UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Discovery of cancer genes patterns using super pathways networks topology

Rodrigo Henrique Ramos

Qualificação de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Rodrigo Henrique Ramos

Discovery of cancer genes patterns using super pathways networks topology

Monograph submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – as part of the qualifying exam requisites of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science.

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Adenilso da Silva Simão

USP – São Carlos
January 2022

Rodrigo Henrique Ramos

Descoberta de padrões de genes de câncer usando a topologia das redes de super pathways

Monografia apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, para o Exame de Qualificação, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional.

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Prof. Dr. Adenilso da Silva Simão

USP – São Carlos
Janeiro de 2022

“Há uma força motriz mais poderosa que o vapor, a eletricidade e a energia atômica: a vontade”

(Albert Einstein)

RESUMO

RAMOS, R. H. **Descoberta de padrões de genes de câncer usando a topologia das redes de super pathways**. 2022. 97 p. Monografia (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Câncer é uma doença complexa caracterizada por mutações genéticas que acontecem na célula e levam a um crescimento e divisão descontrolado. A heterogeneidade do câncer, em que pacientes com diagnósticos semelhantes respondem de maneira diferente aos tratamentos e apresentam evoluções distintas, é bem conhecida a nível clínico, mas pouco compreendida no nível molecular. O sequenciamento de DNA de nova geração criou uma grande quantidade de dados genômicos, promovendo o desenvolvimento de vários métodos computacionais que visam descobrir padrões ocultos em dados genômicos do câncer. Muitos métodos buscam genes *drivers* de câncer e conjuntos de genes mutualmente exclusivos. Além dos dados genômicos de câncer, as abordagens computacionais usam redes de proteínas e *pathways* como entrada, mas frequentemente negligenciam dados qualitativos sobre o tumor e informações clínicas. Os *Super Pathways* agrupam mais de 1.800 *pathways* em 26 funções celulares principais. Redes de *Super Pathways* possuem características topológicas que, em conjunto com dados qualitativos e quantitativos de câncer, podem revelar padrões relativos a genes *drivers* e exclusividade mútua. Neste projeto, propomos utilizar características topológicas das Redes de *Super Pathways* e o papel topológico dos genes *drivers* conhecidos para desvendar padrões ocultos em conjuntos de dados genômicos de câncer, buscando principalmente descobrir genes *drivers* e conjuntos de genes mutuamente exclusivos. Para alcançar os objetos, iremos aprofundar nossos conhecimentos sobre métodos computacionais clássicos e novos para entender suas abordagens e criar um procedimento de validação para comparar e testar os resultados. Vamos combinar dados de cinco fontes diferentes para desenvolver nosso método, incluindo dados genômicos de câncer e Redes de *Super Pathways*. Esta tese tem como objetivo investigar se a combinação de dados qualitativos e quantitativos de câncer, em conjunto com características topológicas de *Super Pathways*, pode ser usada para aprimorar métodos computacionais.

Palavras-chave: Bioinformática do Câncer, Genômica do Câncer, Mutações *Driver*, Exclusividade Mútua, Redes Complexas.

ABSTRACT

RAMOS, R. H. **Discovery of cancer genes patterns using super pathways networks topology**. 2022. 97 p. Monografia (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Cancer is a complex disease characterized by genetic mutations that happen in a cell and lead to uncontrolled growth and division. Cancer heterogeneity, in which patients with similar diagnoses respond differently to treatments and show distinct outcomes, is well known at the clinical level, but poorly understood at the molecular level. The next-generation DNA sequencing created a vast amount of genetic data, promoting the development of numerous computational methods that aim to uncover hidden patterns in cancer genomic data. Many methods seek cancer driver genes and mutual exclusively gene sets. In addition to cancer genomic data, computational approaches use protein networks and pathways as input but often neglect qualitative tumor and clinical data. Super Pathways group more than 1,800 pathways into 26 main cellular functions. Super Pathways Networks possess topological features that, in conjunction with qualitative and quantitative data from cancer, can unveil patterns concerning driver genes and mutual exclusivity. In this project, we propose to use topological characteristics of Super Pathways Networks and the topological role of known driver genes to unveil hidden patterns in cancer genomics data sets, mainly seeking to discover driver genes and mutual exclusively gene sets. To achieve the objects, we will deepen our knowledge about classic and new computational methods to understand their approaches and create a validation procedure to compare and test results. We will combine data from five different sources to develop our method, including cancer genomic data and Super Pathways Networks. This thesis aims to investigate whether a combination of qualitative and quantitative data from cancer, with topological features from Super Pathways, can be used to improve computational methods.

Keywords: Cancer Bioinformatics, Cancer Genomics, Driver Mutations, Mutual Exclusivity, Complex Networks.

LIST OF FIGURES

Figure 1 – Three graphs with the same number of vertices. 24
Figure 2 – Direct graphs and edge weight. 25
Figure 3 – Example Graph. 25
Figure 4 – Data Structures used to store graphs. 27
Figure 5 – Degree distribution and scale-free characterization in eight networks. 29
Figure 6 – Degree distribution and scale-free characterization in eight networks using the Power Law Package. 31
Figure 7 – Scale-free characterization of three random networks with increasing density. 31
Figure 8 – Degree Centrality. 33
Figure 9 – Clustering Centrality. 33
Figure 10 – Kcore Centrality. 34
Figure 11 – Eigenvector Centrality. 34
Figure 12 – Leverage Centrality. 35
Figure 13 – Average Neighbors Degree Centrality. 36
Figure 14 – Closeness Centrality. 37
Figure 15 – Betweenness Centrality. 37
Figure 16 – Bridging Centrality. 38
Figure 17 – Participation Centrality. 39
Figure 18 – Measures Correlation. 40
Figure 19 – All centrality measures - Sample1. 41
Figure 20 – Communities. 42
Figure 21 – Degree Assortativity. 43
Figure 22 – Eccentricity and Diameter. 43
Figure 23 – Density and Clustering. 44
Figure 24 – Attack and Resilience. 45
Figure 25 – Viruses STRING: Interactions Types from *Chikungunya's* network. 51
Figure 26 – PPINs Degree distribution and scale-free characterization. 52
Figure 27 – PPINs Assortativity. 53
Figure 28 – PPINs Clustering and Density. 54
Figure 29 – PPINs Modularity. 55
Figure 30 – PPINs Eccentricity and Diameter. 56
Figure 31 – PPINs Attack. 57
Figure 32 – Reactome's Reproduction Pathway: Three types of representation. 60

Figure 33 – Chosen Super Pathways Networks. 66
Figure 34 – Chosen Super Pathways Networks Measures Correlation. 67
Figure 35 – Chosen Super Pathways Networks Measures Distribution. 68
Figure 36 – Chosen Super Pathways Networks Resilience. 69
Figure 37 – Top 10 most frequently genes in different MAFs 73
Figure 38 – Mutations per gene 74
Figure 39 – Mutations per sample 75
Figure 40 – Top 100 genes coverage over 795 samples 75
Figure 41 – Cancer driver genes databases intersection. 76
Figure 42 – Computational methods for the identifications of cancer driver genes. 77
Figure 43 – Mutual Exclusivity Hypotheses 78
Figure 44 – cBioPortal exclusivity for three genes on Breast Invasive Carcinoma (TCGA, Cell 2015) 79
Figure 45 – Programmed Cell Death Hierarchy 86
Figure 46 – Pipeline 88

LIST OF TABLES

Table 1 – Categorical comparison of four network software 46
Table 2 – Time and space comparison of network software with different datasets . . . 47
Table 3 – PPI Databases 50
Table 4 – Super Pathways Network 65
Table 5 – Mutual exclusivity methods: Input and Output 81
Table 6 – Mutual exclusivity methods: Validation 83
Table 7 – Activity Timetable. 90

LIST OF ABBREVIATIONS AND ACRONYMS

DIP	Database of Interacting Proteins
EG	Example Graph
HINT	High-quality INTeractomes
HPRD	Human Protein Reference Database
HUGO	Human Genome Organisation
ICGC	International Cancer Genome Consortium
IMEx	International Molecular Exchange consortium
MAF	Mutation Annotation Format
NCI	National Cancer Institute
NGS	Next-generation sequencing
PPI	Protein-Protein Interaction
PPIN	Protein-Protein Interaction Network
Reactome FI	Reactome Functional Interaction Network
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TCGA	The Cancer Genome Atlas

CONTENTS

1 INTRODUCTION . . . 19
1.1 Contextualization . . . 19
1.2 Motivation . . . 20
1.3 Objectives . . . 20
1.4 Expected Results . . . 21
1.5 Document Organization . . . 21

2 COMPLEX NETWORKS . . . 23
2.1 Contextualization . . . 23
2.2 Graph Theory . . . 23
2.2.1 Definitions . . . *24*
2.2.2 Topology and Data Structures . . . *26*
2.3 Topological Characterization . . . 28
2.3.1 Scale-Free Property . . . *28*
2.3.2 Centrality Measures . . . *32*
2.3.2.1 Degree . . . 32
2.3.2.2 Clustering . . . 32
2.3.2.3 Kcore . . . 33
2.3.2.4 Eigenvector . . . 34
2.3.2.5 Leverage . . . 35
2.3.2.6 Average Neighbors Degree . . . 36
2.3.2.7 Closeness . . . 36
2.3.2.8 Betweenness . . . 37
2.3.2.9 Bridging . . . 38
2.3.2.10 Participation . . . 38
2.3.2.11 Correlation and Overview . . . 39
2.3.3 Global Network Measures . . . *42*
2.3.3.1 Communities . . . 42
2.3.3.2 Assortativity . . . 42
2.3.3.3 Small World and Diameter - Eccentricity . . . 43
2.3.3.4 Density and Clustering . . . 44
2.3.3.5 Attack and Resilience . . . 44
2.4 Software and Libraries . . . 46

2.5 Final Considerations . . . 48

3 PROTEIN-PROTEIN INTERACTION NETWORKS . . . 49
3.1 Contextualization . . . 49
3.2 PPIN Databases . . . 50
3.3 Protein-Protein Interaction Networks: Topological Characterization 52
3.4 Final Considerations . . . 58

4 PATHWAYS . . . 59
4.1 Contextualization . . . 59
4.2 Pathways Databases and Representation . . . 59
4.3 Pathways Analyses . . . 61
4.4 Super Pathways Networks . . . 64
4.4.1 Topological differences . . . *66*
4.5 Final Considerations . . . 69

5 CANCER GENOMICS . . . 71
5.1 Contextualization . . . 71
5.1.1 Mutation Annotation Format (MAF) . . . *71*
5.2 Heterogeneity: The long tail phenomenon . . . 73
5.3 Cancer Driver Genes . . . 76
5.3.1 Computational Methods . . . *76*
5.4 Mutual Exclusivity . . . 77
5.4.1 Significant Exclusivity . . . *78*
5.4.2 Computational Methods . . . *79*
5.5 Final Considerations . . . 84

6 PROJECT PROPOSAL . . . 85
6.1 Contextualization . . . 85
6.2 Motivation, Hypothesis and Objectives . . . 85
6.2.1 Motivation . . . *85*
6.2.2 Hypothesis . . . *87*
6.2.3 Objectives . . . *87*
6.3 Methodology . . . 87
6.3.1 Activity Timetable . . . *89*
6.3.2 Expected Results . . . *90*
6.3.3 Performed Activities . . . *90*

BIBLIOGRAPHY . . . 93

CHAPTER

1

INTRODUCTION

1.1 Contextualization

Cancer is a complex disease characterized by genetic mutations that happen in a cell and lead to uncontrolled growth and division (HANAHAN; WEINBERG, 2000). Cancer patients with similar diagnoses respond differently to treatments and show distinct outcomes. Albeit known at the clinical level, the heterogeneity is poorly understood at the molecular level (ALLISON; SLEDGE, 2014). With the advent of next-generation sequencing (NGS) technologies, a significant volume of DNA sequencing has been generated (DEMKOW; PLOSKI, 2015). Several databases such as "The Cancer Genome Atlas" (TCGA) and "International Cancer Genome Consortium" (ICGC) make available NGS data. Researchers have widely used data sets with cancer mutation information to study mutations in cancer, genomic instability, and tumor evolution. Such studies utilize computational methods that load and analyze NGS data.

Cancer tumors have many mutations, but only a small portion contributes to oncogenesis and tumor progression. Therefore, cancer mutations are categorized as passenger or driver. Contrary to passenger mutations, driver mutations have a direct impact on oncogenesis, conferring a growth advantage to the cell (STRATTON; CAMPBELL; FUTREAL, 2009). The study of cancer driver genes contributes to understanding the origin and development of the disease. Considering the high heterogenesis in cancer data, the discovery of driver genes is an open challenge in cancer genomics, often explored with computational aid. Cisowski and Bergo (2017) indicates that is intuitive to think the more mutations a tumor has, the faster it progresses. However, large-scale genomics studies show otherwise: driver oncogenes often are mutually exclusive. The authors also say that although this phenomenon is not entirely understood, novel reports indicate mutual exclusivity may be associated with tumor type, pathways, and interactions between drivers' genes. Ding *et al.* (2020) explore how mutual exclusive and co-occurring genes in six types of cancer can significantly divide samples into groups based on clinical data as age, gender, histological type, and pathologic stage.

Protein networks and pathways play an important role in computational methods that aim to discover cancer drivers genes, driver pathways, and mutually exclusive genes. Pathways are sets of genes responsible for the emergence of specific biological functions, thus bringing functional meaning to NGS data, while protein networks represent their interactions, enabling the methods to topological analyze the data (DENG *et al.*, 2019; DING *et al.*, 2020).

1.2 Motivation

A challenge in cancer genomics is the discovery of patterns that can explain the initiation and evolution of the disease, thus enabling personalized therapies. Computational methods have helped in this task, such as searching for cancer driver genes and mutually exclusive gene sets. These methods have evolved over time, improving previous approaches and adding information from different databases. Since many exclusive genes are also drivers, computational methods have used driver information to find mutually exclusive genes and mutually exclusive information to find drivers.

Although there are plenty of computational methods, they often do not consider qualitative data from patients and tumors, and the methods that use networks or pathways frequently only look for neighbors or restrict the topological analyses to degree or shortest path.

The Reactome's Super Pathways are a new concept, since they group more than 1,800 pathways into 26 main cellular functions. The grouping of pathways follows a tree-like structure, where smaller and very specific pathways are leaves, and the 26 Super Pathways are roots. Using Reactome's protein network, it is possible to model the Super Pathways and their sub-pathways as networks. This pathway network hierarchy enables new topological analyses that, in conjunction with quantitative and qualitative NGS data, can be used to uncover patterns in cancer, leading to advances in knowledge about driver genes and mutual exclusivity.

1.3 Objectives

This project's general objective is to discover hidden patterns in cancer genomics data sets using topological characteristics of Super Pathways Networks and the topological role of known driver genes. The discovery of these patterns leads to two specific objectives:

- Finding driver pathways through mutual exclusivity combining quantitative and qualitative data from NGS data with Super Pathways Network topology and driver genes.

- Finding new drivers genes by searching for topological similarities in Super Pathways Networks enriched with driver information.

1.4 Expected Results

We expect the findings of this project contributes to the use of topological information from PPINs to the advancements in Cancer Genomics. We also expect this project to keep collaborating with Barretos Cancer Hospital, which has already resulted in co-authoring six publications with the research group of which the Ph.D. candidate participates.

We intend to continue publishing and participating on two Brazilian bioinformatics conferences: *Simpósio Brasileiro de Computação Aplicada à Saúde* (SBCAS) and *Brazilian Symposium on Bioinformatics* (BSB). In order to further disseminate the outcomes of this Ph.D. project, we also seek to publish the final results in scientific journals.

1.5 Document Organization

This document is organized as follows. In Chapter 2 we present fundamentals about Complex Networks and topological characterization. Chapter 2 offers a background for the network concepts presented in the other chapters and in this thesis proposal. Chapter 3 is an overview of Protein-Protein Interaction Networks, in which we present several topological analyses. In Chapter 4 we present the concept of pathways and how they can be modeled as networks. Chapter 5 explains concepts associated with cancer genomics and computational methods used to discover drivers and mutually exclusive genes. Finally, in Chapter 5 we introduce this thesis proposal, presenting the hypothesis, objects, and methodology. We also present the work plan and some performed activities.

CHAPTER

2

COMPLEX NETWORKS

2.1 Contextualization

The world is full of intricate systems, such as: internet, society, economy, cellular functioning, telecommunications, and others. These systems are call Complex Systems because they are composed of many small parts that interact and cause the emergence of something bigger than the sum of its parts (BARABÁSI, 2015). For example, essential biological functions emerge from molecular interactions, and to understand these functions System Biology studies the associated interactions in different levels of abstraction. (PORRAS, 2016). Each Complex System has on its background a network of interaction between components. For this reason, Network Science is an evolving field used to study Complex Systems (BARABÁSI, 2015).

In this chapter, we discuss fundamentals, approaches, programming tools, and libraries commonly used to study Complex Networks to provide a foundation for the analyses proposed in this project. In Section 2.2 we give a brief introduction on Graph Theory, which is the scaffold behind network science. In Section 2.2.1 we present common terms and definitions used in this work. In Section 2.2.2 we introduce the concept of network topology and data structures used to store network on computer memory. Section 2.3 presents how networks can be studied using topological characterization. Section 2.3.1 explains the importance of degree distribution and the scale-free property. Section 2.3.2 presents ten centrality measures used to characterize nodes, and in Section 2.3.3 we present five global measures used to characterize the whole network. Section 2.4 presents a brief overview of software and libraries used in network characterization and visualization.

2.2 Graph Theory

A graph is a mathematical model that represents relations between objects. Its creation dates back to 1735, when Leonard Euler solved the Königsberg Bridges problem by modeling

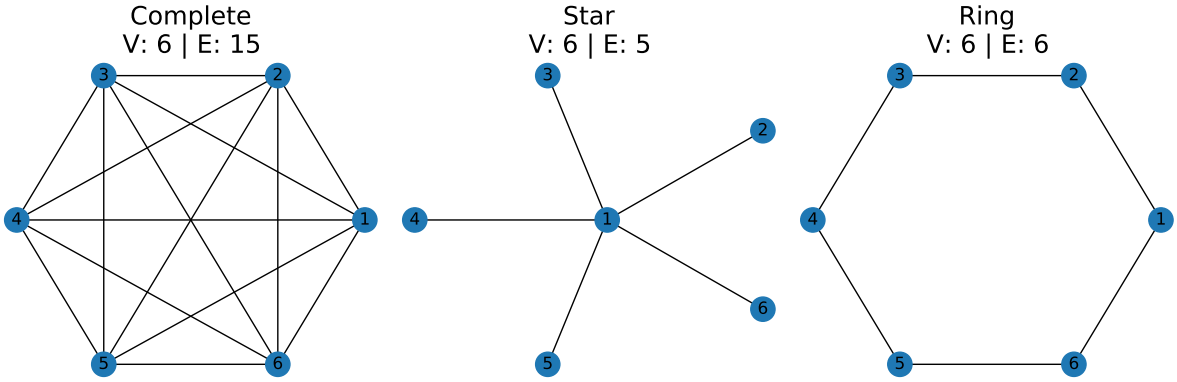
the land areas as vertices and the bridges as edges. Barabási (2015) states that graph theory is the scaffold behind network science, and the terms graph and network are used interchangeably in the scientific literature, with a subtle distinction. Networks commonly refer to real-world systems and use the terms node to represent objects and link to represent the relations, while graph is a the mathematical model and use the terms vertices and edge to represent the objects and relations.

This section presents some concepts and terminologies associated with graphs to provide a background for understanding the following chapters. Since Graph Theory is a vast field, we provide a short introduction only to subjects related to topics present in this thesis proposal.

2.2.1 Definitions

A **graph** G is a set V of **vertices** and a collection E of pairs of vertices from V, called **edges**, and thus can be defined as $G = (V, E)$. Figure 1 shows three graphs with the same number of vertices (V) and a different number of edges (E). The first is a complete graph, where every vertice is interconnected, the second is a star graph, and the third is a ring graph.

Figure 1 – Three graphs with the same number of vertices.

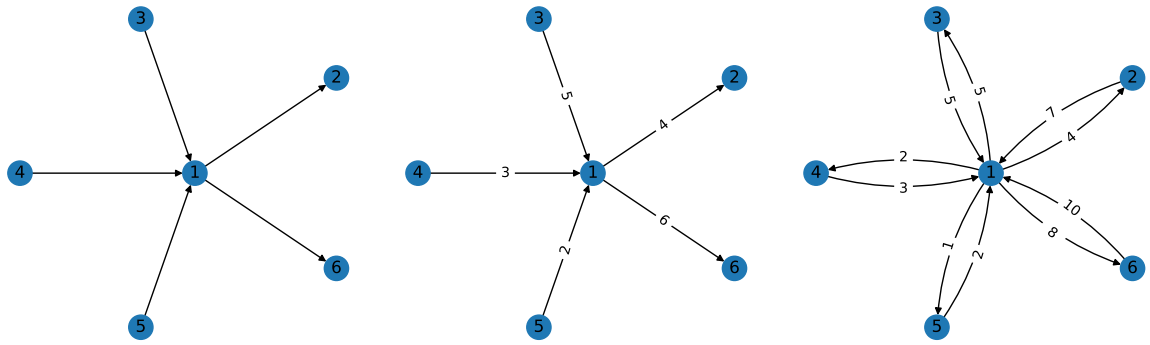


Source: Elaborated by the author.

An important aspect of the graph is the type of information the edges carry. In Figure 1, the graphs have **undirect edges**, meaning the relation goes both ways. When the information flows only in one direction, the graph has **direct edges**, generally represented as arrows. Independent of being direct or undirect, the edge can also bear information about the relation, often called **edge weight**. The edge weight can be, for example, the highway name and distance between two cities. The edge weight between two vertices can vary in direct graphs. For example, the total of flights in a month from city A to B can be different from the flights from B to A. Figure 2 shows a direct graph, a direct graph with edge weight, and a direct graph with different edges weights.

Figures 1 and 2 offers a visual representation of the graphs. A graph can be formally represented as an **adjacency matrix**. An adjacency matrix A is composed of V rows and V

Figure 2 – Direct graphs and edge weight.

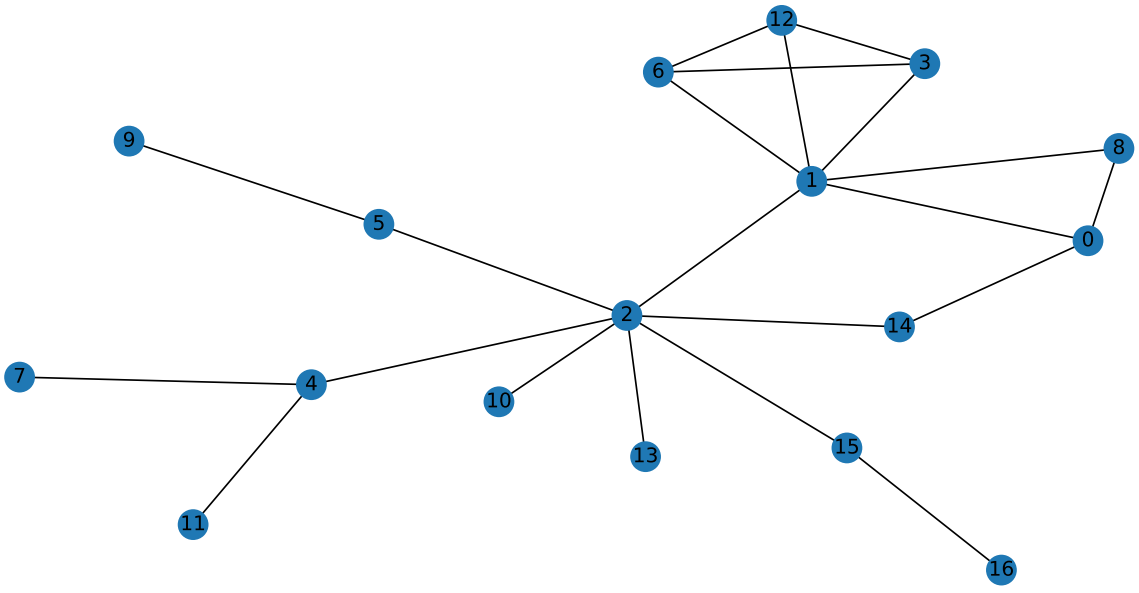


Source: Elaborated by the author.

columns. The presence of a edge between a vertice at row i and a vertice at column j is noted as $A_{ij} = 1$, and the absence of a edge by $A_{ij} = 0$.

We use the Example Graph (EG) in Figure 3 to further define terms associated with Graph Theory. These definitions are based on the work of Goodrich, Tamassia and Goldwasser (2014).

Figure 3 – Example Graph.



Source: Elaborated by the author.

Vertices that share an edge are **neighbors** and the **degree** of a vertice refers to its number of neighbors. Vertice 4 has a degree of three since its neighbors are vertices 2, 7, and 11. In direct graph, the degree is measured as **in-degree**, the number of incomes edges, and **out-degree**, the number of outgoing edges. The vertices with the highest degree in the graph are called **hubs**.

In Figure 3, the average degree is 2.4, since vertice 2 has a degree of seven and vertice 1 has a degree of six, these two vertices are the EG's hubs.

The two hubs have a similar degree but different **neighborhoods**. The vertice 1 neighbors are more interconnected than vertice 2 neighbors (Section 2.3.2.2 discuss the clustering centrality that captures this behavior). The high degree and no interconnection between vertice 2 neighbors is similar to a **star graph**. The high interconnections among vertice 1 neighbors create **cliques**, subsets of completely interconnected vertices. Excluding 2-cliques (edges), the EG have five 3-cliques: {0, 1, 8}, {1, 3, 6}, {1, 3, 12}, {1, 6, 12}, {3, 6, 12}, and a 4-clique: {1, 3, 6, 12}.

A **walk** in a graph is a connected sequence of edges and vertices. The walk allows the repetition of edges and vertices, e.g. $\{2 \rightarrow 1, 1 \rightarrow 3, 3 \rightarrow 1, 1 \rightarrow 6, 6 \rightarrow 1, 1 \rightarrow 12, 12 \rightarrow 1\}$. **Trail** is a walk in which no edge is repeated, e.g. $\{2 \rightarrow 1, 1 \rightarrow 3, 3 \rightarrow 12, 12 \rightarrow 1, 1 \rightarrow 6\}$. Vertice 1 is accessed three times, but no edge is repeated. In a **path** no vertices neither edges can repeat, e.g. $\{2 \rightarrow 1, 1 \rightarrow 3, 3 \rightarrow 12, 12 \rightarrow 6\}$. A **cycle** is a path that end in the starting vertice, e.g. $\{2 \rightarrow 1, 1 \rightarrow 0, 0 \rightarrow 14, 14 \rightarrow 2\}$. The **shortest path** is a path between two vertices in which the sum of edges weight is minimal. EG does not have edges weight, thus we assume all weights are 1. The shortest path between vertices 2 and 8 is $\{2 \rightarrow 1, 1 \rightarrow 8\}$. Albeit other paths start in 2 and end in 8, the sum of weights are greater.

