discrete Fréchet distance (FRÉCHET, 1906) between two polygonal curves $P : [0,N]$ and $Q : [0,M]$ is equal to:

$$\delta_F(P,Q) = \inf_{\alpha,\beta}\{\max_{t\in[0,1]} d(P(\alpha(t)),Q(\beta(t)))\}, \tag{A.1}$$

where $d(a,b)$ the euclidean distance between points a and b and α and β are continuous non decreasing functions with $\alpha(0) = 0$, $\alpha(1) = N$, $\beta(0) = 0$ and $\beta(1) = M$ (ALT; GODAU, 1995).