A **subgraph** is a graph such that its vertices and edges are a subset of another graph. In Figure 1 the Star and Ring graphs are subgraphs from the Complete graph, once their vertices and edges exist within the Complete graph. Cliques are complete connected subgraphs inside a graph. An **induced subgraph** is a subset of vertices and edges that connect these vertices. In Figure 3 a induced subgraph containing the vertices {1,3,6,12} is a 4-clique. **Connected components** are maximal connected subgraphs. EG is composed of one connected component since every vertice is **reachable** by any other vertice. A induced subgraph containing the vertices {2,3,4,5,6,10,13,12} creates two connected components: a triangle and a star.

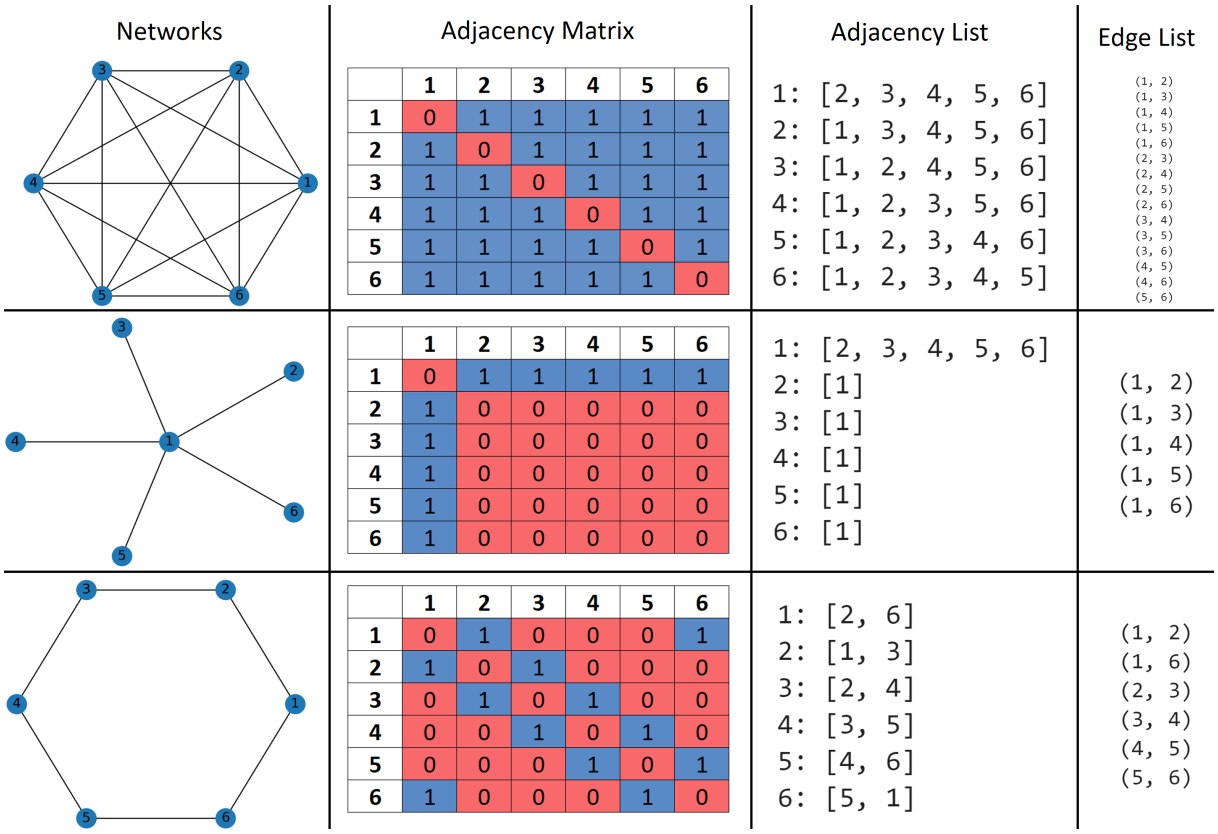
2.2.2 Topology and Data Structures

The graph topology concerns the organization of the vertices and edges within the graph. Figure 1 shows three graphs with the same number of vertices and different topologies. The first topology is a **complete graph**, the second is a **star topology**, and the third is a **ring topology**. The Star and Ring graphs have almost the same number of edges yet have a different topology. As seen in Figure 2, the edge type and weight also change the graph topology. The study of these differences is the main focus of the next chapter.

In computer science, the graphs must be stored in memory. There are three common data structures to store graphs: **adjacency matrix**, **adjacency list**, and **edge list**. Each one has pros and cons concerning the complexity of time and space (GOODRICH; TAMASSIA;

GOLDWASSER, 2014). Figure 4 shows a visual representation of these three data structures applied on the graphs of Figure 1.

Figure 4 – Data Structures used to store graphs.



	1	2	3	4	5	6
1	0	1	1	1	1	1
2	1	0	1	1	1	1
3	1	1	0	1	1	1
4	1	1	1	0	1	1
5	1	1	1	1	0	1
6	1	1	1	1	1	0

	1	2	3	4	5	6
1	0	1	1	1	1	1
2	1	0	0	0	0	0
3	1	0	0	0	0	0
4	1	0	0	0	0	0
5	1	0	0	0	0	0
6	1	0	0	0	0	0

	1	2	3	4	5	6
1	0	1	0	0	0	1
2	1	0	1	0	0	0
3	0	1	0	1	0	0
4	0	0	1	0	1	0
5	0	0	0	1	0	1
6	1	0	0	0	1	0

Source: Elaborated by the author.

The adjacency matrix always use $O(|V|^2)$ memory, even if the graph has few edges. In undirect graphs the adjacency matrix is symmetric, turning halve of the matrix redundant. Barabási (2015) presents a undirected actor graph with 702,388 vertices and 29,397,908 edges. To store this graph an adjacency matrix will create 493,348,902,544 spaces to indicate the presence or absence of edges between all pairs of vertices. 99.994% of the space reserved will be used to indicate the absence of edges. Although useful in mathematical models, the adjacency matrix in computer science is rarely an efficient option in big graphs regarding space complexity. The adjacency list store information about the relations of each vertice, and not about the absence of relation. It offers a balance efficient in time and space complexity, being a good option to store graphs on the primary memory. The edge list only store edges. Its memory efficient, but inefficient in time complexity, being often use as a data structure to save the graph on secondary memory.

2.3 Topological Characterization

Complex Networks differ from "simple" graphs by their non-trivial topology and the ability to represent interconnections found in many real systems (BARABÁSI, 2015). Even though different systems have different networks, many real-world networks share similarities, like being scale-free (BARABÁSI, 2015).

There are many ways to analyze Complex Networks, and since the centrality and global measures are complementary, it is important to use different measurements (OLDHAM *et al.*, 2019). Aiming to provide the means necessary to analyze networks, we present the definition of Scale-Free using six real-world networks and two synthetics. Using two small networks we also describe ten centrality measures, five global network measures, and the concept of network attack and resilience.

2.3.1 Scale-Free Property

A key difference between random networks and networks from real-system is the degree distribution. Real-world networks tend to have a scale-free distribution: few nodes are highly connect (hubs) while most nodes have few neighbors. Thus, the mean and variance do not capture the distribution behavior (BARABÁSI, 2015). Not all real-world networks are scale-free, but the ones that do tend to follow a degree probability similar to a power law:

$$pk \sim k^{-\gamma} \tag{2.1}$$

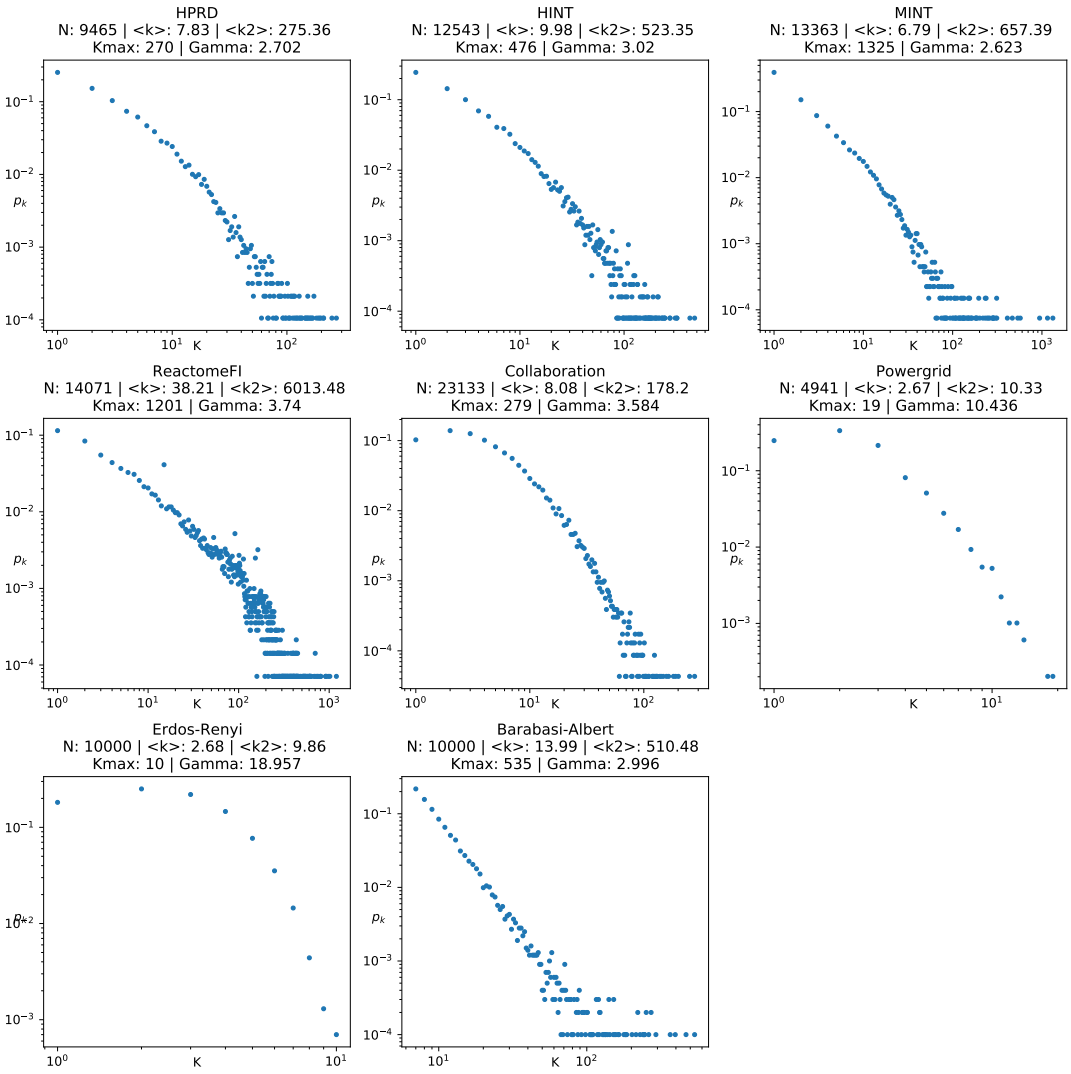
Equation 2.1 captures the phenomenon that is more likely to find small degree nodes than high degree nodes in scale-free networks. To illustrate this property, we use six networks from real-world systems and two synthetics. The first four are Protein-Protein Interactions Networks from *homo-sapiens*, and are further discussed in Section 3.2, since they are used here to exemplify the scale-free property. The next two real-world networks are from scientific collaboration and electrical power grid, both used by Barabási (2015). The last two are synthetic networks. The first represent scale-free networks and the other random networks.

Figure 5 shows the degree distribution probability, in log-log scale, of these eight networks. Bellow the network name, there is the following information: N is the number of nodes; $<k>$ is the average degree; $<k2>$ is the second moment of the degree distribution[1]; *Kmax* is the degree of the biggest hub, and *Gamma* is the exponent from Equation 2.1. To calculate *Gamma* we use the Power Law Package from Alstott, Bullmore and Plenz (2014).

Barabási (2015) states that in scale-free network, as $N \to \infty$, also $<k2> \to \infty$ while $<k>$ stays finite, capturing the scale-free nature of the degree distribution. Additionally, as N increases,

[1] $<k2>$ helps to understand the distribution spread. It is calculated by adding the distribution variance to the square of $<k>$.

Figure 5 – Degree distribution and scale-free characterization in eight networks.



Source: Elaborated by the author.

Gamma should be between two and three. As *Gamma* grows bigger than three, the distribution becomes less similar to scale-free and more similar to a random network. Since the scale-free distribution follows a power law, the degree probability should be close linear in a log-log scale. With some nuances, the five first networks show the expected behavior from scale-free networks. The last network, "Barabasi-Albert", is a random network that follows a probabilistic model that captures the degree distribution found in scale-free networks. This model increases the chances of high-degree nodes receiving new edges, thus, creating hubs while keeping most nodes' degree small (BARABÁSI, 2015). We generate this network so that the value of <*k*> is the average of <*k*> from the five first networks. Even though it is a synthetic network, it shares many similarities with scale-free networks from the real-world. Contrary to "Barabasi-Albert", "Erdos-Renyi" is a random network without preferred attachment. Because of this, even with ten thousand nodes, the biggest degree (Kmax) is ten. We generate this "Erdos-Renyi" network so that the value of <*k*>

is similar to the "Powergrid". The "Erdos-Renyi" network does not have a scale-free distribution: <k> and <$k2$> are finite, and *Gamma* is bigger than eighteen. Although the "Powergrid" network is from the real-world, its behavior is closer to "Erdos-Renyi" than the others. The critical point for this similarity is the absence of hubs. As a "Powergrid" models the electrical wiring of power stations, a physical limitation hinders distant stations' connections, so <k> and <$k2$> are finite.

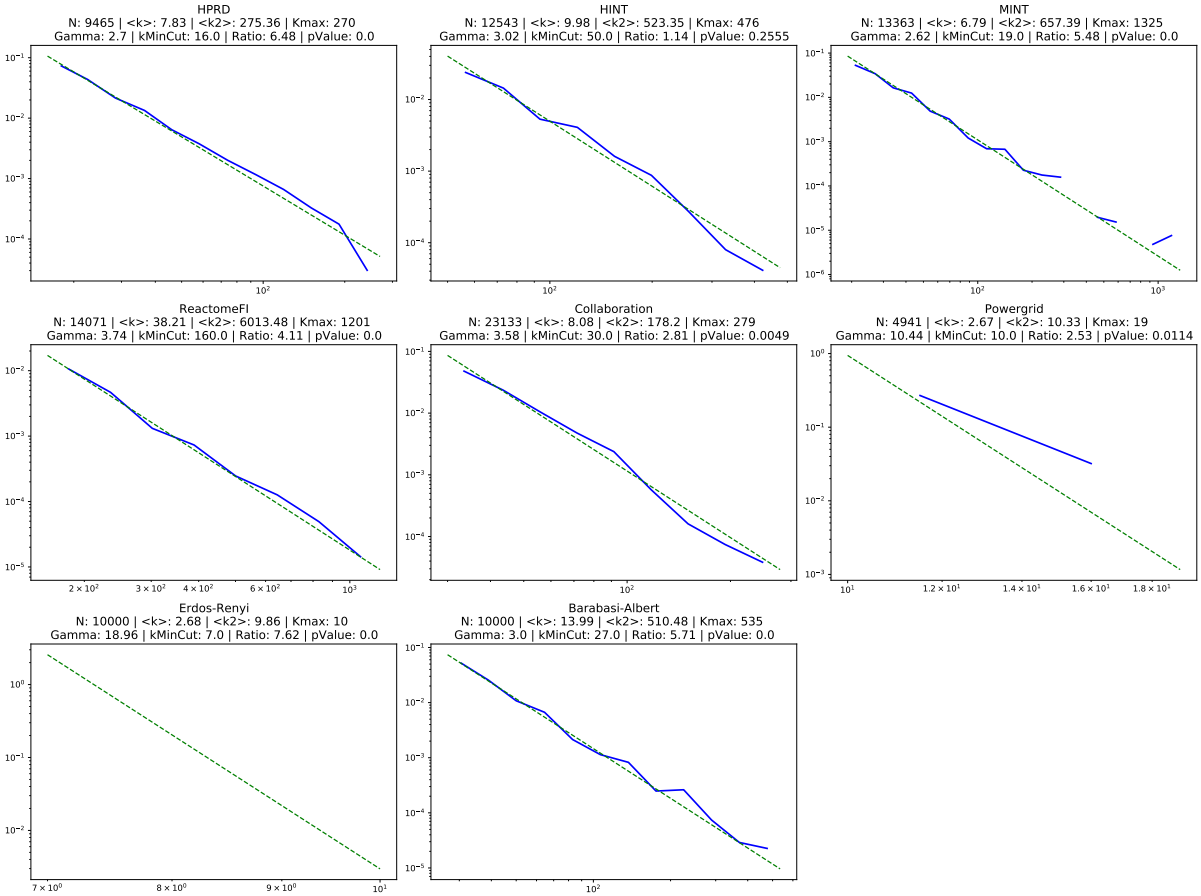
The presence (or absence) of hubs and the degree distribution significantly impacts the network topology. The identification of the *Gamma* exponent helps to characterize the network and to choose a synthetic model that captures its behavior (BARABÁSI, 2015). The difference showed in Figure 5 between "Powergrid" and the other real-world networks, as well as how the two synthetic networks represent the two groups of real networks, exemplifies the importance of the *Gamma* exponent.

Although relevant, the identification of *Gamma* is a complex and often an approximation task (ALSTOTT; BULLMORE; PLENZ, 2014). In most cases this approximation, allied with <k> and <$k2$>, and the distribution plot is enough (BARABÁSI, 2015). To increase our knowledge about the degree distribution is possible to fit it between known distributions, like a power law and an exponential (BARABÁSI, 2015; ALSTOTT; BULLMORE; PLENZ, 2014). In an exponential distribution, the probability of finding high degree nodes decreases exponentially, which is expected from networks with *Gamma* grater than three (random networks). As a general idea, fit the degree distribution is asking if the network is closer to a Barabasi-Albert preferred attachment model or an Erdos-Renyi model. Using the Power Law Package (ALSTOTT; BULLMORE; PLENZ, 2014), Figure 6 shows the degree distribution fitting (blue continuous line) to a power law distribution (dashed green line) for the eight networks used in Figure 5.

For each network in Figure 6 we repeat the N, <k>, <$k2$>, $Kmax$, *Gamma*, and add three new information: $kMinCut$, $Ratio$, and $pValue$. The degree distribution normally does not fit completely in a power law. In these cases it is necessary to ignore small degree nodes and try to fit the distribution after a threshold (BARABÁSI, 2015; ALSTOTT; BULLMORE; PLENZ, 2014). $kMinCut$ represents the minimum degree used in the fitting process. The blue line in "HPRD" starts only after degree sixteen, in "ReactomeFI" its starts only after degree one hundred and sixty neighbors. $Ratio$ is the loglikelihood between the candidate distributions. It is positive if the degree distribution is more likely to be a power law, and negative if it is more likely to be exponential. p-value is the statistical significantly for $Ratio$, 0.0 values means the p-value < 0.05. "HINT" has a $Ratio = 1.14$, pending to the power law side, but the p-value$= 0.26$ does not indicate a significantly stronger fit.

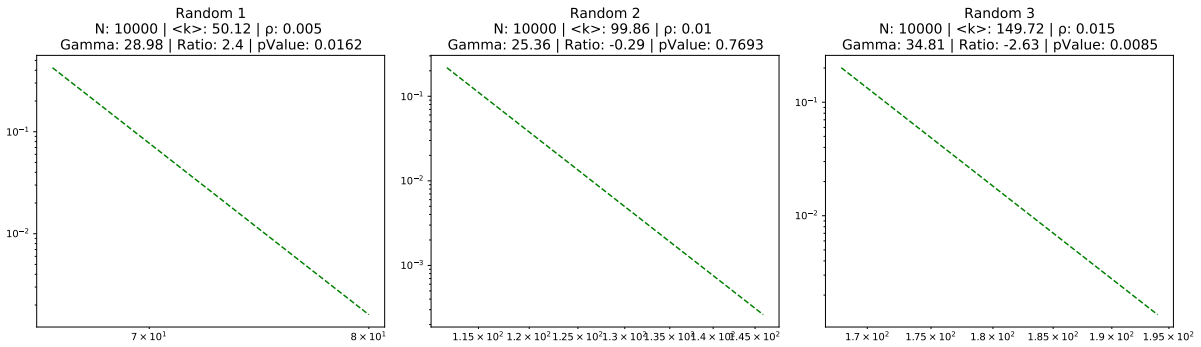
"Powergrid" and "Erdos-Renyi" both have high *Gamma* and finite <$k2$>, yet, they have a $Ratio$ with significant p-value indicating a closer relation to a power law than with an exponential distribution. In Figure 5 these networks also do not show the expected Poisson distribution for random networks. Figure 7 address this issue, and the relation to the network density.

Figure 6 – Degree distribution and scale-free characterization in eight networks using the Power Law Package.



Source: Elaborated by the author using the Python library from Alstott, Bullmore and Plenz (2014)

Figure 7 – Scale-free characterization of three random networks with increasing density.



Source: Elaborated by the author using the Python library from Alstott, Bullmore and Plenz (2014)

Figure 7 presents three random networks with ten thousand nodes created using the Erdos-Renyi model. The networks have an increasing <*k*>, and therefore an increasing density ρ. As the number of edges rises, the *Ratio* and its *pValue* decrease. The problem in fitting "Powergrid" and "Erdos-Renyi" is the small <*k*> and ρ. The *kMinCut* in both networks are closer to <*k*2> than <*k*>. It is important to notice that any network that starts to $\rho \to 1$, and

$<k> \rightarrow (N-1)$ are in the path to become a complete graph, with trivial topology and constant degree distribution.

2.3.2 Centrality Measures

The scale-free property is a topological characterization for the whole network. Centrality measures characterize the network's nodes, assigning values depending on the node's position. Each centrality measure quantifies the nodes within a network by one perspective: flow of information; proximity to all other nodes; neighborhood interconnection; ability to influence (or be influenced) by other nodes, etc. In addition to answering punctual questions, cumulative analysis of different measures helps the understanding of the network.

In this section, we show ten centrality measures. For each one, we present a description explaining how the measure characterizes the nodes and a formal definition using equations. The equations and definition are inspired by the work of Oldham *et al.* (2019). We also plot two sample networks to visually aid the centrality measure impact on nodes. These networks have distinct topologies, and the node size and color, from yellow to red, are defined by the centrality strength. We call the left network S1 and the right network S2.

For equations, we represent the network as an $N \times N$ adjacency matrix A in which the element $A_{ij} = 1$ if nodes i and j are connected and $A_{ij} = 0$ otherwise. We only consider undirected and unweighted networks in this section.

2.3.2.1 Degree

The degree of centrality, Equation 2.2 and Figure 8, is the simplest measure. Defined as the number of edges connected to a node, in other words, the number of neighbors a node has. The degree has a high correlation with other measures, highlighting its importance in the topological analysis of networks (OLDHAM *et al.*, 2019).

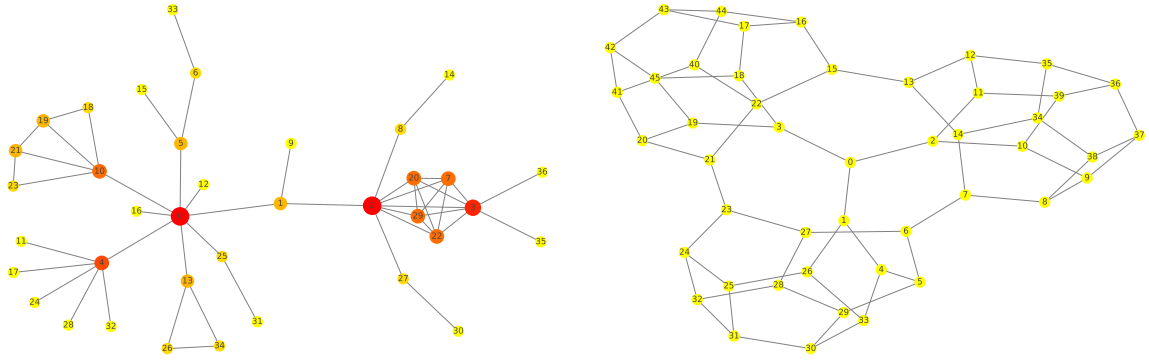
$$DC_i = d_i = \sum_{j \neq i} A_{ij} \tag{2.2}$$

In the S1 network, nodes 0 and 2 have the highest degree: eight. By this centrality measure, these two nodes are equally the most important in the network. Node 3 has a degree of seven, and node 4 has a degree of six. These nodes are the second and the third most important. Nodes 7, 10, 20, 22, and 29 are tied in fourth place. Every node in the S2 network has a degree of 3, meaning that degree centrality considers all nodes equally important.

2.3.2.2 Clustering

The degree clustering, Equation 2.3 and Figure 9, measures the strength in which the neighbors of a node interconnect. For a node i with degree d_i, the clustering centrality is defined

Figure 8 – Degree Centrality.



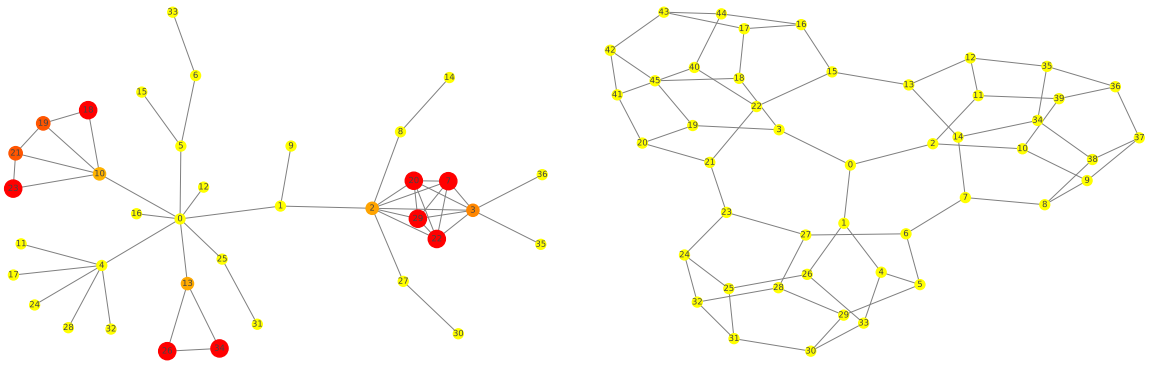
Source: Elaborated by the author.

as:

$$C_i = \frac{2L_i}{d_i(d_i - 1)} \tag{2.3}$$

where L_i is the number of links between the neighbors of node i.

Figure 9 – Clustering Centrality.



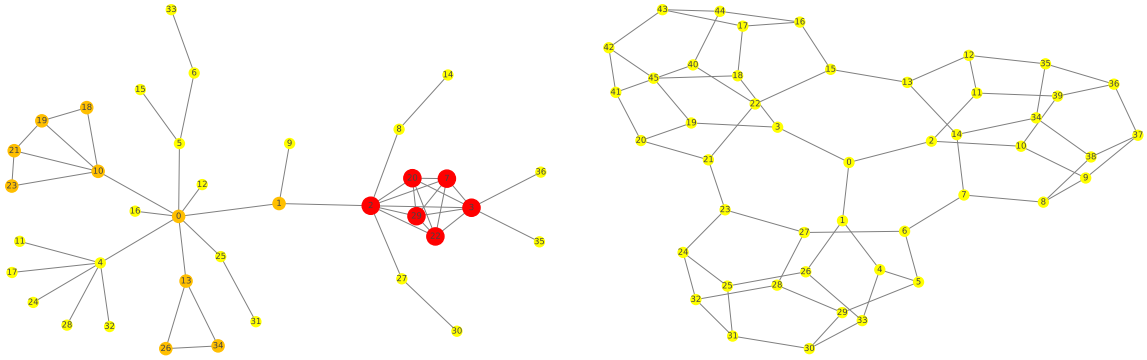
Source: Elaborated by the author.

None of the neighbors in S2 network are directly interconnected, meaning that every node has a clustering of zero and thus are equally important. Contrary to degree centrality, in the S1 network the nodes 0 and 4 are in the group of least important (cluster zero). While nodes 2, 3, 10, 13 have clustering varying between 0.3 to 0.47.

2.3.2.3 Kcore

The kcore of a node, Figure 10, is the minimal degree found in a maximal subgraph that contains nodes with degree k or more. The subgraphs in this case are connected components that remain after the removal of nodes with k or less degree.

Figure 10 – Kcore Centrality.



Source: Elaborated by the author.

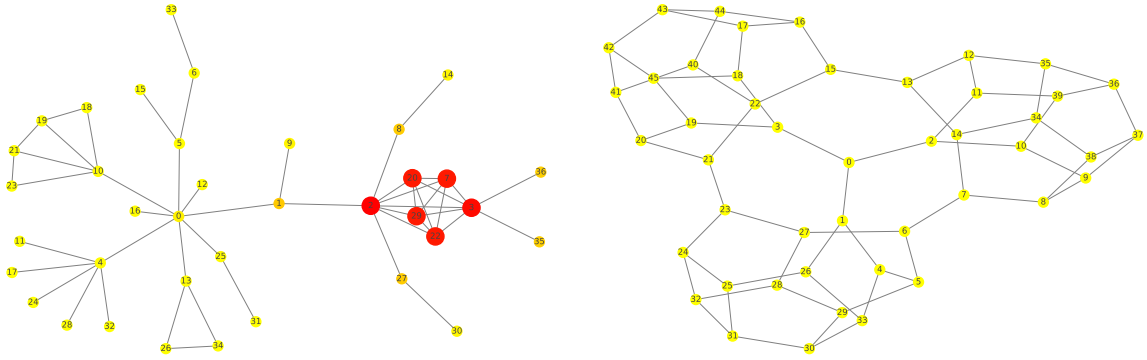
In the S2 network, every node has a kcore of three. Since the degree is always three, removing degrees one and two does not create new connected components, but removing nodes with degree three destroys the network. In the S2 network, nodes 2, 3, 7, 20, 22, and 29 belong to a click, thus having a kcore of 5, even though nodes 2 and 3 have a degree of seven.

2.3.2.4 Eigenvector

Eigenvector centrality, Equation 2.4 and Figure 11, gives a higher score to nodes with high degree and/or that have high degree neighbors. This measure uses the eigenvector, v, associated with the largest eigenvalue λ_1 of the adjacency matrix:

$$EC_i = v_i = \frac{1}{\lambda_1} \sum_j A_{ji} v_j \tag{2.4}$$

Figure 11 – Eigenvector Centrality.



Source: Elaborated by the author.

As the eigenvector centrality depends on the node's degree and neighbors' degree, every node in S2 network has the same value. In the S1 network, nodes at a click of size five receive

the highest values. Nodes 7, 20, 22, and 29 have a score of 0.38, because they have the same degree and are connected with high degree neighbors. Node 2, and 3 have a slightly bigger score of 0.43 and 0.41, since their degree is higher, even having some low degree neighbors. Node 0 has the same degree as node 2, but since it has many low degree neighbors, its score is 0.02.

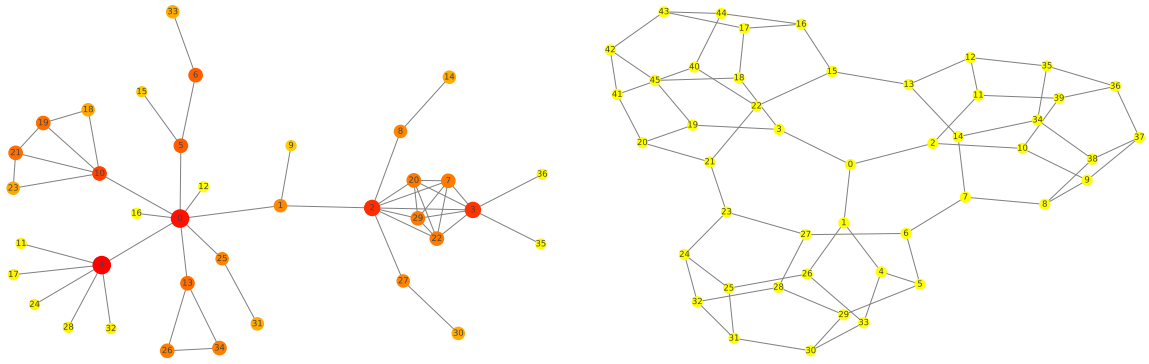
2.3.2.5 Leverage

Leverage centrality, Equation 2.5 and Figure 12, also considers the node's neighbors connections. Leverage identify if a node influences its neighbors or is influenced by them. Leverage centrality assigns a negative value when the node's degree is lower than its neighbors, indicating that they influence it. A positive value means the opposite, and zero indicates a neutral state, which is the case of S2 network. Leverage centrality is defined as

$$LC_i = \frac{1}{d_i} \sum_{j \in \mathcal{N}(i)} \frac{d_i - d_j}{d_i + d_j} \tag{2.5}$$

where $\mathcal{N}(i)$ is the set of node i neighbors and d is the degree.

Figure 12 – Leverage Centrality.



Source: Elaborated by the author.

To maintain the same visual pattern as the other plots, we normalize the centrality values from 0 to 1 only for the plotting of S1 network. The red nodes 4, 0, 2, and 3 have a centrality score of 0.57, 0.48, 0.33, 0.3, meaning that they influence their neighbors. Nodes with shades of orange are close to zero. Node 5 has a score of 0.08, and node 13 has a score of -0.02. Yellow nodes are influenced by their neighbors and have a score tending to -1. The nodes 7, 20, 22, and 29 have a degree of 4 and have neutral leverage (-0.8). This happens because there are neighbors of 2 and 3 that have a higher degree.

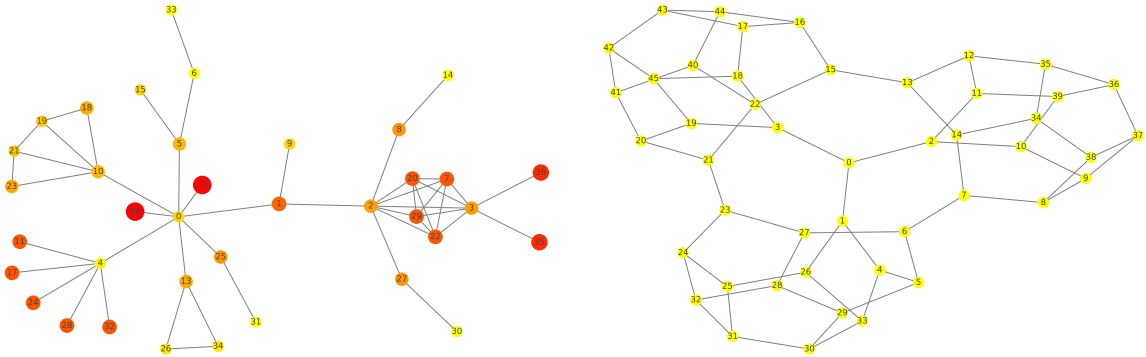
2.3.2.6 Average Neighbors Degree

Average neighbors degree centrality, Equation 2.6 and Figure 13, is a simple measure that sums the neighbors' degree and divides by the node's degree:

$$AvgNC_i = \frac{1}{d_i} \sum_{j \in \mathcal{N}(i)} d_j \tag{2.6}$$

where $\mathcal{N}(i)$ is the set of node i neighbors and d is the degree.

Figure 13 – Average Neighbors Degree Centrality.



Source: Elaborated by the author.

The lack of degree variation on the S2 network makes this centrality assign a score of three to every node. For some nodes at the S1 network we see the opposite score given by Leverage Centrality, like nodes 0, 4, 12, and 16. Albeit both centralities share a strong negative correlation, there are some distinctions. In the previous centrality, nodes 13, 34, and 26 have the same value, but in this measure, they differ.

2.3.2.7 Closeness

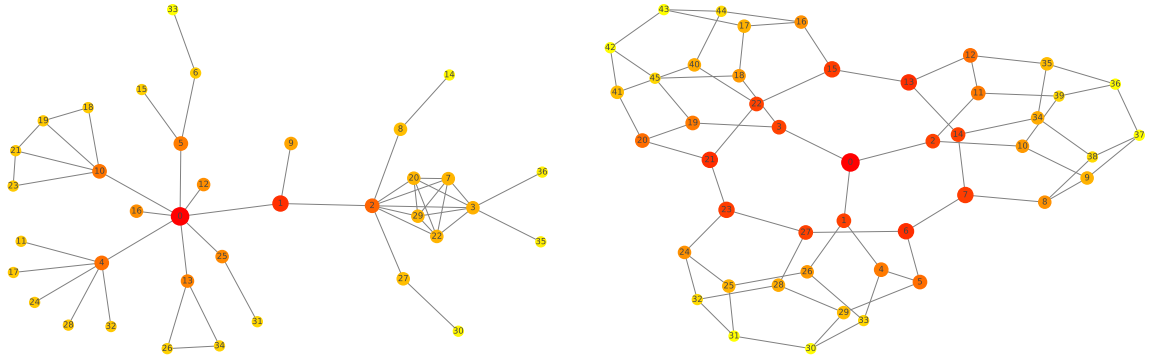
Closeness centrality, Equation 2.7 and Figure 14, gives a high score to nodes that are close to all other nodes in the network. Closeness creates a shortest path matrix from every node to all nodes. Nodes with a small average shortest path are assumed as central.

$$CC_i = \frac{N}{\sum_j l_{ij}} \tag{2.7}$$

where N is the number of nodes in the network, and l_{ij} is the shortest path length between nodes i and j.

Closeness centrality is the first measure to identify differences in S2 network nodes. Node 0 is the most central since from it we can reach any other node with at most five steps. Node 44, which is five steps from node 0, is eight steps from node 33. In the S1 network, the highest value is also node 0. Closeness is sensitive to high degree nodes (OLDHAM *et al.*, 2019)

Figure 14 – Closeness Centrality.



Source: Elaborated by the author.

as hubs normally shorten the network distance and are responsible for the small-world effect (BARABÁSI, 2015). Node 2 and 0 have the same degree, yet its closeness is smaller than node 1, showing that closeness centrality captures nodes' topological centrality more than their degree.

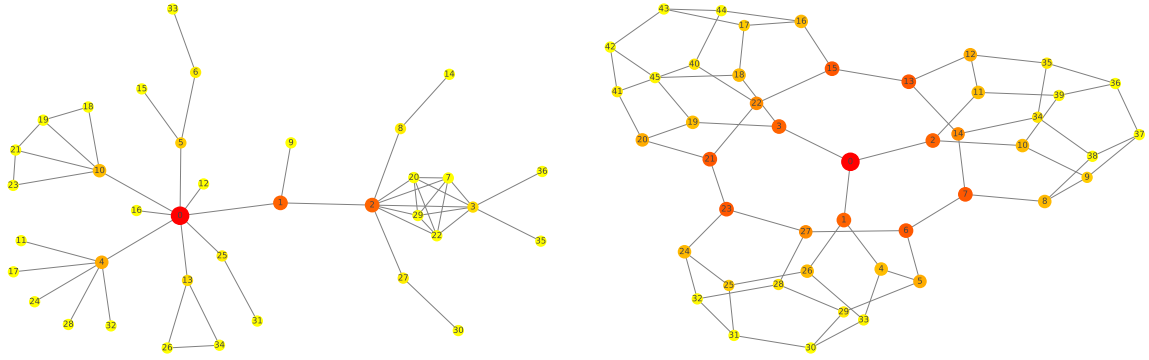
2.3.2.8 Betweenness

Betweenness centrality, Equation 2.8 and Figure 15, also creates a shortest path matrix, but it gives a high score to nodes that appear in other nodes' paths, meaning that this node is frequently used in shortest paths. Betweenness identifies nodes that act as "bottlenecks", concentrating flow of information in the network.

$$BC_i = \sum_{p \neq i, p \neq q, q \neq i} \frac{sp_{pq}(i)}{sp_{pq}} \tag{2.8}$$

where sp_{pq} is number of shortest path between nodes p and q, and $sp_{pq}(i)$ is number of shortest paths between nodes p and q which pass through node i.

Figure 15 – Betweenness Centrality.



Source: Elaborated by the author.

There is a high correlation between this measure and Closeness, since central nodes shorten the distances and thus concentrate flow of information. Betweenness is more prone to outliers than closeness, as can be seen in the S2 network plot which contains fewer red and orange nodes.

2.3.2.9 Bridging

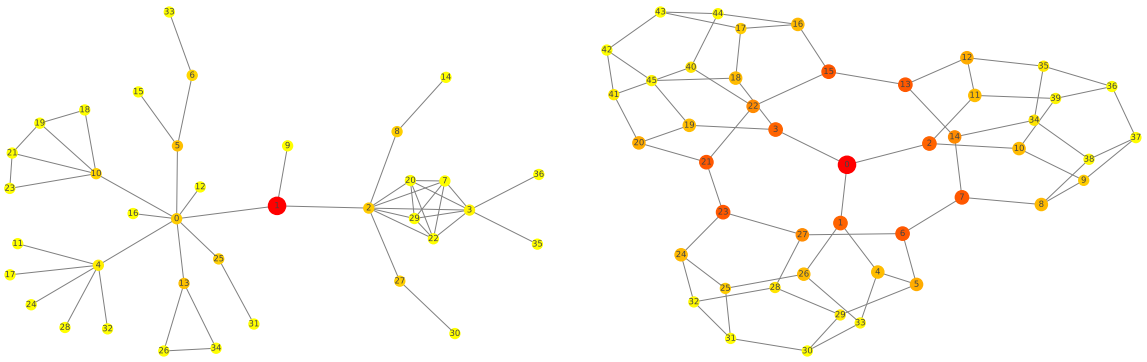
Bridging centrality, Figure 16 and Equations 2.9, and 2.10, aim to modify the betweenness centrality to identify those who connect different communities. The node's score is obtained by multiplying the betweenness centrality of node $_i$ (BC_i) by the bridging coefficient of node $_i$ (Bc_i), Equation 2.9, which is defined as:

$$Bc_i = \frac{d_i^{-1}}{\sum_{j \in \mathcal{N}(i)} d_j^{-1}} \tag{2.9}$$

where $\mathcal{N}(i)$ is the set of node i neighbors and d is the degree. Bc_i quantifies by how much the neighbors have a higher degree.

$$BridC_i = BC_i \, x \, Bc_i \tag{2.10}$$

Figure 16 – Bridging Centrality.



Source: Elaborated by the author.

Bridging Centrality is even more prone to outliers than betweenness. In the S1 network just node 1 have a high score because it has a lower degree than its neighbors and an average betweenness. The constant degree distribution at S2 network makes the score similar to the betweenness.

2.3.2.10 Participation

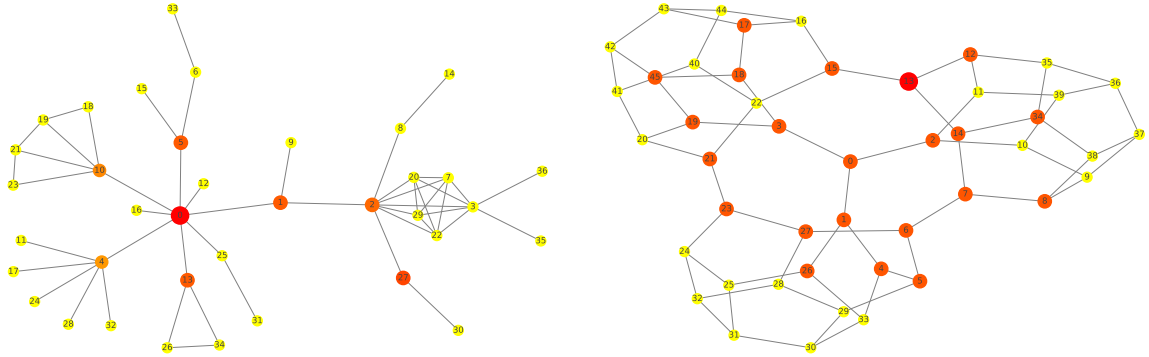
Participation centrality, Equation 2.11 and Figure 17, also works with the notions of communities. Based on the idea that most real-world networks have communities (or modules),

participation centrality gives a high score to nodes that participate in different communities. PC_i is defined as:

$$PC_i = 1 - \sum_{c=i}^{C} \left(\frac{d_i(c)}{d(i)} \right)^2 \tag{2.11}$$

where C is the number of network's communities and $d_i(c)$ is the number of node i neighbors in community c, and $d(i)$ is the node i degree.

Figure 17 – Participation Centrality.



Source: Elaborated by the author.

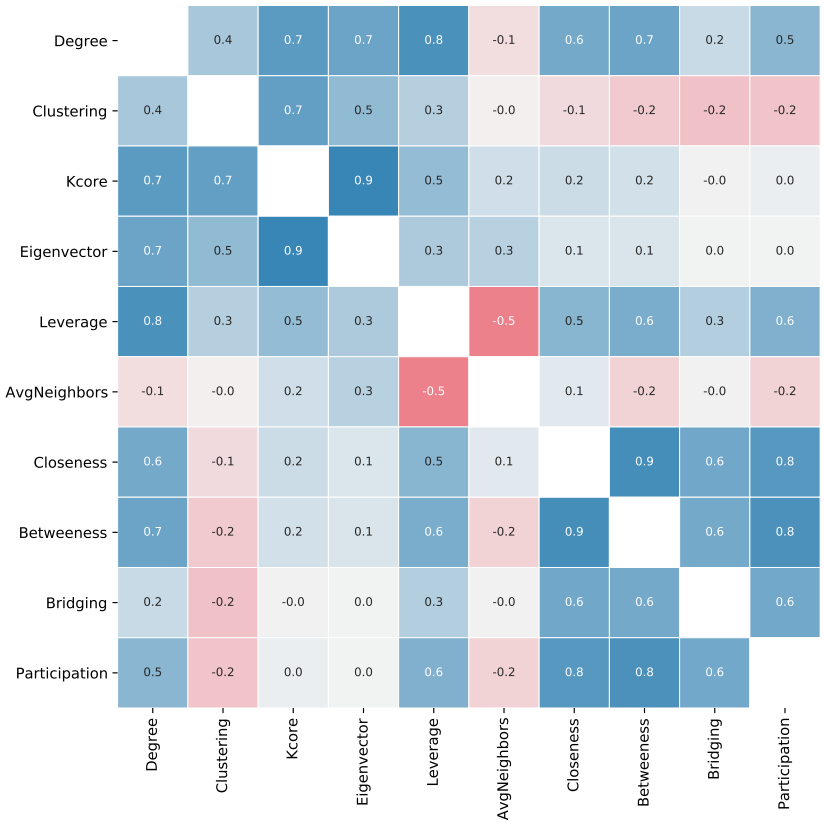
Node 0 is the most prominent in S1 network because it has eight links that participates in five different communities. Node 2 also has eight links, but participates in three communities. Node 1 has three links and participates in two communities, thus having a score similar to node 2. Participation centrality depends on the definition of communities, which is an NP problem (BARABÁSI, 2015). There are many heuristics to identify communities, which one of them may result in different types of communities and therefore different scores in participation centrality. Node 13 in S2 network receives the highest score, even tough there are others nodes that are topological identical. The role of communities in complex networks are further discuss in Section 2.3.3.1.

2.3.2.11 Correlation and Overview

Many centrality measures have a high correlation to each other, especially degree (OLDHAM *et al.*, 2019). Figure 18 shows the Person's correlation between the ten centrality measures used in this section.

Kcore and Eigenvector share a 0.9 correlation. Clustering correlates 0.7 with Kcore and 0.5 with Eigenvector, showing how measures may correlate differently even with a pair of highly correlated measures. It is also important to notice that distinct networks may share different correlations.

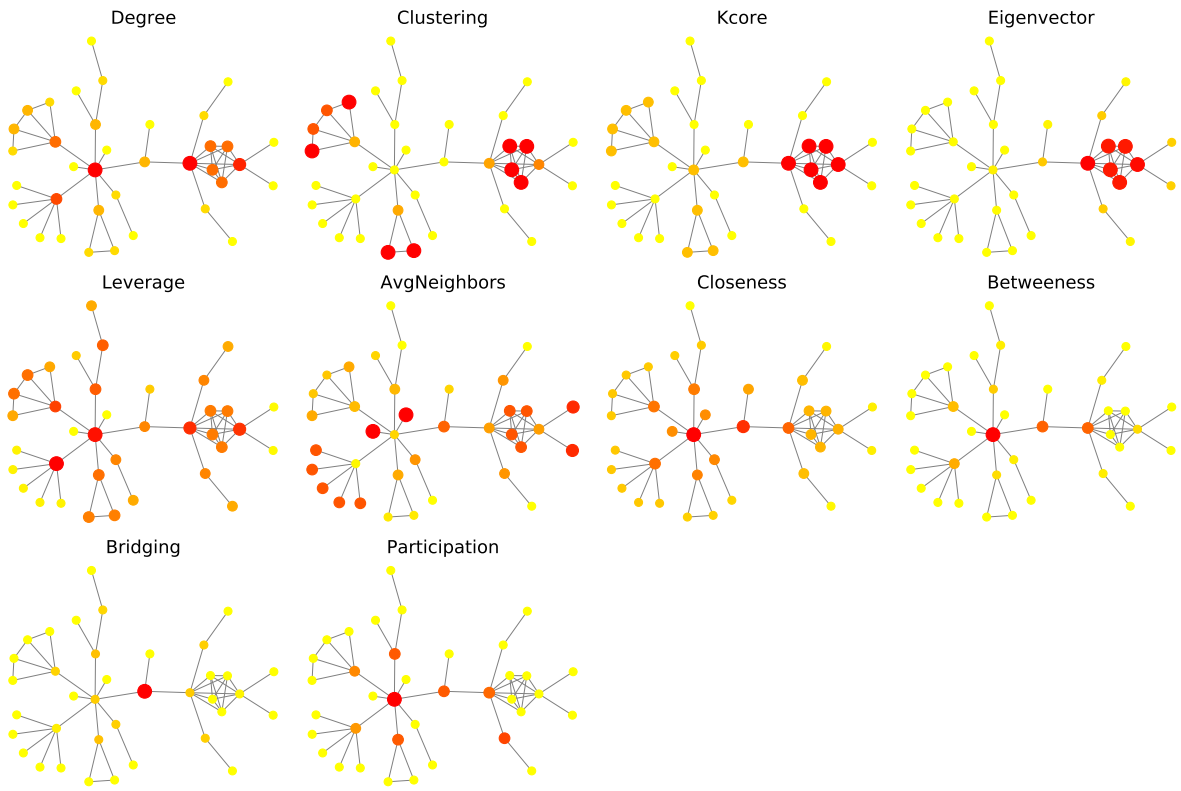
Figure 18 – Measures Correlation.



Source: Elaborated by the author.

Figure 19 visually summarizes the scores given by the centrality measures. There are ten plots of the S1 network, and each one of them is different, showing that each network measure "look" the network by one perspective.

Figure 19 – All centrality measures - Sample1.



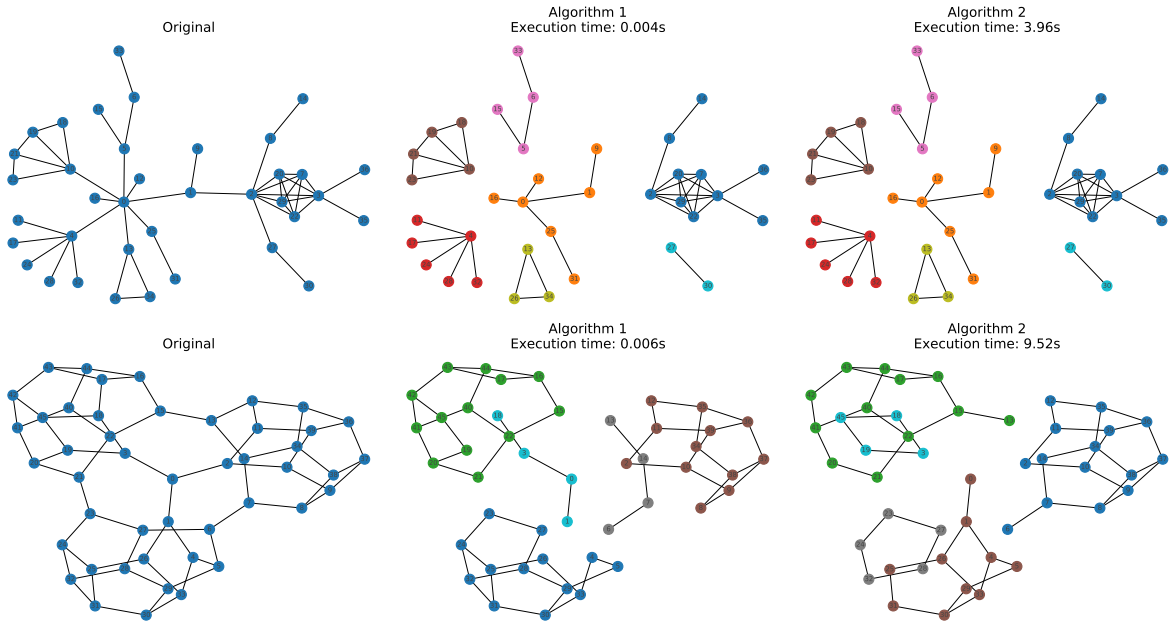
Source: Elaborated by the author.

2.3.3 Global Network Measures

2.3.3.1 Communities

Network communities, or modules, are sub-networks within the network that interconnect in a way that forms clusters (RADICCHI *et al.*, 2004). The identification of communities, albeit relevant to understanding the network structure, is an NP problem. The algorithms used to discover these modules rely on heuristics and may vary the result (BARABÁSI, 2015). Figure 20 shows how two different algorithms extract communities from the networks S1 and S2.

Figure 20 – Communities.



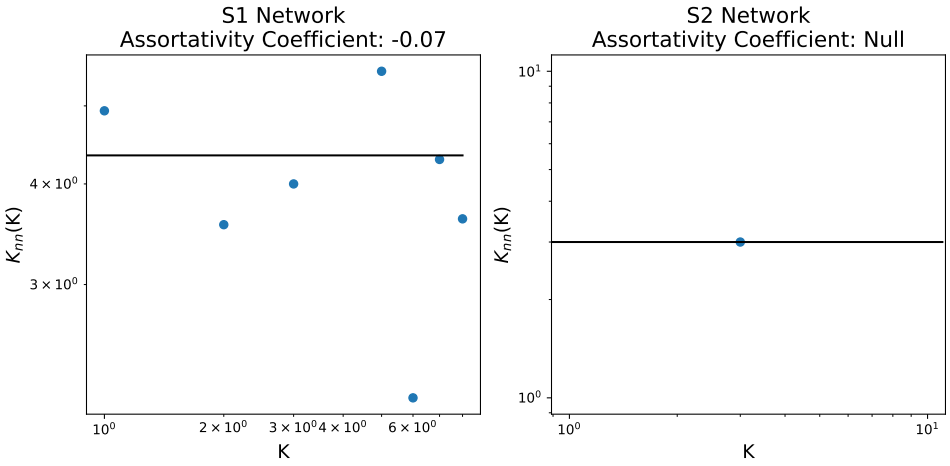
Source: Elaborated by the author.

The communities are presented as connected components, varying in color, but keeping the original node position. Both algorithms are implemented in the NetoworkX library, first is a greedy approach proposed by Clauset, Newman and Moore (2004), while the second is a $O(n^4)$ implementation. Although the algorithms find the same communities in S1 network, algorithm 2 is 990 times slower. Both algorithms find five communities in S2 networks, but they are different. Algorithm 2 performs 1586 times slower.

2.3.3.2 Assortativity

The assortativity measures preferences nodes use to connect with other nodes. Figure 21 shows the degree assortativity among nodes in the networks S1 and S2. Barabási (2015) explains that in social networks, famous people tend to know others famous. Thus hubs connect to hubs, creating a positive degree assortativity.

Figure 21 – Degree Assortativity.



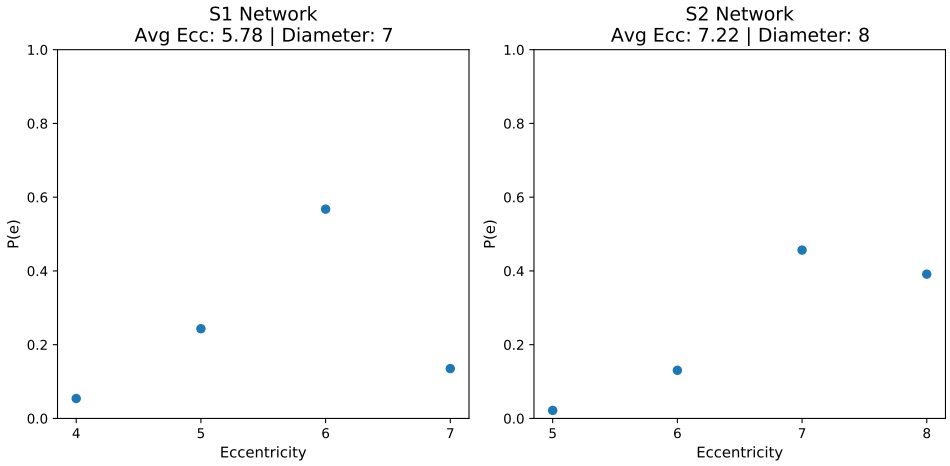
Source: Elaborated by the author.

In Figure 21, K (x-Axis) represent all nodes with degree K, and K_{nn}(K) (y-Axis) is the average neighbors' degree of nodes with degree K. In S2 Network, all nodes have degree 3. Therefore we have only one point in the plot, and the assortativity coefficient is null. The right-most point in the S1 Network represents all nodes with degree 8, which are nodes 0 and 2. The average neighbors' degree of node 0 is 3, and for node 2 is 4.25; Thus the $K_{nn}(8)$ is 3.62.

2.3.3.3 Small World and Diameter - Eccentricity

Eccentricity of a node is maximum shortest path from it to all nodes. The diameter is the maximum eccentricity find in the network. Figure 22 shows the eccentricity and diameter for network S1 and S2.

Figure 22 – Eccentricity and Diameter.



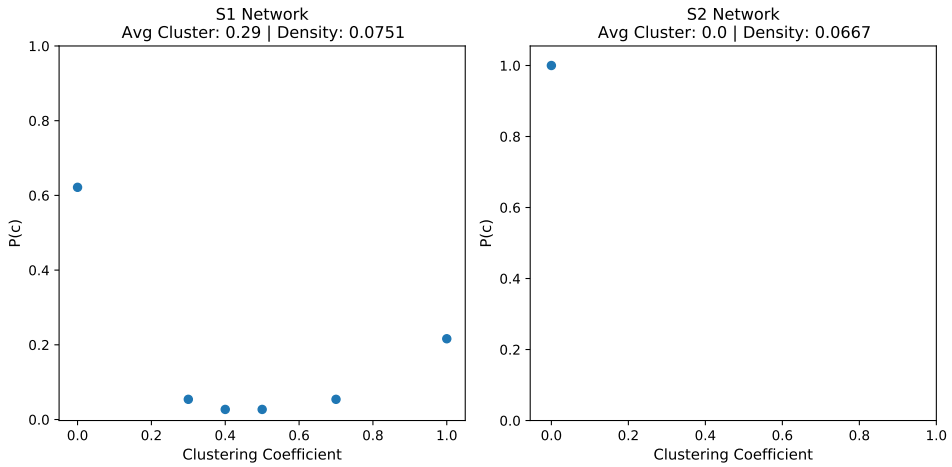
Source: Elaborated by the author.

The concept of a network be a “small-world” come from the fact that many real-world networks, albeit with thousands of nodes, have a relatively small diameter and average eccentricity. For example, as shown in Figure 30, ReactomeFI has 14k nodes and an average eccentricity of 7.72, while S2 network has 46 nodes an average eccentricity of 7.22. The small-world phenomenon happens due to the presence of hubs that shorten the distances.

2.3.3.4 Density and Clustering

Another important feature concerning the interconnections within a network is clustering. Section 2.3.2.2 discusses this centrality, but here we show the average cluster from the entire network the percentual of nodes that have a given clustering coefficient. Figure 23 shows this information for networks S1 and S2.

Figure 23 – Density and Clustering.



Source: Elaborated by the author.

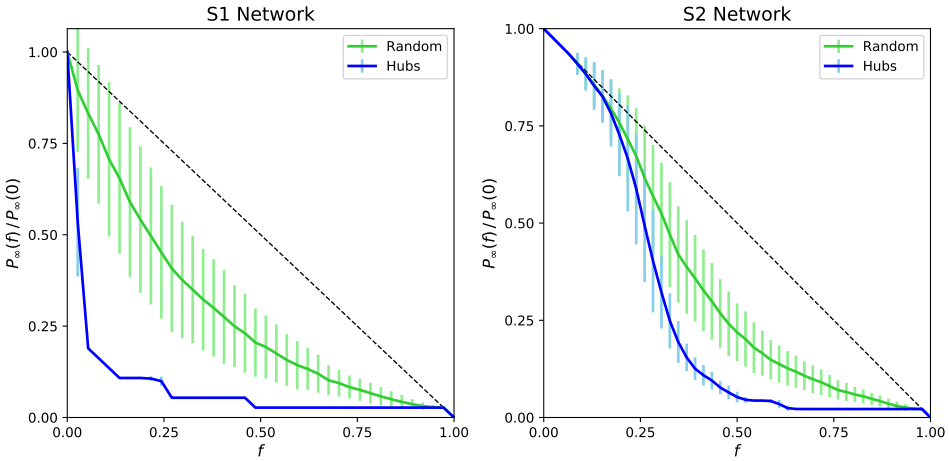
All nodes in S2 have 0 clustering. In S1 60% of the nodes have 0 clustering, and around 25% of the nodes have a clustering of 1. Figure 23 also show the network density, which is the ratio between the actual number of edges a network has by the maximum number of edges the network can have.

2.3.3.5 Attack and Resilience

Albert, Jeong and Barabási (2000) explores the network resilience to random removal of nodes and removal of hubs. Random attacks test the system modeled by the network to random failures, while the hub attack simulates intentional attacks. Figure 24 shows the impact of hubs and random attacks on S1 and S2.

In Figure 24 the x-Axis (f) indicates the percentual of nodes removed, and the y-Axis ($P_\infty(f) \,/\, P_\infty(0)$) is the percentual size of the largest connected component. We run 100 executions for random removal and 100 for hub removal. A vertical line over the impact line represents

Figure 24 – Attack and Resilience.



Source: Elaborated by the author.

the standard deviation from these executions, indicating the variance between executions. The dotted black line indicates zero impact. Removal does not break the network in more than one connected component. The constant degree distribution, and the consequent absence of hubs, in S2 make both intentional and random attack impact the same until around 25% of nodes removal. After 25% random removal starts to remove nodes with degree 2, 1, or 0, while hub removal keeps choosing degree 3 nodes. Random removal in S1 has a high standard deviation because hubs may be selected and cause a greater impact than the average.

2.4 Software and Libraries

Many software and programming libraries have been developed with the increased interest in network science, especially social and biological networks. Pavlopoulos *et al.* (2017) and Faysal and Arifuzzaman (2018) reviewed the most famous software used to analyse networks, focusing in the visualization aspects of theses tools. Albeit the software have some topological measures implemented, the programming libraries, like NetworkX[2] for Python and iGraph[3] for R, are best suited to topological characterize networks. NetworkX and iGraph are capable of plotting networks, all the network plotting from the previous section was made with NetworkX, but these packages can not handle large network plotting and do not have all the visualization options found in software like Gephi and Cytoscape. The library Graphviz[4] has an API for eight programming languages and is focused on visualization of graphs, networks, and diagrams.

Pavlopoulos *et al.* (2017) build a network of 202,424 nodes and 354,468 edges to test four network tools. Table 1 is an adaption of a similar table presented in their work. The tools are ranked between each other in ten attributes. Gephi is the most well-rounded. Cytoscape is the best to work with biological networks and has a large selection of plugins. Pajek is useful for massive networks, it can hold up to 2 billion nodes. Tulip main advantage is be easy to use and beginner-friendly (PAVLOPOULOS *et al.*, 2017).

Table 1 – Categorical comparison of four network software

	Cytoscape	Tulip	Gephi	Pajek
Scalability	3	4	2	1
User friendliness	3	1	2	4
Visual styles	1	3	2	4
Relevance to biology	1	3	2	4
Memory efficiency	4	3	2	1
Layouts	2	3	1	4
Plugins	1	3	2	4
Stability	2	4	1	2
Speed	3	4	2	1
Documentation	1	4	3	2

Faysal and Arifuzzaman (2018) review the same four network software. They also mention another 15 tools and libraries related to network analysis and visualization, included the previously mentioned NetworkX, iGraph, and Graphviz. They introduced some JavaScript libraries but did not mention the vis.js[5], that only not handle dynamic network plotting but also animated 3D graphics and timelines.

[2] https://networkx.org/
[3] https://igraph.org/r/
[4] http://www.graphviz.org/
[5] https://visjs.org/

To compare the four software Faysal and Arifuzzaman (2018) measure the time and space each tool use to run the same layout algorithm with five different datasets. Table 2 is a copy from the table presented in their paper that summarize the comparison.

Table 2 – Time and space comparison of network software with different datasets

Dataset	# Vertices	# Edges	Cytoscape		Gephi		Pajek		Tulip	
			Mem (MB)	Time (sec)	Mem (MB)	Time (sec)	Mem (MB)	Time (sec)	Mem (MB)	Time (sec)
No Graph	N/A	N/A	960	N/A	288	N/A	60	N/A	70	N/A
email-Enron	36,692	367,662	3,200	14	1,100	13	88	3	658	434
web-Stanford	281,903	2,312,497	12,300	125	3,970	101	-	-	-	-
web-Google	875,713	5,105,039	-	-	6,350	280	-	-	-	-
wiki-Talk	2,394,385	5,021,410	-	-	7,742	612	-	-	-	-

Gephi was the only tool that handled all datasets. Cytoscape could not run the two largest datasets and consume significantly more memory than Gephi. As expected by the work of Pavlopoulos *et al.* (2017), Tulip only works with small networks. Pajek's strength lies in working with large networks, but it was unable to run any of the networks with millions of edges.

2.5 Final Considerations

An intricate set of relations characterizes Complex Systems. The Complex Networks models these relations using nodes and links. Once the system is modeled as a network, there are many ways to analyze and extract information about the system using topological measures. This analysis can focus on the whole network or each node and can be made using software or libraries.

In this chapter, we presented some methods to topological characterize networks, starting with the degree distribution and the scale-free property that many real-world networks possess. We also presented ten centrality measures that characterize nodes and five global measures that characterize the whole network. Finally, we presented a summary based on two review articles that explore software and libraries to visualize and characterize networks. In the next chapter, we present the Protein-Protein Interaction Networks and characterize them using the methods described in this chapter.

CHAPTER

3

PROTEIN-PROTEIN INTERACTION NETWORKS

3.1 Contextualization

Advances in large-scale methods to identify the functional relationship between genes (such as gene expression correlations, protein-protein interactions (PPI), text mining associations, and others[1]) made possible the creation of Protein-Protein Interactions Networks (PPINs). PPI are present in many, if not all, biological processes, and the network analyses of these interactions can lead to unexpected biology (KOH *et al.*, 2012; LAGE, 2014). Genomics advances on mapping most humans' genes allowed the identification of genetic variations and their association with diseases. A consequent challenge is understanding how variations in some genes interfere in complex biological functions with thousands of interactions (LAGE, 2014).

Proteins operate in groups to develop complex tasks (LAGE, 2014). From a network perspective, group activities form clusters, and large clusters form communities. In PPINs, the communities (or modules) play an essential role since its often associated with functional modules, like pathways (PORRAS, 2016). The analysis of PPIN and its functional modules have already been used to study complex diseases, like cancer and diabetes (KOH *et al.*, 2012; SAFARI-ALIGHIARLOO *et al.*, 2014)

In this chapter presents several PPINs databases and the differences in their creation in Section 3.2. In Section 3.3 we topological characterize four PPINs, a social network, and a synthetic scale-free network, comparing how each network differs and associates between measures. This knowledge about PPINs and their topological features is important to this project since many computational methods use PPINs as inputs, including the proposal of this project.

[1] https://www.ebi.ac.uk/training/online/courses/protein-interactions-and-their-importance/where-do-the-data-come-from/

3.2 PPIN Databases

The junction between curation of scientific literature and computational methods resulted in a vast collection of protein interaction databases (SZKLARCZYK; JENSEN, 2015). Even though the interaction data quality was questioned in the beginning, advances in the last decade approaches significantly increased the data quality. Furthermore, data standards and curation practices between databases improve the interactions data confidence and availability (LAGE, 2014).

The International Molecular Exchange consortium[2] (IMEx) aims to standardize curation rules concerning the identification of interactions. It also defines a standard data format, site search interface, and free access through the Creative Commons Attribution License. Table 3 list twelve PPI Databases, how many organisms the database cover, the total number of interactions, if the database follows the IMEx, and the database link.

Table 3 – PPI Databases

PPI Database	Organisms	Interactions	IMEx	Link
DIP	-	-	Yes	https://dip.doe-mbi.ucla.edu/dip
IID	18	7,369,019	Yes	http://iid.ophid.utoronto.ca
InnateDB	3	462,421	Yes	http://www.innatedb.com
IntAct	16	1,144,360	Yes	https://www.ebi.ac.uk/intact
MatrixDB	10	106,393	Yes	http://matrixdb.univ-lyon1.fr
MINT	587	76,501	Yes	https://mint.bio.uniroma2.it
HPRD	1	37,039	No	http://hprd.org
BioGRID	82	2,285,361	No	https://thebiogrid.org/
HINT	12	-	No	http://hint.yulab.org/
Reactome FI	1	268,857	No	https://reactome.org/
STRING	14094	> 20 billions	No	https://string-db.org/

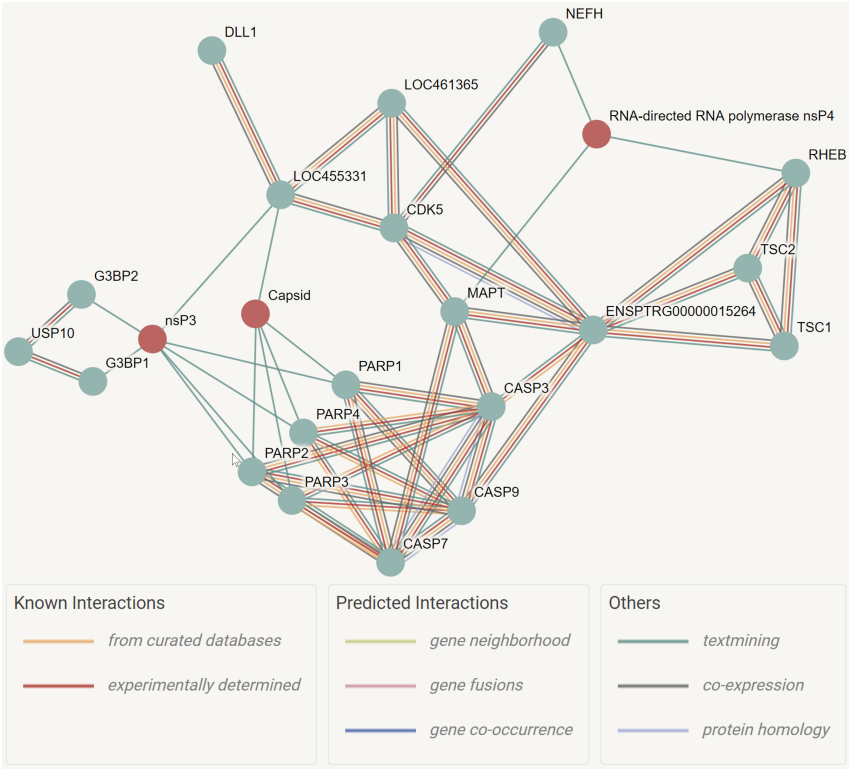
The DIP (Database of Interacting Proteins) database, although listed at IMEx site, seems long outdated. None of the download links worked. HPRD (Human Protein Reference Database) is also outdated. The last version date from 2010, albeit the download links are still functional. HINT (High-quality INTeractomes) differs from others PPI databases since its interactions come from a consensus between eight databases: BioGRID, MINT, iRefWeb, DIP, IntAct, HPRD, MIPS, and the PDB. Five of these databases are presented in Table 3, the others are: iRefWeb also use a consensus between databases, and aim to simplify queries concerning Functional Enrichment (topic further explored in Section 4.3); MIPS is also outdated, the last version is from 2004, and their interactions came only from manually curated high-quality scientific literature; PDB stands for Protein Data Bank, and its main focus is more about the protein (3D Shape, nucleic acids, and complex assemblies) than their interactions. Reactome is a well known pathways database, the Reactome FI (Functional Interaction) is a PPIN created to reflect

[2] http://www.imexconsortium.org/

functional events present in cellular pathways. They use their pathway database in junction with predicted interactions from IntAct, HPRD, and BioGrid (WU; FENG; STEIN, 2010). STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is the largest database we know. Their goal is to catalog all PPIN in a wide range of species (SZKLARCZYK *et al.*, 2021). Many of the organisms covered by STRING are bacteria and viruses. They even have a specific site for viruses: http://viruses.string-db.org/

A distinction between the databases and their PPINs are the interactions (edges) score and types. Reactome FI, for example, has three types of interactions: '–>' for activating/catalyzing; '–|' for inhibition; '-' solid line for complexes or inputs. Interactions score ranges from 0.5 to 1. A score of 1 means the interaction came from a pathway, while the others values are the level of confidence from predicted interactions. The work of Reyna, Leiserson and Raphael (2018) uses the Reactome FI from 2016 and only considers edges with a score of 0.75 or higher[3], therefore increasing the network confidence. STRING uses eight types of interactions, and the score ranging from 0 to 1, representing the probability of an interaction be true. Figure 25 shows the STRING's interactions types. The network is from the virus *Chikungunya*, where the red nodes are from the virus and the gray nodes from the host (*homo sapiens*).

Figure 25 – Viruses STRING: Interactions Types from *Chikungunya's* network.



Source: Extracted from http://viruses.string-db.org/cgi/network.pl?taskId=rgyN8lsRc0Gr.

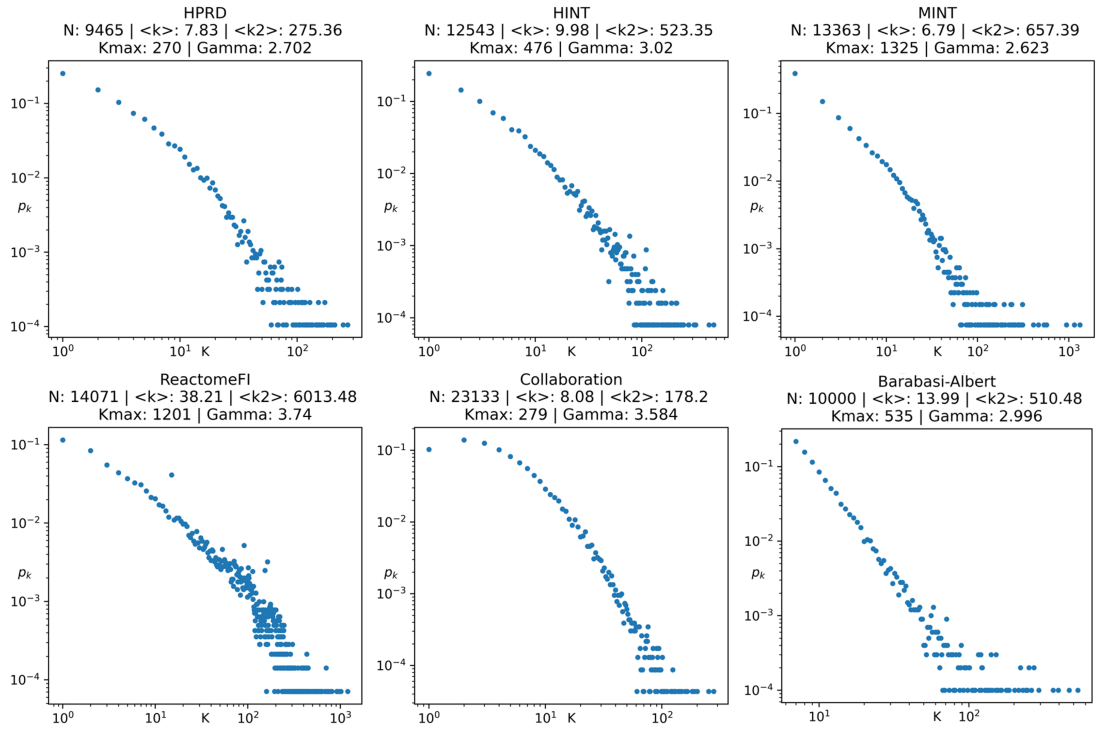
[3] This information is on the Supplementary Material

3.3 Protein-Protein Interaction Networks: Topological Characterization

We use metrics and approaches presented in Section 2.3 Topological Characterization to analyse four *homo-sapiens* PPINs, a social network and a synthetic scale-free network. These analyses enable us to characterize and compare the networks, which is relevant, since many computational methods use PPINs as input.

We choose HPRD that, albeit outdated, is used as a foundation for others networks, like HINT and Reactome FI. HINT was chosen because it is a consensus network between eight databases. MINT follows the IMEx standards and does not use HPRD as a background. Reactome FI is a PPIN constructed upon pathways and predicted interactions. The social network is a scientific collaboration network used by Barabási (2015). We choose this social network to contrast with the four PPINs that model complex systems from the cellular realm. We also use a synthetic scale-free network to analyze the similarities and differences between networks further. Figure 26 shows the degree distribution for these six networks.

Figure 26 – PPINs Degree distribution and scale-free characterization.



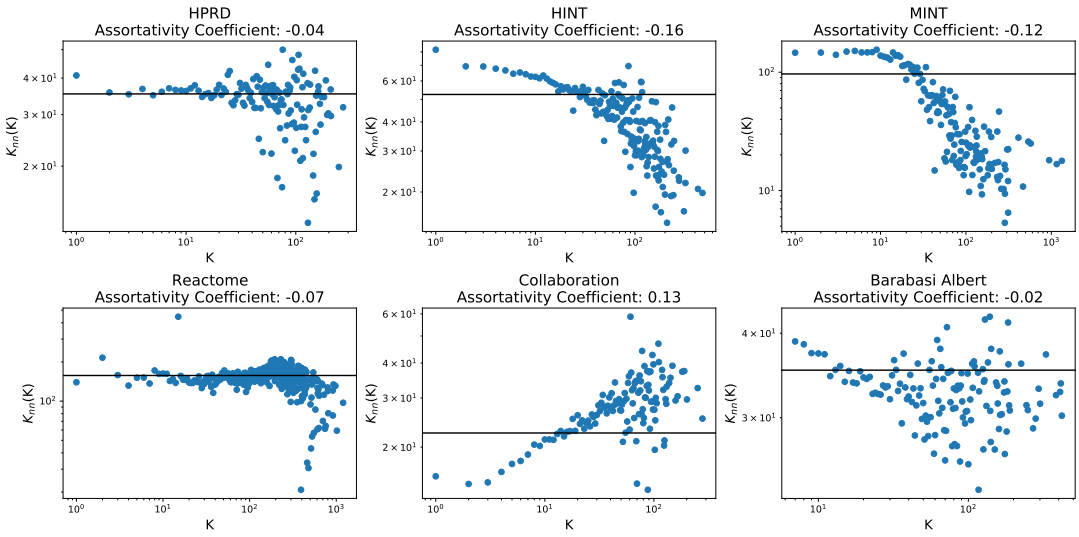
Source: Elaborated by the author

Collaboration has the highest number of nodes (N). A little more than 23k. The others networks nodes range from 9.4k to 14k. HPRD, HINT, MINT, and Collaboration have an average degree (<k>) between 6.79 and 9.98. Reactome FI has the highest average degree, of 38. Collaboration has, relative to N, a small biggest hub (kmax). In a network with 23K nodes, the biggest hub is connected with around 1.2% of the nodes, in MINT and Reactome this relation

is 9.9% and 8.5%. Gamma is the exponent from Equation 2.1 and helps to characterize the scale-free nature of a network (topic further discussed in Section 2.3.1). Overall, the networks have singularities in their degree distribution, but all have a scale-free distribution.

Figure 27 shows the degree assortativity for the six networks. If a node has a degree (K) and its neighbors an average degree (K_{nn}) close to K, this node has a positive assortativity. In Figure 27 the Assortativity Coefficient indicates the whole network assortativity, and the solid black line the expected behavior of a neutral network. Each point in the plot represents all the nodes with degree K. Considering that in a scale-free network, few nodes have a high K, and most nodes have a small K, a point at the beginning of the plot represents more nodes than a point at the end.

Figure 27 – PPINs Assortativity.

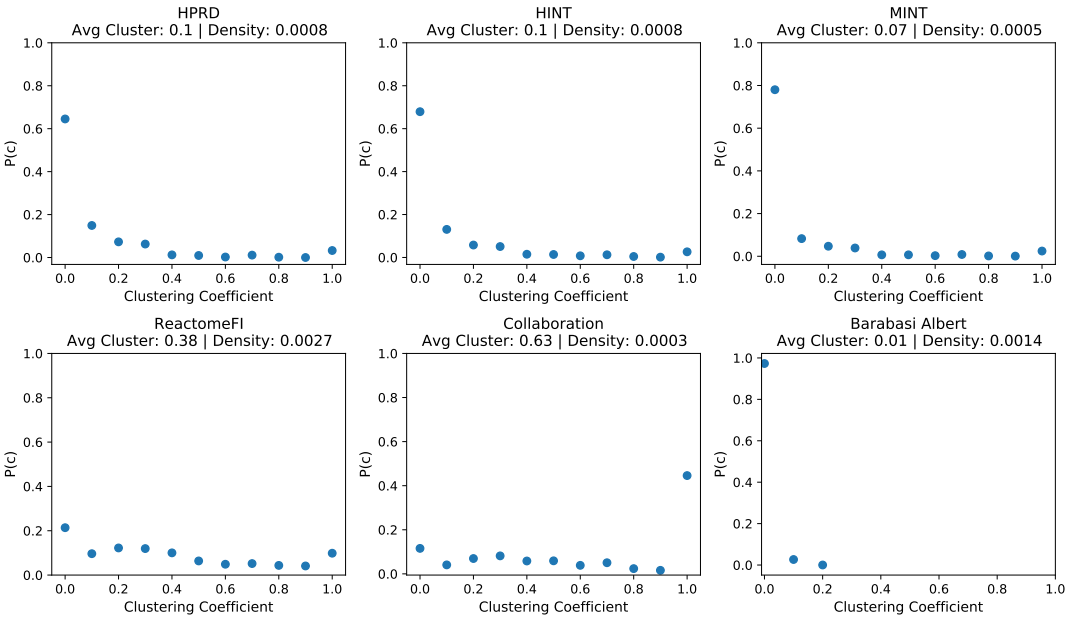


Source: Elaborated by the author

The Barabasi-Albert network has the most neutral assortativity, with a coefficient of -0.02 and points spread across to the solid black line. HPRD and Reactome FI share a similar behavior, neutral assortativity within the majority of nodes and negative assortativity among hubs. In other words, hubs tend to connect with small degree nodes, and small degree nodes do not show a clear pattern. HINT and MINT plots follow a downward line, with a clear negative assortativity in hubs and a mixed pattern among small degrees in MINT and around the average degree in HINT. Social networks are expected to have a positive assortativity, where famous people (hubs) know a lot of people, including other famous, and non-famous people (small degree) know non-famous (BARABÁSI, 2015). The Collaboration network has an inverse pattern found in PPINs: a clear behavior within small degree nodes and an unclear pattern among hubs. A possible interpretation for this network is that students (small degree nodes) publish articles with their advisors, thus have a positive assortativity. Famous researchers that publish more articles with other researchers than with students have a positive assortativity. Researchers that publish more articles with students than with other researchers have a negative assortativity.

Figure 28 shows the clustering and density for the six networks. HPRD and HINT have the same average clustering and density. They also have a similar distribution, in which more than 60% of the nodes have zero clustering, meaning that these nodes do not have neighbors, or that neighbors do not share links. MINT follows the same path as HPRD and HINT, but with a smaller average clustering and density, with almost 80% of the nodes with zero clustering. Looking at the assortativity of these three networks in the previous figure, we see that nodes with a degree of one are connected with high degree nodes, which indicates a star topology among them and thus zero clustering.

Figure 28 – PPINs Clustering and Density.



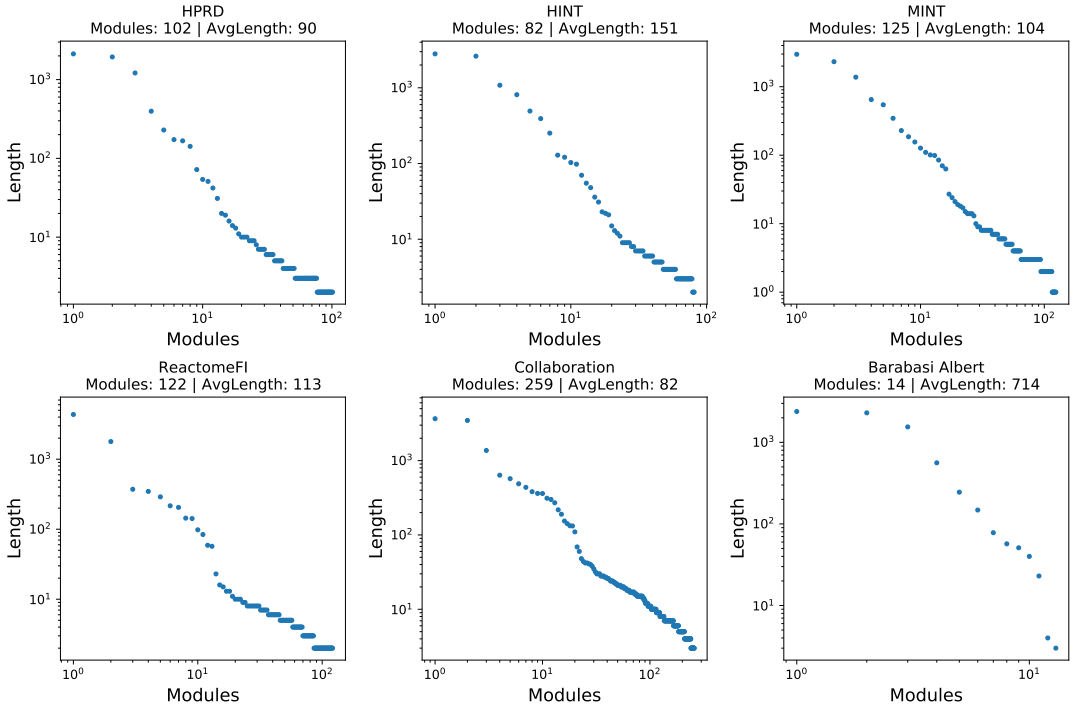
Source: Elaborated by the author

Reactome FI has the highest density and the second-highest clustering. Also, it is the only network which an almost uniform clustering distribution. This considerable difference with the other PPINs may come from its creation that favors functional modules (WU; FENG; STEIN, 2010). The use of pathways (groups of proteins that work together to develop a biological function) may be responsible for the higher average clustering and distinguish distribution compared with others PPINs. Collaboration has the opposite behavior from the first three networks. More than 40% of the nodes have 100% clustering coefficient. Even though this network has the least density, the average clustering is 63%. As expected, Barabasi-Albert network has almost no clustering (BARABÁSI, 2015). Compared with Collaboration, Barabasi-Albert is 4.5 times denser but has an average cluster of 1%, against 63%. This happens because the Barabasi-Albert model uses a probabilistic preferred attachment to capture the scale-free degree distribution found in many real-world complex systems. However, this model does not capture the “inner” preferred attachment among nodes. In a social network, it is expected that people with similarities cluster together. The high clustering found on the Collaboration network

may come from the notion that researchers from the same field know each other and tend to publish articles together. The same logic applies to research groups.

The interconnection among neighbors form clusters, and high clustering areas in the network form modules (or communities). The identification of modules is an NP problem, thus the number and sizes of modules found in a network may vary depending on the algorithm used (BARABÁSI, 2015). Figure 29 shows the number of modules and the average module length in each of the six networks. To find the modules we use the greedy algorithm propose by Clauset, Newman and Moore (2004) and implemented by the NetworkX library.

Figure 29 – PPINs Modularity.



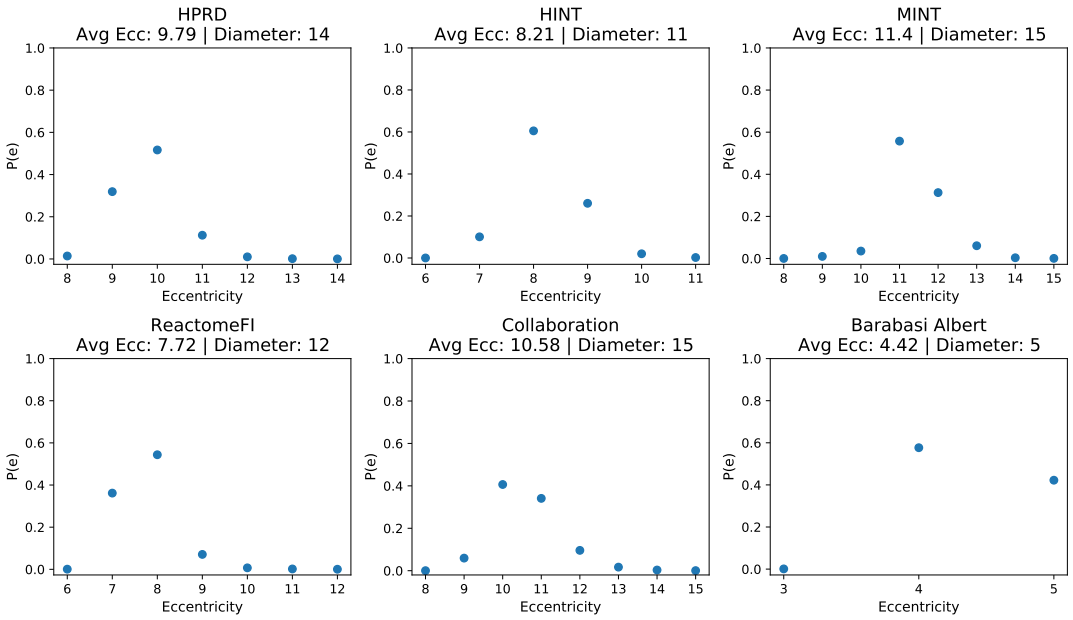
Source: Elaborated by the author

Following the idea that modules emerge from high clustering areas, the Collaboration network, with average clustering of 63%, has 259 modules, while the second network with most modules has 125. The four PPINs have the number of modules ranging from 82 to 125, even though three of them have low clustering. This behavior may be associated with the negative assortativity found in hubs (star topology) or with interconnection that does not form triangles but forms modules (which is the case of the CARSPER module find the Reproduction Super Pathways, Figure 32). Contrary to the PPINs, the lack of clustering in Barabasi-Albert network does correlate with the lack of modules. With 10k nodes, the network only has 14 modules.

Another way to characterize and compare networks is to measure the distances between nodes. Here we use the eccentricity, which measures the maximum shortest path from all nodes

to all nodes. We also show the diameter, which is the maximum eccentricity. Figure 30 present these measures.

Figure 30 – PPINs Eccentricity and Diameter.



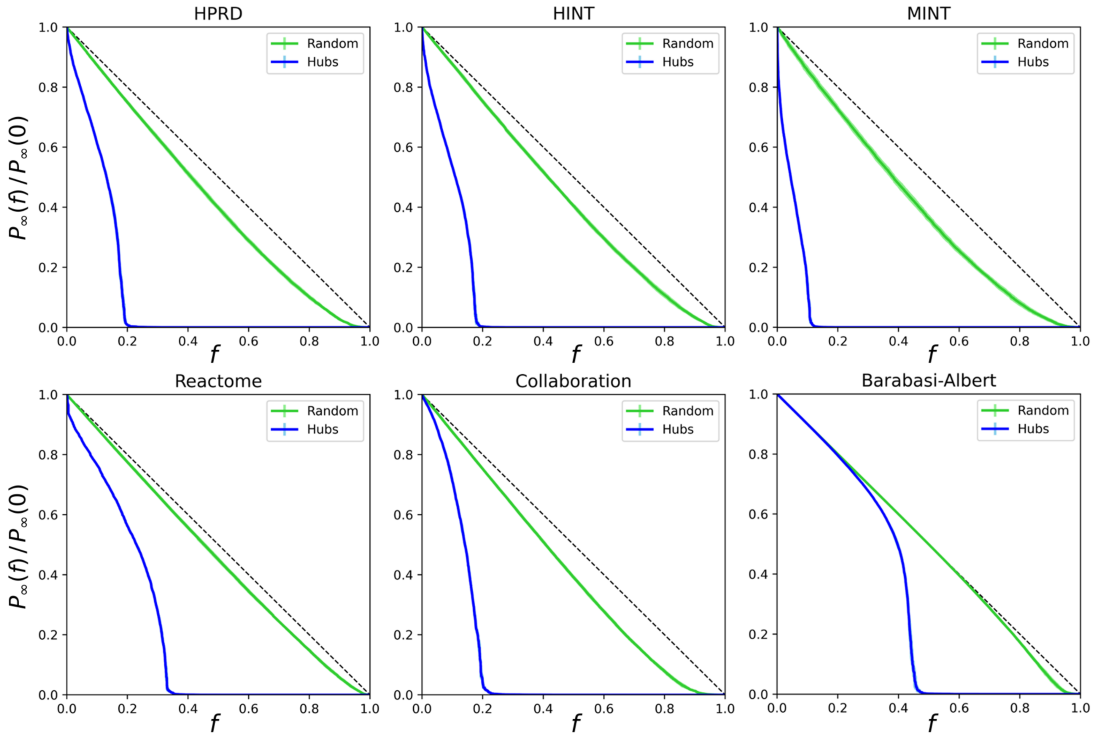
Source: Elaborated by the author

All five real-world networks have a diameter that does not represent the average eccentricity. HPRD, for example, has almost 80% of its nodes at most 9 or 10 steps to any other node and a diameter of 14. Overall, the five networks show a "small-world" behavior since the average eccentricity and diameter are small compared with the total number of nodes. This happens due to the hubs. These high degree nodes interconnect different parts (modules) of the network and shorten the distance. Collaboration has an average eccentricity and diameter similar to the PPINs, albeit it has 23k nodes, and the biggest PPIN has 14k. This may be due to the positive assortativity, where hubs are connected to hubs and thus shorten even more the network distances. Barabasi Albert network has 10k nodes and a diameter of 5, Barabási (2015) describes this phenomenon as "ultra-small property". The probabilistic model used to create this network favors the connection of new nodes to the high degree nodes in the network, creating a neutral assortativity where hubs connect to hubs and with low-degree nodes, keeping all nodes close to each other. The ultra-small diameter in conjunction with the lack of modules found in the Barabasi Albert network evidence a non-spread pattern among nodes.

Lastly, we compare the six networks' resilience to random and hub attacks. Many real systems have high resilience to error but may show vulnerability to intentional attacks. As the networks models systems, random and intentional removal of nodes can be used to test the system resilience (ALBERT; JEONG; BARABÁSI, 2000). Attack and resilience have been used to compare the impact cancer driver genes have on the pathways networks. In some pathways, the removal of driver genes causes more impact than the removal of hubs (RAMOS *et al.*, 2021).

Figure 31 shows the attack on the six networks. We incrementally remove nodes in each network randomly choosing nodes, the green line represents the impact of this removal. We also remove the biggest hub in each interaction, the blue line represents the impact of this removal. We run 30 executions for random removal, and 10 for hub removal. The standard deviation from these executions is represented by a vertical line over the impact line, albeit barely visible, indicating a slight variance between executions. The dotted black line indicates zero impact, removal does not break the network in more than one connected component. The x-Axis (f) indicates the percentual of nodes removed, and the y-Axis ($P_\infty(f)\,/\,P_\infty(0)$) is the percentual size of the largest connected component.

Figure 31 – PPINs Attack.



Source: Elaborated by the author

HPRD, HINT, and Collaboration show a similar resilience. High resistance to random attacks but susceptible to hub removal. In these networks, the blue line touch the x-Axis around 20%, meaning that the network is complete without edges if we remove 20% of the hubs. MINT is the least resilient network. Figure 26 shows that MINT has high degree hubs. This feature allied with the negative assortativity, small density, low clustering, and the highest average eccentricity explains the observed impact on hubs removal. Reactome FI also has high degree hubs but is the most resilient real-world network. This may be due to the high clustering, small average eccentricity, and neutral assortativity among most nodes. Barabasi Albert shows remarkable resilience to hubs and random attacks, despite almost no clustering. The reason for the resilience lay in the tiny diameter. Every node is so close to each other that even with the removal of hubs

the network keeps a large connect component due to the alternative paths the small diameter confer.

3.4 Final Considerations

Improvements in the methods to identify functional relations between proteins made possible the creations of PPINs. Many databases were created, and the PPINs evolved with time. This chapter presents the topological characterization of four PPINs and shows they have a scale-free degree distribution and are resilient to random attacks, but fragile to hub removal. When comparing measures like assortativity, clustering, eccentricity, and modularity, the PPINs display some differences. The similarity in some measures and differences in others show the need to diversify the topological characterization when analyzing and comparing networks. It also raises the question of how methods that use PPINs as input deal with the topological differences between networks. In the next chapter, we present the concept of pathways and how they can be modeled as induced sub-graphs of PPINs.

CHAPTER

4

PATHWAYS

4.1 Contextualization

Pathways are sets of genes that interact and are responsible for the emergence of specific biological functions. Each pathway works as building blocks of a cell's complex system. Some examples of pathways are: Circadian Clock, DNA Repair, and Programmed Cell Death (JASSAL *et al.*, 2020).

The PPINs aim to describe the whole-cell interactions. The PPINs explored in the previous chapter have an average of 12.360 nodes (genes), while the Programmed Cell Death pathway has 216 genes. These 216 genes are a considerably smaller subset and carry meaningful information about the genes' functional role. Pathways analyses reduce the PPINs complexity and enable a more thorough investigation of the studied genes (GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015).

4.2 Pathways Databases and Representation

With the increase of genomic data available and the interest in the use of pathways to functional analyze sets of genes, many databases were created (KHATRI; SIROTA; BUTTE, 2012). These databases are also called "knowledge bases" since they store the current knowledge of molecular interactions into lists or networks of genes (GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015). Pathguide[1] contains 702 databases related to molecular interaction for 24 different organism. Pathguide also groups these databases into 9 categories concerning the type of interaction: Protein-Protein Interactions; Metabolic Pathways; Signaling Pathways; Transcription Factors / Gene Regulatory Networks; Protein-Compound

[1] http://pathguide.org/

Interactions; Pathway Diagrams; Genetic Interaction Networks; Protein Sequence Focused; and Others.

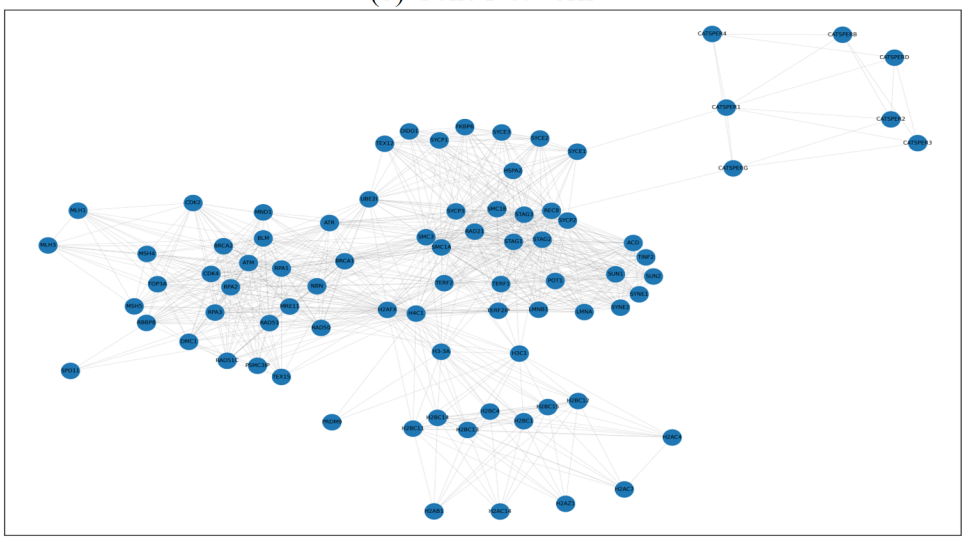
Another difference in these databases is the curation approach. In this regard, García-Campos, Espinal-Enríquez and Hernández-Lemus (2015) indicates that Reactome[2] is manually curated from literature by expert biologists working together with their editorial staff, and cross-referenced it with several others pathway related databases. Figure 32 shows the genes that Reactome associates to the Reproduction Pathway in three levels of abstraction/representation.

Figure 32 – Reactome's Reproduction Pathway: Three types of representation.

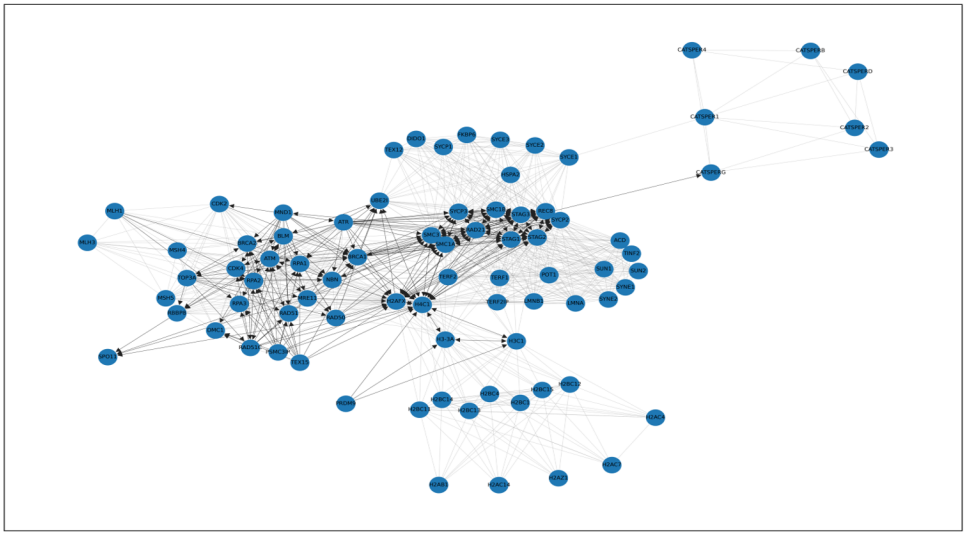
(a) Gene Set

ACD	ATM	ATR	BLM	BRCA1	BRCA2	CATSPER1	CATSPER2	CATSPER3	CATSPER4	CATSPERB	CATSPERD	CATSPERG	CDK2
CDK4	DIDO1	DMC1	FKBP6	H2AB1	H2AC14	H2AC4	H2AC7	H2AFX	H2AZ1	H2BC1	H2BC11	H2BC12	H2BC13
H2BC14	H2BC15	H2BC4	H3-3A	H3C1	H4C1	HSPA2	LMNA	LMNB1	MLH1	MLH3	MND1	MRE11	MSH4
MSH5	NBN	POT1	PRDM9	PSMC3IP	RAD21	RAD50	RAD51	RAD51C	RBBP8	REC8	RPA1	RPA2	RPA3
SMC1A	SMC1B	SMC3	SPO11	STAG1	STAG2	STAG3	SUN1	SUN2	SYCE1	SYCE2	SYCE3	SYCP1	SYCP2
SYCP3	SYNE1	SYNE2	TERF1	TERF2	TERF2IP	TEX12	TEX15	TINF2	TOP3A	UBE2I			

(b) Gene Network



(c) Gene Network with Edge Information



Source: Elaborated by the author.

[2] https://reactome.org/

In Figure 32-a the Reproduction Pathway genes consist only as a set of 81 genes. Considering that there are around 19.000 known protein-encoding genes (PIOVESAN *et al.*, 2019), a sub-set of 81 is meaningful and helps to identify a gene set concerning their functional role in the cell. Figure 32-b increase the information that can be extracted from the pathway by representing it as a network. Figure 32-c adds information to the edges. Interactions represented by "-" are from complexes or inputs, while "->" indicates activation or catalyzation.

The databases' types of representation have a direct impact on how the pathways can be analyzed (KHATRI; SIROTA; BUTTE, 2012; GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015).

4.3 Pathways Analyses

Pathways Analyses, or functional enrichment analysis, is an increasing field in genomic research. The analyses are implemented by tools that accept groups of genes and associate them with biological functions. Considering the large number of genes and pathways, these tools are fundamental in providing meaning to the input genes, enabling interpretation and hypothesis generation (GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015).

Overall, all tools combine biological function knowledge from pathways databases with statistical testing, mathematical analyses and computational algorithms (GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015). As the approaches evolved with the years, they can be categorized into three generations that share the same input (KHATRI; SIROTA; BUTTE, 2012):

- **Input:** The data usually came from High Throughput Genomic studies. For example, the study from Ciriello *et al.* (2015) makes available cancer mutation data from 817 breast tumors. We can get the altered genes in each tumor and analyze their functional role using pathways analyses.

- **First Generation - Over-Representation Analysis:** Calculates the probability of the intersection between the input genes and the pathways be a random event or not. For example, an input set with 10 genes in which 7 of them appear in the list shown at Figure 32-a. This means that 7 genes are associated with the Reproduction Pathway, and 3 are not. Since there are 81 genes in this pathway and 71 are not associated with the input, it is possible to define the number of "background genes" as all other genes not associated with the input or the pathway. The notion and size of the background genes changes from method to method (POIREL; OWENS; MURALI, 2011; GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015; PACZKOWSKA *et al.*, 2020). It is possible to set the background gene as 19.000 since this is the expected number of protein-encoding genes (PIOVESAN *et al.*, 2019), or a smaller set, that consider the number of genes

that exist in the pathway database (PACZKOWSKA *et al.*, 2020). Naming the input set genes as A and the genes in the pathways as B, and using the above example, we have four parameters: $AandB = 7$; $AnotB = 3$; $BnotA = 71$; $neither = [19.000, 14.000, 9.000]$. Using Fisher's exact test and changing the background genes with values in $neither$, to show the impact of these changes in the result, we have the following p-values: $1.72e^{-15}$; $1.43e^{-14}$; $3.08e^{-13}$. Changes in the background genes impact the results, but as A and B are small, all results are $\ll 0.05$. The methods that use Over-Representation Analysis usually do this for every pathway and show an ordered list by the p-values, indicating the pathways that are more associated with the input set (KHATRI; SIROTA; BUTTE, 2012).

There are two main limitations of Over-Representation Analysis. First, they give the same weight for every gene, without considering their interactions and topological role in the pathway network, and without considering any statistics from the input set (High Throughput Genomic data) (KHATRI; SIROTA; BUTTE, 2012). Second, they consider the pathways to be independent of each other, without acknowledging that their networks overlap (BARABASI; OLTVAI, 2004; GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015).

- **Second Generation - Functional Class Scoring:** These second generation methods solve one of the limitations of the previous approach: they consider statistical measurements from the High Throughput Genomic data used as input to calculate the functional enrichment (GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015). The types of statistical measurements used are highly dependable on how the researchers acquire the data. Cancer mutation data associated with samples and patients can be found in MAF files (further details about MAFs are discussed in Section 5.1.1). To a MAF file be publicly available, they already pre-process the raw data making some of the analyses proposed by Khatri, Sirota and Butte (2012) unfeasible. However, the Functional Class Scoring can also use descriptive statistics over the pre-processed data to improve their analyses (GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015). For example, the study from Ciriello *et al.* (2015) contains 817 samples with a total of 69.968 mutations on 16.178 genes in a scale-free distribution, where few genes are very frequently and most of the genes are below the average mutation (long tail effect). The genes PIK3CA, TP53, and TTN appear altered in 282, 280, and 134 samples (respectively 35%, 34%, and 16%) while the mean and the median for this distribution are 3.43 and 2. Concerning the relation between genes and mutations, there are also different distributions in the SNV classes and the mutation classification from sample to sample (RAMOS *et al.*, 2020). Some methods that calculated genes and mutation probability consider the number of nucleotides in a weighted approach (LEISERSON; REYNA; RAPHAEL, 2016; CANISIUS; MARTENS; WESSELS, 2016). All these different measures about the input data can be used to increase the significance of the functional enrichment in a way that two

sets of the same length that would result in the exact p-value using Over-Representation Analysis may result different using Functional Class Scoring.

Although Functional Class Scoring is an evolution over the first generation, it still considers the genes as a "list", ignoring the network topology. Like the previous one, this approach also admits the pathways as independent units instead of building blocks of the cell complex system. (GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015).

- **Third Generation - Pathway Topology:** Following the increased interest in network science and its ability to model complex systems, Pathway Topology joins network measures with Functional Class Scoring to improve functional enrichment (KHATRI; SIROTA; BUTTE, 2012; GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015). To illustrate the difference between the three generations and how Pathway Topology is considered a evolution over the others, we define two input genes set with the same size: GS1={BRCA1, BRCA2, SMC1A, STAG1, STAG2} and GS2={CATSPER1, CATSPER2, CATSPER3, CATSPER4, CATSPERB}, and apply them to each method.

  Since GS1 and GS2 have the same size and both have 100% intersection with the list in Figure 32-a, Over-Representation Analysis methods will score the exact same p-value regarding the functional enrichment with the Reproduction Pathway.

  If we cross GS1 and GS2 with the MAF found in Ciriello *et al.* (2015), Functional Class Scoring methods may weigh the fact that none of the genes in GS2 appear in any altered sample, while the genes in GS1 appear respectively in 18, 19, 9, 10, 11 samples. The methods can also use the information that all genes in GS1 are known Cancer Driver Genes (more about drivers in Section 5.3), while none of GS2 are.

  Pathway Topology methods will add information over Functional Class Scoring. Using the network in Figure 32-b and calling all its nodes GN, we discover that the average[3] degree for GN is 21, while the GS1 is 35.6 and GS2 is 4.7. Respectively, the average clustering are 0.77, 0.6, 0.49, and the average closeness are 0.49, 0.59, 0.32, and the average betweenness are 0.014, 0.021, 0.013. The GS2 degree is 4.5 times smaller than GN, while the GS1 is 1.7 greater. The degree is a simple but important measure that greatly impacts the network (BARABÁSI, 2015), which gives a topological difference between GS1 e GS2. The behavior of GS1 being greater than GN is also found in the closeness and betweenness measures. GS1 may be interpreted as hubs (degree) that are in the middle of the network (closeness) and concentrate flow of information (betweenness). The opposite can be said about GS2. The bigger clustering in GN indicates that GS1 connects different parts (communities) in the network, which corroborates with the high betweenness. A visual inspection on Figure 32-b shows that GS2 are in a community, a

[3] We use the average of sets to illustrate the topological analyses, but the methods can use the information from each node, since we are using centrality measures.

high clustering area, but GS2 clustering is smaller than GS1 and GN. This happens because in that community the interconnections are formed by squares, not triangles. Changing the clustering algorithm from triangles to squares, we have the following values for GN, GS1 and GS2: 0.38, 0.12, 0.71.

Methods that analyze the pathway network can also consider the edge type, as shown in Figure 32-c. All the 16 edges involving GS2 are "-", indicating they are in the same complex. GS1 has 171 have edges, which are 103 of the type "-", 35 of the type "->" and 32 of the type "<->". Since the direct arrow indicates activation, this is another difference between GS1 and GS2 that can be used to functional enrich an input set.

The last example, using GN, GS1, and GS2, helps to show the differences in the pathways analyses types and how they work. Two sets that would be equal using Over-Representation Analysis are distinct when we add information about the input data and are also topological different concerning their role in the pathway network. Functional enrichment analyses show a more significant result when they combine Functional Class Scoring and Pathway Topology (KHATRI; SIROTA; BUTTE, 2012; GARCÍA-CAMPOS; ESPINAL-ENRÍQUEZ; HERNÁNDEZ-LEMUS, 2015).

4.4 Super Pathways Networks

The previous section showed how the network topology can increase the functional enrichment analyses. One advantage of using pathways networks over PPINs is that they are considerably small and carry significant functional information. However, there is a problem: Reactome has one PPINs on its database and 1.803 pathways networks (JASSAL *et al.*, 2020), and different and complementary measurements should be used to analyze each network (OLDHAM *et al.*, 2019).

In the most recent paper about Reactome (JASSAL *et al.*, 2020), the authors present the concept of Super Pathways, which represent 26 biological functions that group 1803 pathways. Using 26 Super Pathways Networks (SPN) instead of 1803 makes the topological analyses more feasible.

Following the pipeline proposed in Figure 46, we create 26 SPN enriched with cancer driver genes. As the cancer driver genes are topological different than non-driver genes (RAMOS *et al.*, 2021), we use the drivers to choose the four SPNs most associated with cancer and explore their differences. Table 4 presents a summary of all 26 Super Pathways and the corresponding SPN. LenSet, which is the length of Super Pathway gene set; LenCC, which is the size of the resulting nodes (genes) in the induced network that may have more than one Connect Component (CC); LenLCC, which is the size of nodes in the Largest CC; Driver LCC, which is the number

of cancer driver genes in the LCC; Driver %, which shows the percentual presence of drivers in the LCC.

Table 4 – Super Pathways Network

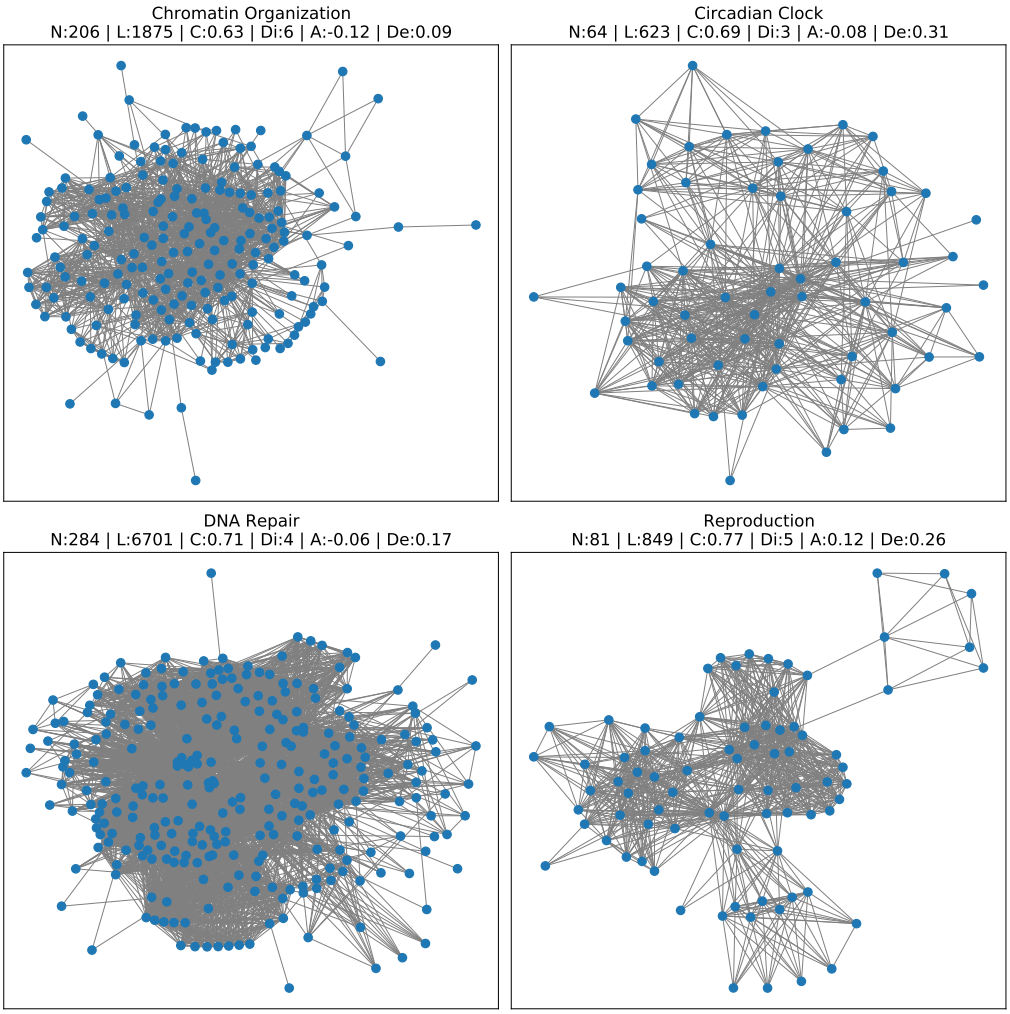
Super Pathway Name	LenSet	LenCC	LenLCC	Driver LCC	Drivers %
Chromatin organization	240	218 (91%)	206 (86%)	45	22
Circadian Clock	70	69 (99%)	64 (91%)	12	19
DNA Repair	312	290 (93%)	284 (91%)	48	17
Reproduction	114	95 (83%)	81 (71%)	14	17
Gene expression	1536	1392 (91%)	1367 (89%)	194	14
Developmental Biology	1097	972 (89%)	962 (88%)	137	14
Programmed Cell Death	216	208 (96%)	201 (93%)	27	13
Cell-Cell communication	122	117 (96%)	115 (94%)	15	13
Signal Transduction	2542	2399 (94%)	2363 (93%)	285	12
Cellular responses	665	619 (93%)	604 (91%)	73	12
Cell Cycle	662	636 (96%)	632 (95%)	78	12
Immune System	2088	1910 (91%)	1818 (87%)	174	10
Hemostasis	687	608 (89%)	554 (81%)	54	10
Metabolism of proteins	2017	1816 (90%)	1725 (86%)	133	8
Organelle biogenesis	282	271 (96%)	265 (94%)	22	8
Autophagy	132	131 (99%)	129 (98%)	10	8
Metabolism of RNA	721	651 (90%)	607 (84%)	42	7
Extracellular matrix	296	287 (97%)	272 (92%)	20	7
Vesicle-mediated transport	730	687 (94%)	661 (91%)	40	6
Neuronal System	393	368 (94%)	339 (86%)	20	6
Metabolism	2125	1650 (78%)	1371 (65%)	74	5
Muscle contraction	192	174 (91%)	156 (81%)	8	5
DNA Replication	128	128 (100%)	128 (100%)	4	3
Sensory Perception	605	543 (90%)	498 (82%)	8	2
Protein localization	164	155 (95%)	150 (91%)	3	2
Digestion and absorption	28	16 (57%)	3 (11%)	0	0

The four chosen Super Pathways have specific biological functions and are associated with cancer. **Chromatin organization** refers to the composition and conformation of complexes between DNA, protein, and RNA (REACTOME, 2011), and have been reported to have a significant influence on regional mutation rates in human cancer cells (SCHUSTER-BÖCKLER; LEHNER, 2012). **Circadian Clock** is a master regulator of mammalian physiology, regulating daily oscillations of crucial biological processes and behaviors. Notably, circadian disruption has recently been identified as an independent risk factor for cancer and classified as a carcinogen (SHAFI; KNUDSEN, 2019). **DNA Repair** is responsible for the integrity of the cellular genome, and its malfunction is a notorious cancer hallmark (JIN; OH, 2019). **Reproduction** pathway mixes the genomes of two individuals creating a new organism (REACTOME, 2006). The mutations in BRCA, that belong to the Reproduction pathway, and their role in fertility is studied by (DAUM; PERETZ; LAUFER, 2018).

4.4.1 Topological differences

The four chosen SPN have distinct topologies. To start exploring these differences, we present the Figure 33 that offers a visual representation of the networks and some global network measures.

Figure 33 – Chosen Super Pathways Networks.

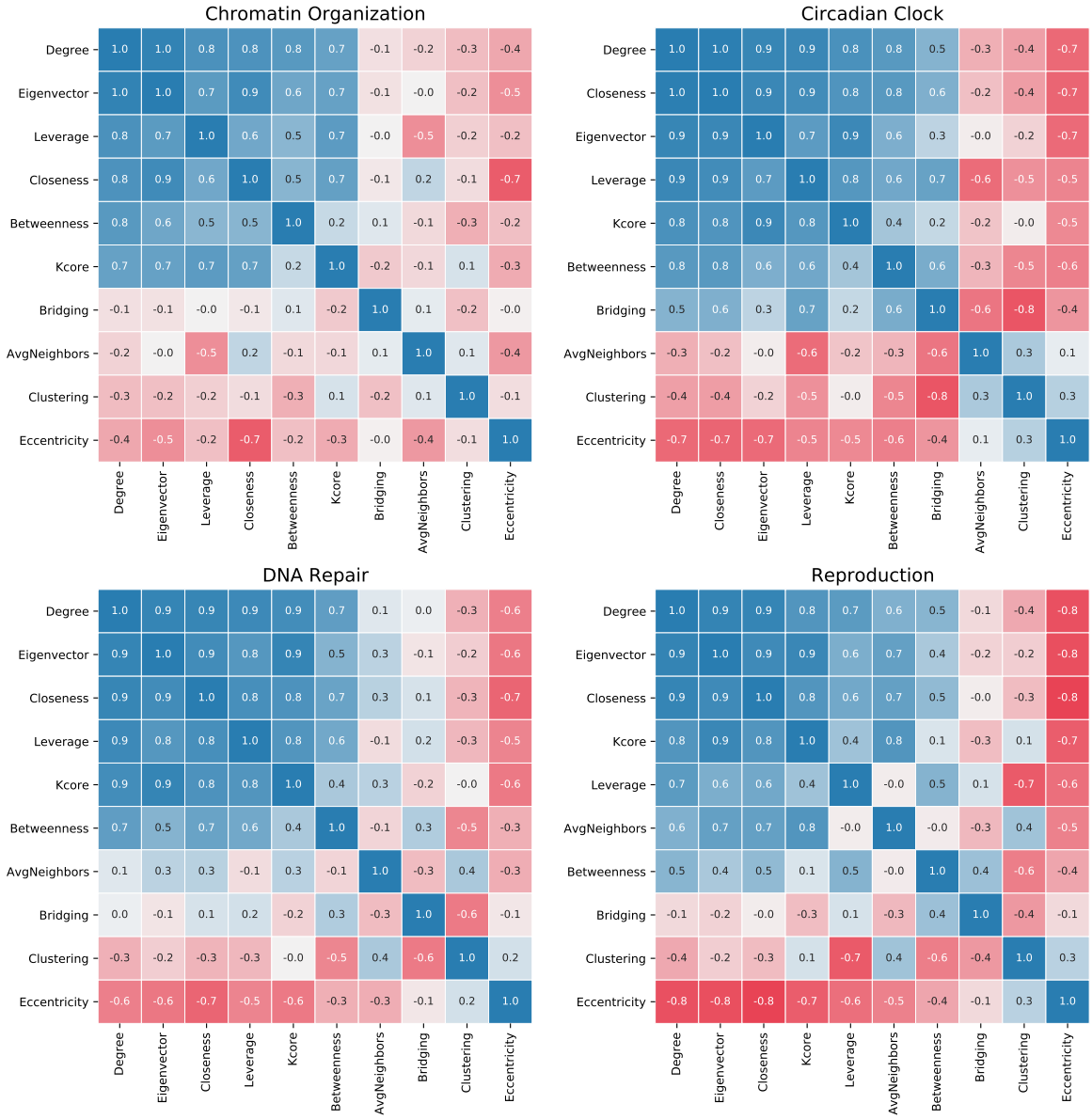


Source: Elaborated by the author.

On each subplot, the initials stand for N, the number of nodes; L, the number of links; C, the average clustering; Di, the diameter; A, the assortativity coefficient; and De, the density. Section 2.3.3 explain and exemplify these measures. Albeit the number of nodes and links differ, all networks have a high clustering and a small diameter. A visual inspection shows the presence of high-density areas, allied with clustering, these areas form communities that represent functional complexes (PORRAS, 2016). The small values in assortativity indicate that there is no clear preferred attachment regarding degree similarity.

We use the same 10 centrality measures showed in Section 2.3.2 to analyse the networks. Figure 34 shows the correlation between measures. As expected (OLDHAM *et al.*, 2019), degree is the most correlated and we order the plot axis by this association.

Figure 34 – Chosen Super Pathways Networks Measures Correlation.



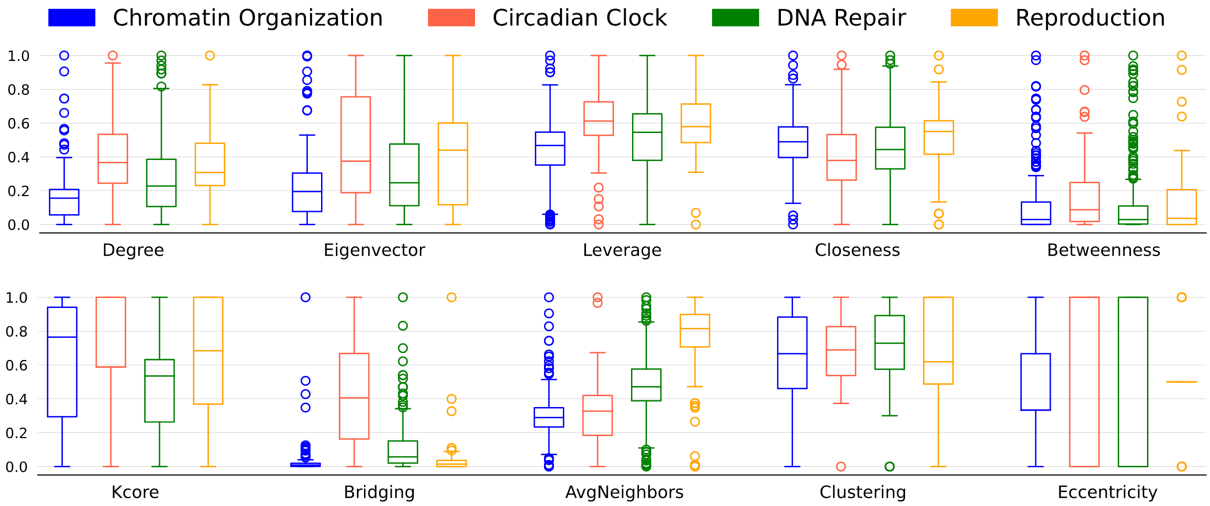
Source: Elaborated by the author.

Chromatin Organization shows a high correlation among Degree, Eigenvector, Leverage, Closeness, Betweenness, and Kcore. These six measures are also correlated in the other networks. Eccentricity has the most negative correlation, followed by clustering, although this one correlates with Average Neighbors and Eccentricity. The correlation matrix works as a network fingerprint and serves as an overview among centralities in the network (OLDHAM *et al.*, 2019).

Another way the overview the centralities is to analyze their distributions. Figure 35 shows the box plot for the ten centralities used in Figure 34. Looking at the median, Clustering

and Closeness are the most uniform measures among the four SPNs. Eccentricity is the measure with the least variation. All nodes in Circadian Clock have an eccentricity of 2 or 3, while in DNA Repair these values are 3 or 4. In Reproduction SPN the majority of nodes are at most 4 steps to any other nodes, but some are 3 or 5 steps. In Chromatin Organization, the eccentricity range from 3 to 6.

Figure 35 – Chosen Super Pathways Networks Measures Distribution.



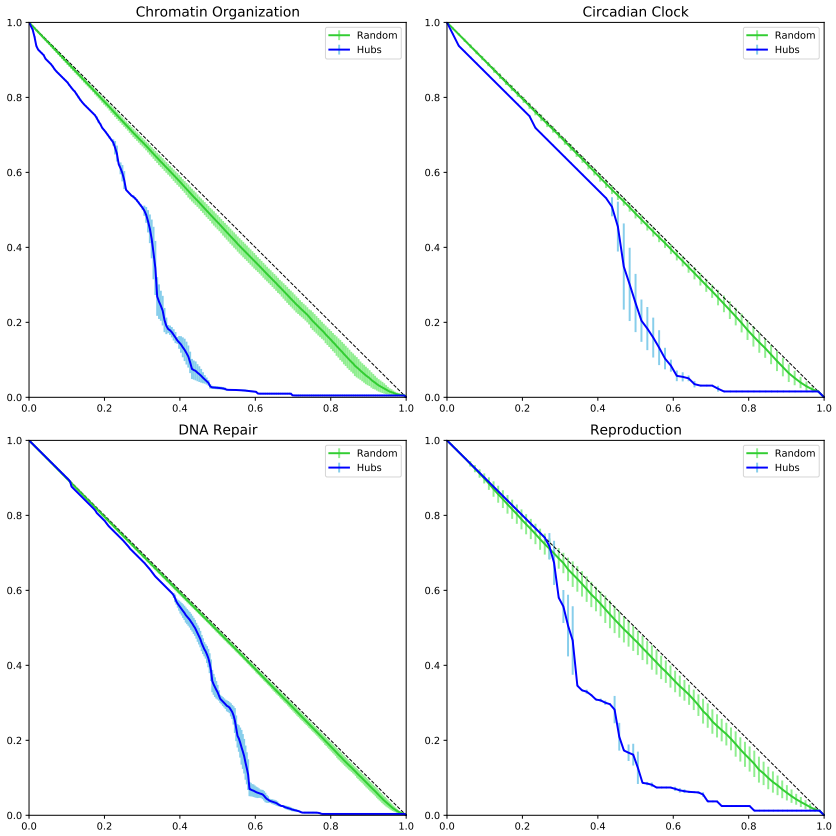
Source: Elaborated by the author.

The degree distribution, especially Circadian Clock and Reproduction, has fewer outliers than expected in scale-free networks. This may be due to the relatively small size. DNA Repair, the biggest SPN analyzed, has 284 nodes, while the Reactome FI has 14k. Overall, the four SPNs share similarities and differences among their centralities. In conjunction with global measures, these nuances can be used the compare and categorize networks. The centrality comparison between groups of nodes is an approach to topological characterize and distinguish them. (RAMOS *et al.*, 2021).

Finally, we analyze the SPNs' resilience to random and intentional attacks. The network resilience shows how networks sustain failure, and it is a way to characterize and differentiate them (ALBERT; JEONG; BARABÁSI, 2000). Intentional attacks was successfully used to study the topological role of driver cancer genes (RAMOS *et al.*, 2021), and can be explored to discover the impact functional genes have on the network structure.

We incrementally remove nodes from the four SPNs randomly choosing from the set of genes. We also remove the biggest hub in each interaction. Figure 36 shows these attacks. As we run 100 executions for each group (random and Hubs), a vertical line over the impact line represents the standard deviation from these executions. The dotted black line indicates zero impact when removal does not break the network in more than one connected component. The x-Axis indicates the percentual of nodes removed, and the y-Axis is the percentual size of the largest connected component.

Figure 36 – Chosen Super Pathways Networks Resilience.



Source: Elaborated by the author.

All networks show a high resilience to random and hubs attacks, specially Circadian Clock and DNA Repair, that maintain the hub impact close to the dotted-line until around 40% removal.

4.5 Final Considerations

The identification of sets of genes associated with biological functions motivated the development of pathways' databases and Functional Enrichment approaches. Recent methods model the pathways as networks and consider the genes' topological role in the enrichment process. Reactome database groups more than 1.800 pathways into 26 Super Pathways. The use of Super Pathways Networks confers specific biological functions to a smaller set than the whole PPIN, and allows the topological characterization of each gene in each Super Pathway.

CHAPTER

5

CANCER GENOMICS

5.1 Contextualization

Cancer is a complex disease characterized by genetic mutations that happen in a cell and lead to uncontrolled growth and division (HANAHAN; WEINBERG, 2000). The investigation of such genetic variations can explain the initiation and evolution of the disease, thus enabling personalized therapies. With the advent of next-generation sequencing (NGS) technologies, a significant volume of DNA sequencing has been generated (DEMKOW; PLOSKI, 2015). Several databases such as "The Cancer Genome Atlas" (TCGA) and "International Cancer Genome Consortium" (ICGC) make available NGS data. Researchers have widely used data sets with cancer mutation information to study mutations in cancer, genomic instability, and tumor evolution. Such studies utilize computational methods that load and analyze NGS data.

This chapter presents the MAF file (a commonly used data set to analyze cancer mutation), and discuss the high heterogeneity found in cancer data, and the phenomenon of driver mutation and mutual exclusivity.

5.1.1 Mutation Annotation Format (MAF)

Cancer studies, such as Ciriello *et al.* (2015), made available many data files. One of these files is the Mutation Annotation Format (MAF), a tab-delimited file containing mutations found in samples. Each patient in the study has one or more samples, and each sample has one or more genes with one or more alterations (mutations). The MAF file is frequently used in many analyses and computer methods to discover driver genes and mutual exclusivity (MAYAKONDA; KOEFFLER, 2016; DENG *et al.*, 2019).

The Genomic Data Commons, from the National Cancer Institute (NCI), defines the pipeline used to create the MAF and the 126 columns found in the file[1]. Many of these columns

[1] https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

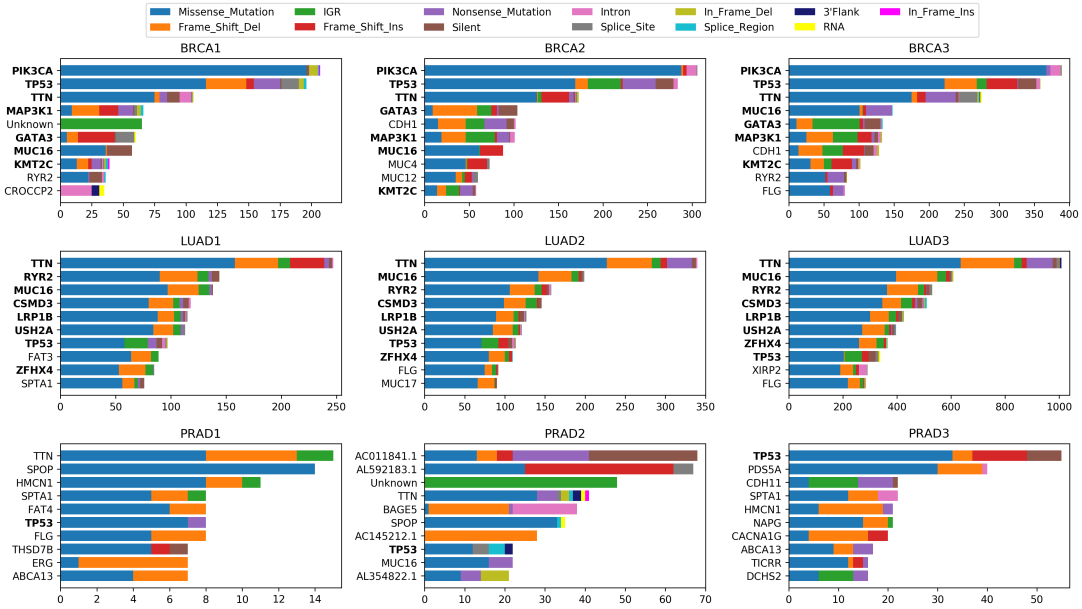
store duplicate information used to be indexed by different systems and databases, as well as some meta-data fields. MafTools is an R package that processes MAF files and offers a graphical overview, performing commonly used analyses in cancer genomics (MAYAKONDA; KOEFFLER, 2016). From the 126 fields present in the file, MafTools only requires 9:

- Hugo Symbol: Gene symbol following the Human Genome Organisation (HUGO) standards.
- Chromosome: The affected chromosome.
- Start Position: The mutation start coordinate.
- End Position: The mutation end coordinate.
- Reference Allele: The strand reference allele, including the deleted sequence for a deletion or "-" for an insertion.
- Tumor Seq Allele: Primary data genotype for tumor sequencing (discovery) allele
- Variant Classification: The translational effect of variant allele, example: Missense, Silent, frameshift deletion.
- Variant Type: The mutation type, example: TNP (tri-nucleotide polymorphism), DNP (di-nucleotide polymorphism), ONP (oligo-nucleotide polymorphism).
- Tumor Sample Barcode: The barcode for the tumor sample. It is a unique identifier to the sample and the patient.

Another tool to analyze MAFs is the cBioPortal[2], with the advantage over MafTools of being online and store many data sets. cBioPortal is an open-access and open-source portal for cancer genomics that aim to lower the barriers between complex genomic data and cancer researchers. They make cancer studies available from TCGA and from the literature, which can be downloaded from the portal. There are online interactive tools to explore data sets, mainly MAF and clinical data files.

Considering the massive amount of cancer data available and the vast number of MAFs for the same type of cancer, we investigated how the mutational characteristics presented in different cancer mutation data sets of the same type of cancer change between studies (RAMOS *et al.*, 2020). The analyses showed that BRCA (a type of breast cancer) and LUAD (a type of lung cancer) have evidence of similarity among their data sets, while PRAD (a type of prostate cancer) is likely heterogeneous. We used three MAFs for each type of cancer and explored the similarities concerning: mutation class, SNV classes, drivers mutations, network comparison

Figure 37 – Top 10 most frequently genes in different MAFs



Source: Elaborated by the author and present in article (RAMOS *et al.*, 2020)

using Mapper (VEEN *et al.*, 2019), and analyses on the top 10 most frequently genes. Figure 37 shows the former analysis.

The top 10 most mutated genes in each MAF are shown with the color representing the mutation type in each gene. Genes within the top 10 of all MAFs in the same cancer type are marked as bold. BRCA has seven common genes in the top 10, while LUAD has eight. Contrary to the previous cancer types, in PRAD only one gene is common.

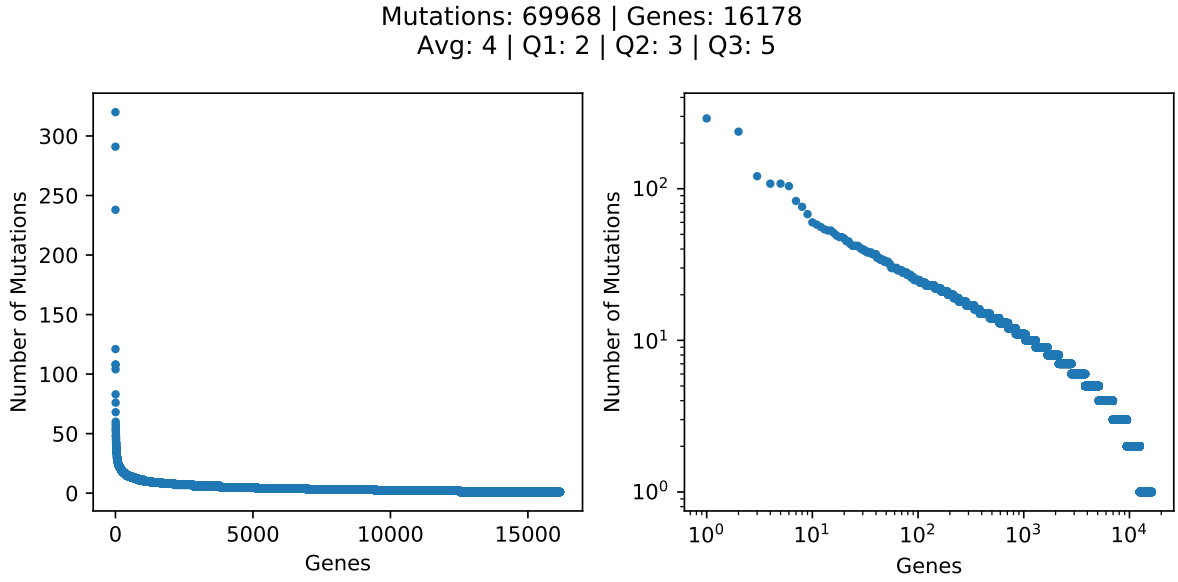
Another important file present in cancer studies is the Clinical Data File. It contains information about the patients concerning age, gender, race, tumor type, and others. The combined analyses of MAF and Clinical Data may categorize the genes' mutations using quantitative and qualitative patient data.

5.2 Heterogeneity: The long tail phenomenon

Cancer is a heterogeneous disease. Patients with similar diagnoses respond differently to treatments and show distinct outcomes. Albeit known at the clinical level, the heterogeneity is poorly understood at the molecular level (ALLISON; SLEDGE, 2014). Genetics studies on large cancer data sets show high variability between tumors, as well as primary and metastasis (TURAJLIC *et al.*, 2019). An exploratory analysis of the MAF from Ciriello *et al.* (2015) shows that few genes are highly frequently mutated, while most genes are seldom altered. Figure 38 shows this distribution that follows a scale-free behavior, known in cancer studies as "long tail" (ARMENIA *et al.*, 2018).

[2] https://www.cbioportal.org/

Figure 38 – Mutations per gene



Source: Elaborated by the author

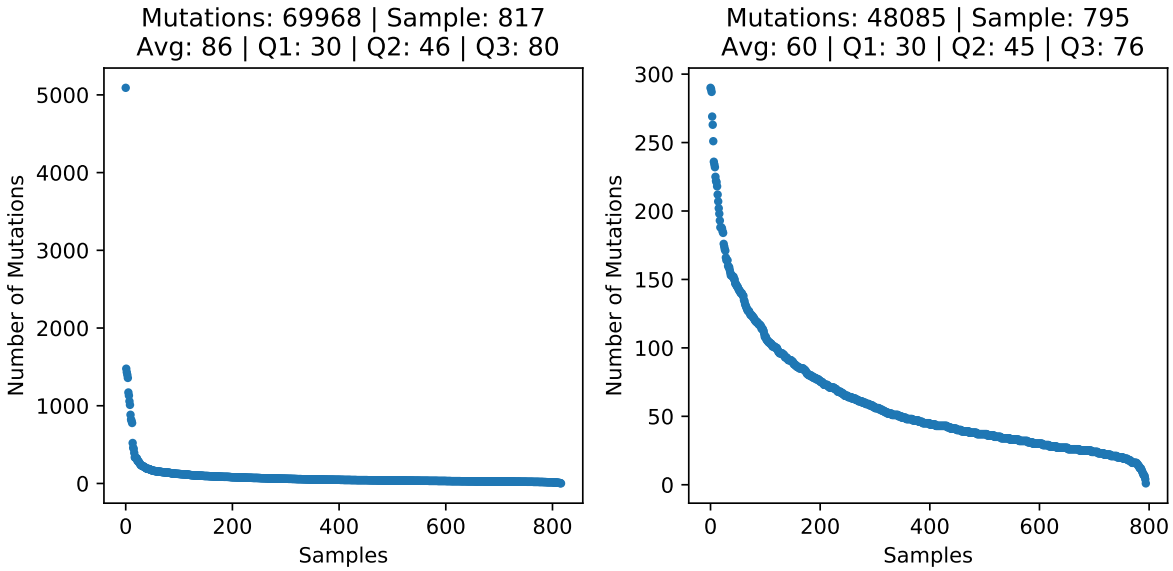
Figure 38 presents the number of mutations per gene in a long-tail perspective, first subplot, and in a log-log, second subplot. The whole MAF has 16,178 genes that sums 69,968 mutations. The average mutation per gene is 4, while the quartiles are 2, 3, and 5. Although the values from the average and quartiles are small, the top five mutated genes and their number of mutations are PIK3CA: 320, TP53: 291, TTN: 238, MUC16: 121, CDH1: 108.

A group of patients with the same type of cancer and similar diagnoses harbor distinct sets of altered genes and number of mutations. The heterogeneity among samples (patients) and number of mutations are show in Figure 39, and the heterogeneity among samples and genes are show in Figure 40.

The MAF has 69,968 mutations across 16,178 genes (as seen in Figure 38) that pertain to 817 samples. The number of mutations per sample is also a long-tail since most samples have few mutations, while the minority of samples have a high number of mutations. A known phenomenon is the presence of "hypermutated samples" (TAMBORERO *et al.*, 2013). The first subplot of Figure 39 shows that the most mutated sample has more than 5,000 mutations, whereas the average is 86 and quartiles are 30, 46, and 80. Following the proposal of Tamborero *et al.* (2013), we remove samples with more than $(Q3 + 4.5 \times IQR)$[3] mutations. The second subplot of Figure 39 shows the impact of these removals. With the elimination of 22 hypermutated samples, the total number of mutations decreased 31%, while the first quartile does not chance and the second quartile drop from 46 to 45, and the average became smaller than the third quartile.

[3] Interquartile range times 4.5 plus the third quartile

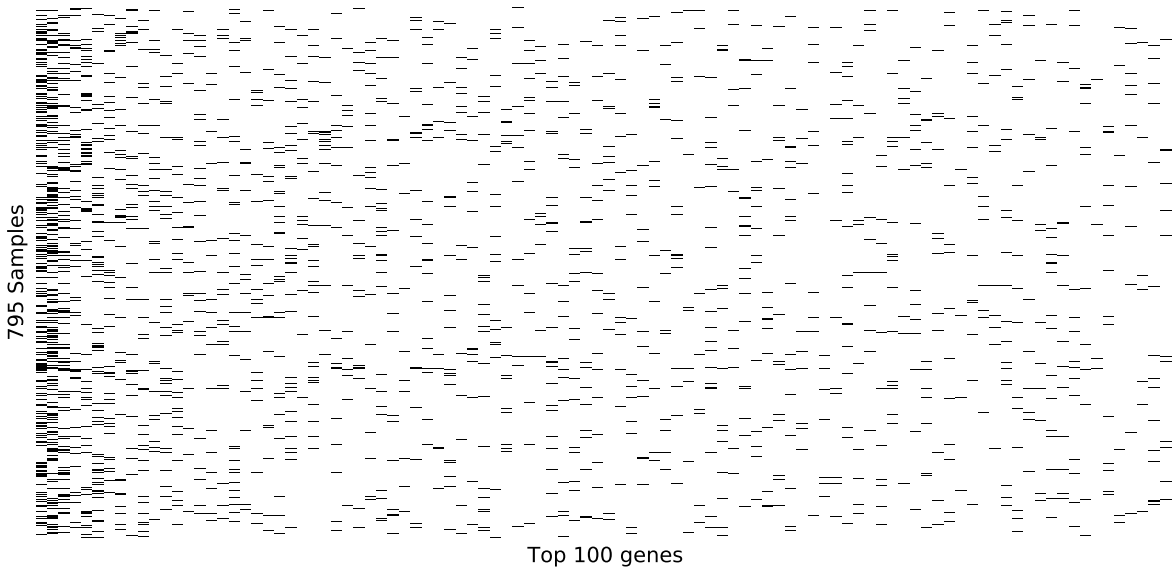
Figure 39 – Mutations per sample



Source: Elaborated by the author

Figure 40 shows the top 100 most frequently genes coverage over the samples without hypermutated. The figure is a heatmap. The first column corresponds to the gene PIK3CA and its presence on 306 samples. Albeit disperse, the long-tail effect found on Figures 38 and 39 can be noticed, as the right side of the heatmap is denser than the left side, and as some lines (samples) have more dots (genes) than others.

Figure 40 – Top 100 genes coverage over 795 samples



Source: Elaborated by the author

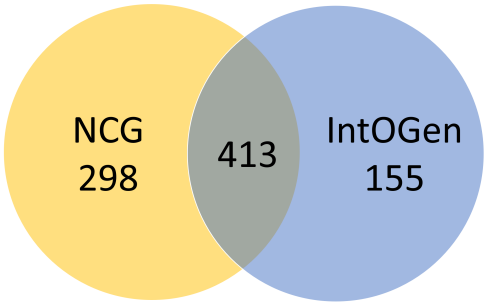
Finding mutational patterns within this heterogeneous scenario is a problem that computational approaches have sought to resolve. Among the contributions of computing to cancer genetics, it is worth highlighting the search for drivers genes and mutual exclusivity sets.

5.3 Cancer Driver Genes

Cancer tumors have many mutations, but only a small portion contributes to oncogenesis and tumor progression. Therefore, cancer mutations are categorized as passenger or driver. Contrary to passenger, driver mutations have a direct impact on oncogenesis, conferring a growth advantage to the cell (STRATTON; CAMPBELL; FUTREAL, 2009; RAMOS *et al.*, 2021).

The study of cancer driver genes contributes to understanding the origin and development of the disease. Considering the high heterogenesis in cancer data, the discovery of driver genes is an open challenge in cancer genomics, often explored with computational aid. It is known that some driver genes are commonly involved in different types of cancer (NUSSINOV *et al.*, 2019), and they tend to be mutual exclusive (PULIDO-TAMAYO *et al.*, 2016). Two databases that make avaliable lists of cancer driver genes[4] are NCG (REPANA *et al.*, 2019), and IntOGen (MARTINEZ-JIMENEZ *et al.*, 2020). NCG provides a list with 711 genes, while IntOGen has 568. The union of both databases sums 866 driver genes. Figure 41 shows the intersection between the two databases.

Figure 41 – Cancer driver genes databases intersection.



Source: Elaborated by the author

5.3.1 Computational Methods

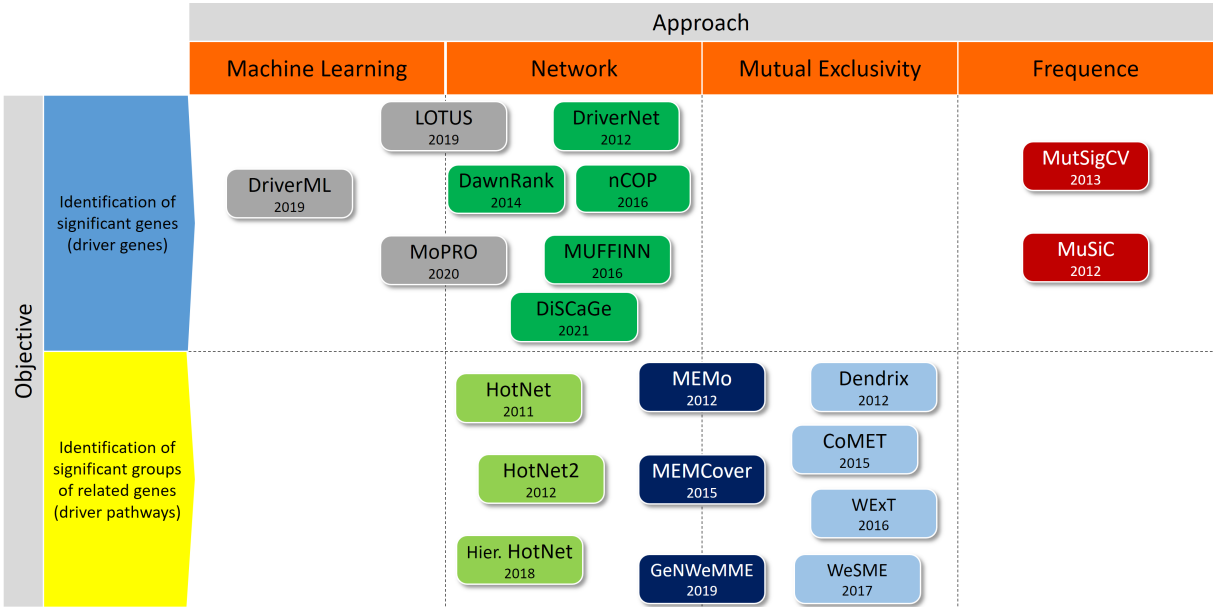
Cutigi, Evangelista and Simao (2020) review approaches for the identification of driver mutations in cancer, where they categorized many computational methods. In its Ph.D. thesis, the author updated the categorization with more methods. Figure 42 shows a adaptation from a figure present in Cutigi (2021). The methods are categorized by what they aim to find: driver genes or driver pathways, and the approach used to reach the objective. Twelve of the twenty

[4] http://ncg.kcl.ac.uk/ and https://www.intogen.org/

methods use PPINs in their pipelines. Six methods search for mutual exclusivity genes to that find driver pathways.

Analyzing the methods' publication dates in Figure 42, we notice that the frequency approach was only used in the beginning. The previous section shows that the data frequency is highly heterogenous and follows a scale-free distribution. Methods that rely on frequency are not efficient in finding drivers on the distribution "tail". On the other hand, the adoption of Machine Learning approaches began only in 2019. The use of PPINs and mutual exclusivity has persisted over time, with many methods evolving over past contributions. HotNet, HotNet 2, and Hirarquical HotNet are a family of methods, the same way that Dendrix, CoMET, and WeXT also are. The six previously mentioned methods belong to the same research group, named Raphael Lab at Princeton University.

Figure 42 – Computational methods for the identifications of cancer driver genes.



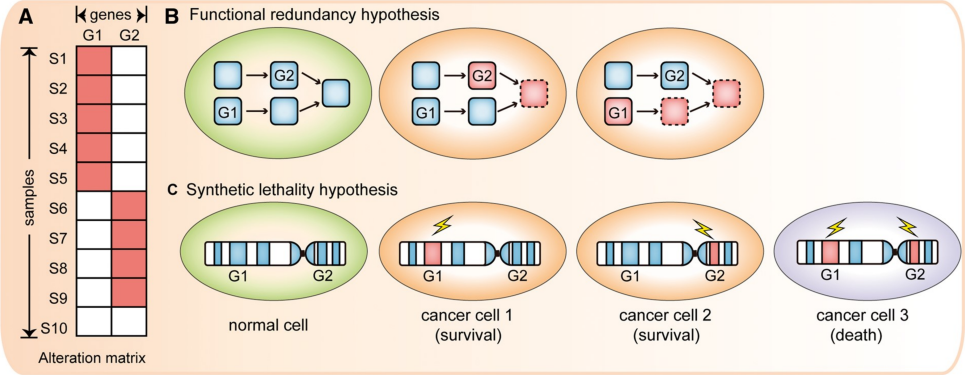
Source: Adapted from the work of Cutigi (2021)

5.4 Mutual Exclusivity

In their work Cisowski and Bergo (2017) explains that cancer is driven by mutations in pathways associated with cell proliferation and survival, and it is intuitive to think the more mutations a tumor has, the faster it progresses. However, large-scale genomics studies show otherwise: driver oncogenes often are mutually exclusive. The authors also say that although this phenomenon is not entirely understood, novel reports indicate mutual exclusivity may be associated with tumor type and interactions between drivers' genes. Ding *et al.* (2020) explore how mutual exclusive and co-occurring genes in six types of cancer can significantly divide samples into groups based on clinical data as age, gender, histological type, and pathologic stage.

The two main hypotheses for mutual exclusivity are Functional Redundancy and Synthetic Lethality (DENG *et al.*, 2019). Functional Redundancy Hypothesis is based on pathway topology and the downstream effect: mutation in one gene on the stream is enough to corrupt the entire pathway; thus, there is no selective pressure for mutations in other genes (BABUR *et al.*, 2015). Following this hypothesis, identifying mutated genes that corrupted the pathway can help understand which biological function (pathway) corruption leads to cancer. Synthetic Lethality Hypothesis states that simultaneous mutation in a pair of genes leads to cell death, thus these genes do not co-occur in the MAF. Synthetic Lethality can be used in cancer therapy, which is the case of the genes KRAS and EGFR in lung adenocarcinomas (DENG *et al.*, 2019). Figure 43 exemplify both hypotheses.

Figure 43 – Mutual Exclusivity Hypotheses



Source: Extracted from the work of Deng *et al.* (2019)

Figure 43 A shows the alteration matrix with ten samples and two genes. If we consider that each gene is equally likely to be mutated, which is not the case (LEISERSON; REYNA; RAPHAEL, 2016; MANESCU; KEICH, 2016), there is a probability of 50% that a sample is mutated in G1, and 40% in G2, therefore the joint probability of G1 and G2 is 20%. Since the expected number of samples with mutations in G1 and G2 simultaneously is two, but we find none, we see evidence that G1 and G2 are mutually exclusive.

5.4.1 Significant Exclusivity

In a breast cancer MAF with 816 samples, 288 (35%) have mutations on gene TP53, 308 (38%) on PIK3CA, and 141 (17%) on TTN. The expected number of samples with mutations on TP53 and TTN is 0.35 x 0.17 = 0.06 x 816 = 49, but we find 65 samples, indicating co-occurrence. Following the same logic, for TP53 and PIK3CA we expected 109 samples but found 90, indicating mutual exclusivity. For PIK3CA and TTN we expected 53 samples but found 58, indicating a weak co-occurrence. The cBioPortal uses Odds Ratio and Fisher Exact Test to measure the observed association strength and statistical significance. Figure 44 shows the above example using cBioPortal Mutual Exclusivity tool.

Figure 44 – cBioPortal exclusivity for three genes on Breast Invasive Carcinoma (TCGA, Cell 2015)

A	B	Neither	A Not B	B Not A	Both	Log2 OR	p-Value	q-Value ▲	Tendency
TP53	**TTN**	452	223	76	65	0.794	0.002	**0.004**	**Co-occurrence**
TP53	**PIK3CA**	310	198	218	90	-0.630	0.003	**0.004**	**Mutual exclusivity**
PIK3CA	**TTN**	425	250	83	58	0.248	0.206	0.206	Co-occurrence

Source: Extracted from cBioPortal (GAO *et al.*, 2013)

For a given set of genes and a MAF, cBioPortal Mutual Exclusivity tool creates all possible combinations pairs, and for each pair presents: the genes name (fields A and B); total of samples that does not have mutation on A or B (Neither); total of samples only mutated on gene A (A not B); total of samples only mutated on gene B (B not A); total of samples with mutation in A and B (Both); The association strength using log_2 Odds Ratio (Log2 OR), where negative values indicate exclusivity and positive values Co-occurrence; the significance of the association strength using one-sided Fisher Exact Test (p-value); a p-value correction using Benjamini-Hochberg FDR (q-value); and the association tendency the pair of genes have of being mutually exclusive or co-occurrent (Tendency). As mentioned before, PIK3CA and TTN have a weak co-occurrence. Figure 44 shows an odds ratio close to zero and a high p-value indicating that although this pair's observed intersection is different from the expected, it is not significant.

Using the MAF from (CIRIELLO *et al.*, 2015) and 55 genes, the cBioPortal Mutual Exclusivity tool generates 1485 pairs of genes, with 519 showing mutual exclusivity tendency and 966 showing co-occurrence tendency. Keeping only significant associations (p-value < 0.05) the mutual exclusive pairs drop from 519 to 6, and the co-occurrence drop from 966 to 46. Overall, only 52 (3.5%) of all pairs show statistical significant association.

5.4.2 Computational Methods

Deng *et al.* (2019) did a review on 21 computational methods developed to discover mutual exclusivity. Although these methods vary on the approaches and evolved with time (the review cover methods from 2011 to 2016), the authors separate them into two groups based on the input data: ***De novo***, that only use genomic data (mainly MAF), and **Knowledge-Based**, that integrate data from pathways, PPINs and/or phenotype data. Extending the analysis made by Deng *et al.* (2019), we present Table 5, with information about the input and output used on the 21 computational methods. Reviewing the methods' articles is not always clear the pipeline used, and there are nuances. For example, some methods remove the hypermutated samples, and others do not. There are also differences in the PPIN and Pathways database used. Table 5 offers an overview to categorize and compare the methods from a high-level perspective.

From the 21 methods, 19 use the Mutation Matrix as an input. Derived from the MAF, in the Mutation Matrix samples are rows and genes are columns, Figure 43-A shows a muta-

tion/alteration matrix. Since the MAF came directly from cancer studies, the Mutation Matrix offers a *De novo* view on genes' mutations co-occurrence and exclusivity, making possible an unbiased analysis from using previous knowledge. All *De novo* methods have an output of genes set or network modules. The three methods that use networks create it by establishing pairwise mutual exclusivity between genes and using significant pairs as network edges. The output in these methods are cliques or modules that represent a group of mutually exclusive genes. Knowledge-based methods also use the Mutation Matrix as input, but combine it with other inputs. Using the hypothesis that drivers genes tend not to co-occur, some knowledge-based methods try to find new interactions and pathways or associate interactions and pathways to cancer. It is worth mentioning that some methods search for exclusivity aiming to find drivers genes and driver pathways, which is, for example, the case of Dendrix, Mult-Dendrix, CoMET, WExT, WeSME, SAGA, and SSA-ME.

Given the variety of methods, we analyzed how each one validates results. Table 6 presents the same 21 computational methods from Table 5 and their approach to validate results, the first ten are *De Novo* and the last eleven are Knowledge-Based. Excluding SAGA and SSA-ME, all other methods apply some statistical significance test. Six of the seven earlier methods used the Fisher Exact Test or similar approaches (Chi-Square or G-Test for examples). More recent methods point to the limitation of these tests and compare the results to random sampling to calculate p-values. CoMET, albeit using Fisher Exact Test, weights the test with gene frequency to diminish the importance of high frequent genes. MutExSL creates a p-value that encompasses the Synthetic Lethality Hypothesis. WExT condition the number of mutated samples with mutation probabilities from per-event and per-sample. It also considers the length of the genes in the significance test. WeSME main feature is a fast method to estimate p-values, since earlier methods frequently rely on permutation tests, which are computationally infeasible. Knowledge-Based methods mainly develop unique ways to statistical test the significance of results. REVEALER uses many biological function information to validate the mutual exclusive genes sets. As Knowledge-Based methods use different types of input data, they often use synthetic input data that maintain statistical characteristics of the real data to extract statistical significance. SAGA does not develop a significance test, just compare the results with other methods. SSA-ME is a reinforce machine learning method that validates the result by it models accuracy.

The Result Comparison part of Table 6 shows that recent methods tend to compare their result with previous methods. Mult-Dendrix is an evolution of Dendrix, and WExT is an evolution of CoMET. These four methods belong to the same research group[5]. Some methods do not compare results because the pipeline and output significantly differ. Thereby they rely only on the Significance Test. Overall, the *De Novo* methods are more similar in the pipeline, but the Knowledge-Based methods have mostly unique approaches.

[5] http://compbio.cs.brown.edu/software/

Table 5 – Mutual exclusivity methods: Input and Output

-		INPUT						OUTPUT		
Type	Year	Methods	Mutation Matrix	Gene Expression	Pathway	PPIN	Others	Set of Genes	Network Modules/Cliques	Driver Pathways
De novo	2011	RME	X						X	
De novo	2012	Dendrix	X					X		
De novo	2013	Mult-Dendrix	X					X		
De novo	2014	MutExSL	X	X				X		
De novo	2015	GAMToc	X						X	
De novo	2015	CoMET	X					X		
De novo	2016	TiMEx	X						X	
De novo	2016	MEGSA	X					X		
De novo	2016	WeSME	X					X		
De novo	2016	WExT	X					X		
Knowledge based	2012	MEMo	X		X	X			X	
Knowledge based	2012	MDPFinder	X	X	X					X
Knowledge based	2013	iMCMC	X	X			X		X	
Knowledge based	2014	DAISY		X			X	X		
Knowledge based	2014	SAGA	X	X	X					X
Knowledge based	2015	MEMCover	X		X	X				X
Knowledge based	2015	Mutex	X		X					X
Knowledge based	2016	DISCOVER	X		X	X		X		
Knowledge based	2016	SSA-ME	X			X			X	
Knowledge based	2016	REVEALER					X	X		
Knowledge based	2016	C3	X		X			X		
Total:			19	5	7	4	3	11	6	4

Deng *et al.* (2019) end their review on computational methods for mutual exclusivity pointing the confounders of mutual exclusivity. The authors address key features that can cause false-positive discovery: cancer sub-type, intra-tumor heterogeneity, and imbalance of mutual exclusivity. A group of genes can be mutually exclusive in a sub-type of cancer, while it can be dissociated or co-occurring in others. When the method analyzes the entire MAF, this group of genes probably will not have a significant association without considering sub-types. The authors suggest grouping the samples by cancer-type, but indicate that not all methods can perform well with a small group of samples. The imbalance of mutual exclusivity refers to the long-tail effect. Few genes have a high sample cover and may disrupt the method statistics. It is important to weigh the genes coverage as well the cancer sub-types. The intra-tumor heterogeneity confounder arises when the method analyzes the entire MAF. The MAF represents all the mutations found in the tumor, while studies have identified mutual exclusivity at the single-cell level.

Table 6 – Mutual exclusivity methods: Validation

Methods		Significance Test		Results Comparison							
		Fisher and similars	Others	RME	Dendrix	Mult-Dendrix	Mutex	MEMo	CoMET	TiMEx	MDPFinder
2011	RME	X									
2012	Dendrix	X									
2013	Mult-Dendrix	X		X	X						
2014	MutExSL		X								
2015	GAMToc	X									
2015	CoMET	X			X	X	X	X			
2016	TiMEx	X				X					
2016	MEGSA		X		X		X			X	
2016	WeSME		X								
2016	WExT		X						X		
2012	MEMo		X								
2012	MDPFinder		X								
2013	iMCMC		X								
2014	DAISY		X								
2014	SAGA				X						X
2015	MEMCover		X								
2015	Mutex	X		X	X	X		X			X
2016	DISCOVER		X				X	X	X	X	
2016	SSA-ME										
2016	REVEALER		X								
2016	C3		X						X		

5.5 Final Considerations

The large amount of data from NGS allowed the identification of cancer heterogeneity, which was already known at the clinical level, also at the molecular level. Using genomic files, like the MAF, PPINs, and pathways, computation aim to find patterns in this complex mutational scenario. The search for cancer driver genes and mutually exclusive sets of genes are areas that rely on computational methods. Many of these methods evolved over time, which is the case of the HotNet and Dendrix families. New methods seek to improve over previous attempts by comparing results, considering new information (gene length, network topology, tumor information, etc.), and statically measuring the significance of results.

CHAPTER

6

PROJECT PROPOSAL

6.1 Contextualization

Cancer is a complex disease characterized by genetic mutations that happen in a cell and lead to uncontrolled growth and division. NGS's great volume of data reveals a high heterogeneous scenario, where patients harbor different sets of mutated genes, even in the same type of cancer. A challenge in cancer genomics is the discovery of patterns that can explain the initiation and evolution of the disease, thus enabling personalized therapies. Computational methods have helped in this task, such as searching for cancer driver genes and mutually exclusive gene sets.

Only few of the many mutations a tumor undergoes are responsible for oncogenesis and tumor progression, which are known as driver mutations, and the associated genes as driver genes. A phenomenon related to cancer heterogeneity is the presence of mutually exclusive genes: genes that, albeit frequent, are statistically disassociated in the same tumor. Since many exclusive genes are also drivers, computational methods have used driver information to find mutually exclusive genes and mutually exclusive information to find drivers.

Computational methods have evolved over time, improving previous approaches and adding information from different databases. In addition to genomic data, the methods also use information from PPINs and pathways.

6.2 Motivation, Hypothesis and Objectives

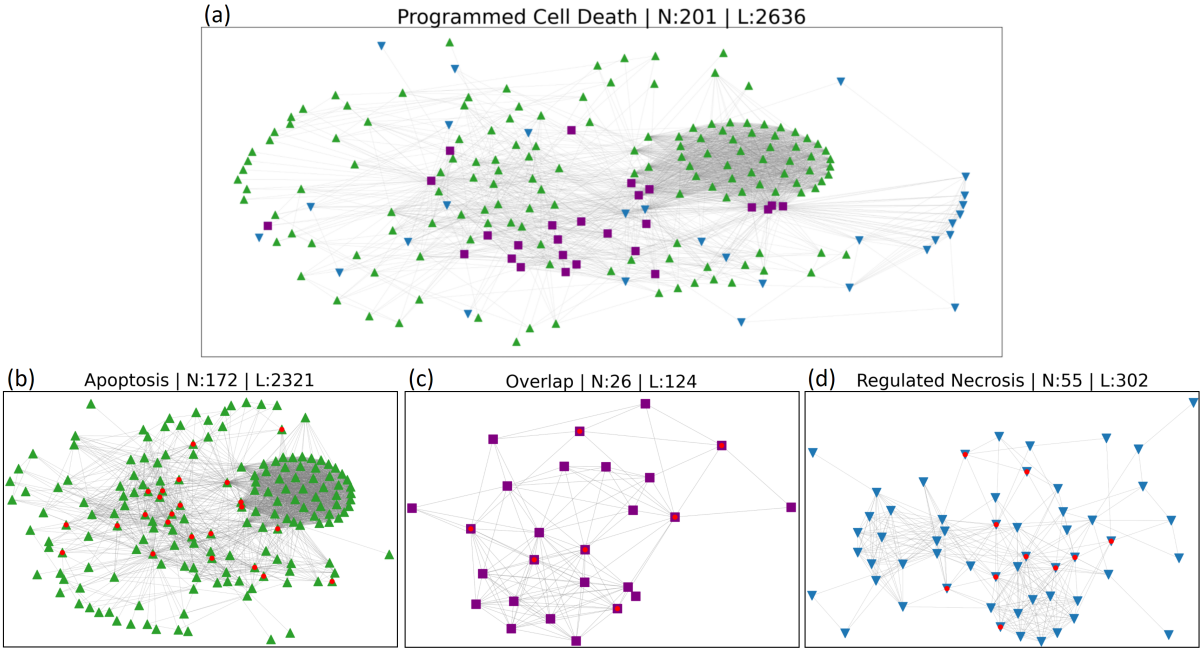
6.2.1 Motivation

Although there are plenty of computational methods that search for drivers and mutual exclusivity genes, they often do not consider qualitative data from patients and tumors. The

methods that use networks or pathways frequently look for neighbors or restrict the topological analyses to degree or shortest path.

The Reactome's Super Pathways group more than 1,800 pathways into 26 main cellular functions. The grouping of pathways follows a tree-like structure, where smaller and very specific pathways are leaves, and the 26 Super Pathways are roots. The Programmed Cell Death Super Pathway groups a total of 42 pathways. None of the methods use Super Pathways or the topological characteristics of sub-pathways inside a larger pathway. Figure 45 shows the hierarchy for a root pathway (Programmed Cell Death) and the two first-level pathways (Apoptosis and Regulated Necrosis).

Figure 45 – Programmed Cell Death Hierarchy



Source: Elaborated by the author.

The Figure 45(a) shows the Programmed Cell Death Super Pathway Network, where green nodes represent the genes associated with the Apoptosis pathway and the blue ones to the Regulated Necrosis pathway. Each of these pathways also have sub-pathways, meaning that it is possible to extract 42 sub-networks from Programmed Cell Death. There are intersections among sub-pathways. In the plot, genes associated with Apoptosis and Regulated Necrosis are represented as purple. The Figures 45(b), (c), and (d) are sub-pathways networks and the intersection between them, nodes with red marks are known cancer driver genes. The topological characterization of drivers and non-drivers nodes may be used to discover new drivers.

The Super Pathways and their sub-pathways are smaller subsets of whole PPINs and bring biological function to associated genes. Since the mutual exclusivity phenomenon is characterized by the disassociation of drivers genes on pathways in the same sample, the topological features

of Super Pathways Networks and their sub-networks may be used to study the "Functional redundancy mutual exclusivity hypothesis".

6.2.2 Hypothesis

The main goal of this proposal is to investigate the following hypothesis: combining the topological characterization of Super Pathways Networks enriched with driver genes information, in conjunction with quantitative data from MAF and qualitative data from patients, it is possible to find new driver pathways through mutual exclusivity and new driver genes through topological similarity.

6.2.3 Objectives

This project's general objective is to discover hidden patterns in cancer genomics data sets using topological characteristics of Super Pathways Networks and the topological role of known driver genes. The discovery of these patterns leads to two specific objectives:

- Finding driver pathways through mutual exclusivity combining quantitative and qualitative data from NGS with Supper Pathways Networks topology and driver genes.
- Finding new driver genes by searching for topological similarities in Super Pathways Networks enriched with driver information. This objective is a continuation of the approach developed in Ramos *et al.* (2021)

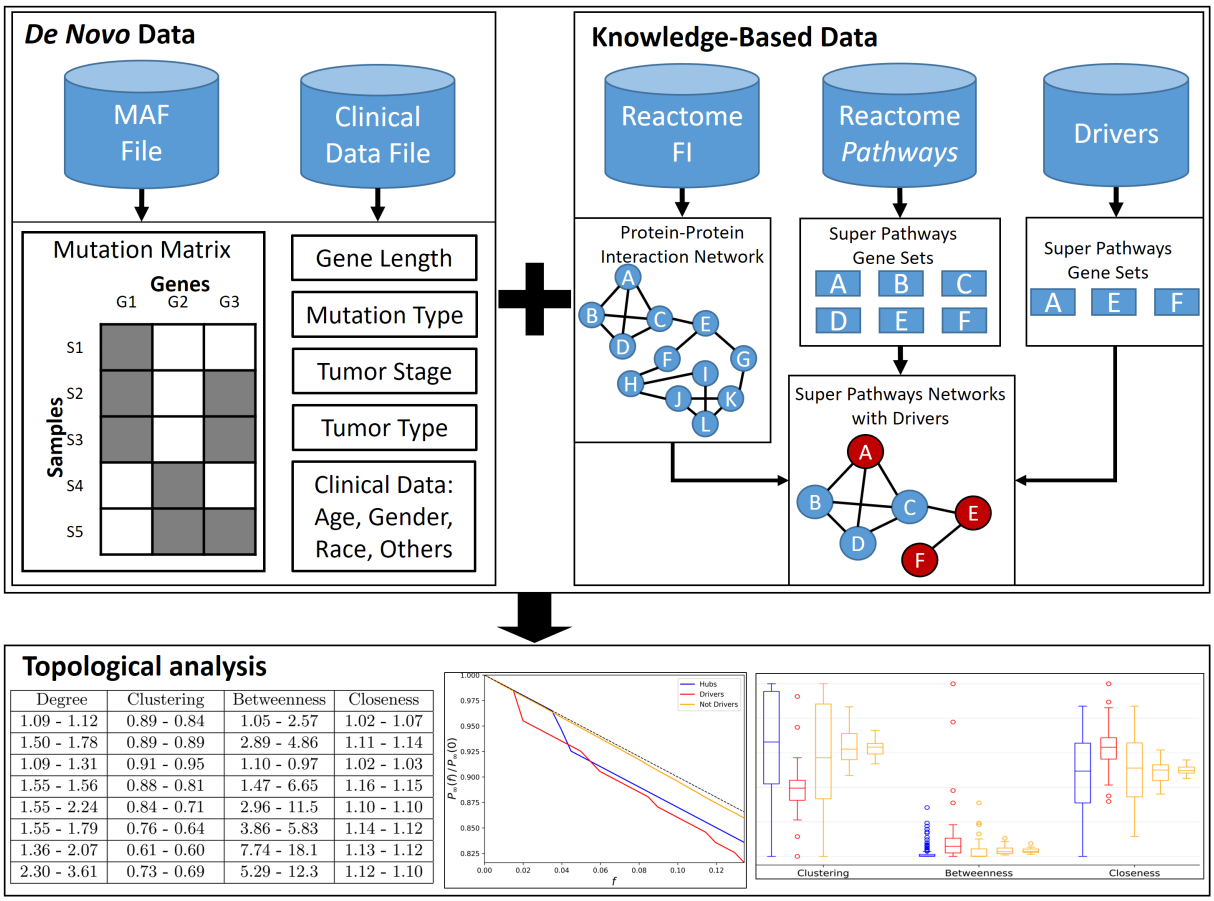
6.3 Methodology

To achieve the objectives of this thesis proposal, we will execute new and classic methods that search for drivers (Figure 42) and mutual exclusive genes (Tables 5 and 6) in order to deepen our understand and develop a way to compare and test results. The *De Novo* methods Dendrix, Mult-Dendrix, CoMET, WExT and WeSME, search for driver genes using mutual exclusivity. The knowledge-based methods MDPFinder, SAGA, MEMCover, and Mutex combine data from MAF files and pathways to find driver pathways. The execution of these nine methods can reveal how the approaches evolved and how to establish a combined pipeline for results evaluation.

After executing methods and defining a process to evaluate results, we intend to develop a computational method that contemplates the objectives of this thesis. To do so, we will follow the pipeline described in Figure 46. We join data from five different fonts, two *De Novo* and three Knowledge-Based. The MAF file is used to generate the Mutation Matrix and extract information about gene length, and mutation type. The Clinical Data file contains information about tumor stage and type and patient data like age, gender, race, and others. To create Super Pathways Networks, we extract induced subgraphs from the Reactome FI PPIN using the genes sets from

Reactome Super Pathways database. The networks are enriched with driver information from bases like NCG and IntOGen, thus creating Super Pathways Networks with Drivers.

Figure 46 – Pipeline



Source: Elaborated by the author.

The Knowledge-Based part of the pipeline was developed and used in Ramos *et al.* (2021), where we compare de topological differences between drivers and non-drivers on seven Super Pathways Networks. The MAF file was explored in Ramos *et al.* (2020). The combination of *De Novo* and Knowledge-Based will be used to achieve the two specific objectives of this thesis proposal: find driver pathways and drivers genes.

Finding driver pathways: We will look for driver pathways searching for mutual exclusivity pathways and sub-pathways. Following the procedure used in Figure 45, we will enrich the Super Pathways Networks and their sub-pathways network with driver information. Using per-sample mutation information, we will trace a mutation path within the Super Pathways hierarchy and significantly associate sample mutation to pathways and search for mutual exclusivity among drivers. We expect a qualitative categorization of samples to lead to groups of mutual exclusivity drivers pathways. We also expect that the topological categorization of drivers and sample's genes will aid the significance of driver pathways. **Validation:** The association of gene set and pathways is well explored in "functional enrichment methods" (Section 4.3). We can

use these methods to develop our method and compare results. The resulting driver pathway significance can be statically measured and compared to similar methods.

Finding driver genes: We will use the topological differences found between drivers and non-drivers in Ramos *et al.* (2021) to search for topological patterns among known drivers genes in Super Pathways Networks. We will compare drivers' patterns with topological measures found in candidate cancer drivers from databases (NCG and IntOGen) in order to discover new drivers. We can also compare the drivers' patterns with the output of drivers methods. **Validation:** we will replicate the validation make by Cutigi *et al.* (2021), where they use an automated literature-based analysis with CancerMine[1] tool.

6.3.1 Activity Timetable

This section lists the main activities performed and expected for the proposed research. The activity schedule is presented in Table 7. The activities are described below:

1. **Conclusion of courses:** the courses taken and concluded were: Artificial Intelligence I, Artificial Intelligence II, Biostatistics, Pedagogical Preparation, Research Methodology, Software Testing and Validation. Besides these courses, the candidate was a listener student in Complex Networks course.

2. **Literature review:** we did a literature review in order to define a relevant and feasible problem to work in this Ph.D. project. The review will continue throughout the project, as similar methods and approaches may emerge.

3. **Study of Cancer's Data Sets:** databases as TGCA and cBioPortal were analyzed, and from their many data sets, we selected the MAF and Clinical Data files.

4. **Study of PPINs and Pathways:** with help from the results found in the literature review, PPINs and Pathways databases were studied.

5. **Writing the Qualification Exam:** time dedicated to writing this project. Many of the analyses presented in this project were made during the two previous activities.

6. **Finding Drivers Topological Patterns:** in order to achieve one of the specific objectives, we will search for topological patterns among drivers in Super Pathways Networks to find new driver genes.

7. **Execution of Mutual Exclusivity Methods:** following the proposed methodology, we will deepen our knowledge about new and classical methods for the discovery of mutual exclusivity in mutated genes.

[1] http://bionlp.bcgsc.ca/cancermine/

8. **Finding Driver Pathways:** in order to achieve one of the specific objectives, we will search for driver pathways using topological features and mutual exclusivity.

9. **Writing Reports:** the first mark comprises the report Ramos *et al.* (2021). The other two are reports we will make regarding this proposal objectives results.

10. **Writing Thesis:** time that will dedicate to writing the Ph.D. thesis.

	2020	2021		2022		2023		2024
	2º S	1º S	2º S	1º S	2º S	1º S	2º S	1º S
1. Conclusion of Courses								
2. Literature review								
3. Study of Cancer Studies Data Sets								
4. Study of PPINs and Pathways								
5. Writing the Qualification Exam								
6. Finding Drivers Topological Patterns								
7. Execution of Mutual Exclusivity Methods								
8. Finding Driver Pathways								
9. Writing Reports								
10. Writing Thesis								

Table 7 – Activity Timetable.

6.3.2 Expected Results

We expect this Ph.D. project contributes to Cancer Genomics, especially in using PPINs for cancer studies. Moreover, we expect this project to continue collaborating with Barretos Cancer Hospital, which has already resulted in co-authoring six publications with the research group of which the Ph.D. candidate participates.

In order to disseminate the outcomes of this Ph.D. project, we intend to continue publishing and participating on two Brazilian bioinformatics conferences: *Simpósio Brasileiro de Computação Aplicada à Saúde* (SBCAS) and *Brazilian Symposium on Bioinformatics* (BSB). We also seek to publish this project's final results in scientific journals.

6.3.3 Performed Activities

In addition to the studies that compose the writing of this document, we have performed other activities, which are presented below:

- **Research papers:** the Ph.D. candidate wrote and presented two papers as first author and one as a collaborator:

1. A full paper published at the *Simpósio Brasileiro de Computação Aplicada à Saúde* (SBCAS 2020) (RAMOS *et al.*, 2020). This article predates the official entry to the Ph.D. program and explores the mutations similarities and differences between MAFs from three types of cancers.
2. A full paper published at the Brazilian Symposium on Bioinformatics (BSB 2020) (CUTIGI *et al.*, 2020), which presents a machine learning method to classify candidate cancer drivers as real drivers or false drivers.
3. A full paper published at the Brazilian Symposium on Bioinformatics (BSB 2021) (RAMOS *et al.*, 2021), which topological differentiate drivers and non-driver in Super Pathways Networks.

- **Scientific initiation advisor**: the Ph.D. candidate advised three scientific initiations in Complex Networks between 2020 and 2021. The latter explored the resilience of nine PPINs from resistant bacteria and won second place in the *ComunICa IFSP* contest.

- **Talk about Complex Networks**: the Ph.D. candidate gave a speech about Complex Networks in the *II Semana Acadêmica de Computação (II SeComp)* at *Universidade Tecnológica Federal do Paraná - Campus Apucarana* in 2020/12.

- **Code repository about Complex Networks:** Durant the studies of this proposal, the Ph.D. candidate elaborated a code repository containing fundamentals about Graph Theory, Complex Networks, and network visualization. Repository link: <https://github.com/RodrigoHenriqueRamos/Complex-Networks>

- **Other types of topological characterization:** With members of the research team, the Ph.D. candidate studied others methods to topological characterize data. The Scikit-tda groups Topological Data Analysis Python libraries (SAUL; TRALIE, 2019), including Persistant Homology and Mapper. Persistant Homology can detect global structural features in the network that persist across multiply interactions, like holes or strong connect components. Mapper is used to visualize the topological structures in high-dimensional data point cloud data. The mapper can also be used to compare data sets, as used in Ramos *et al.* (2020).

BIBLIOGRAPHY

ALBERT, R.; JEONG, H.; BARABÁSI, A.-L. Error and attack tolerance of complex networks. **nature**, Nature Publishing Group, v. 406, n. 6794, p. 378–382, 2000. Citations on pages 44, 56, and 68.

ALLISON, K. H.; SLEDGE, G. W. Heterogeneity and cancer. **Oncology**, UBM LLC, v. 28, n. 9, p. 772–772, 2014. Citations on pages 19 and 73.

ALSTOTT, J.; BULLMORE, E.; PLENZ, D. powerlaw: a python package for analysis of heavy-tailed distributions. **PloS one**, Public Library of Science San Francisco, USA, v. 9, n. 1, p. e85777, 2014. Citations on pages 28, 30, and 31.

ARMENIA, J.; WANKOWICZ, S. A.; LIU, D.; GAO, J.; KUNDRA, R.; REZNIK, E.; CHATILA, W. K.; CHAKRAVARTY, D.; HAN, G. C.; COLEMAN, I. *et al.* The long tail of oncogenic drivers in prostate cancer. **Nature genetics**, Nature Publishing Group, v. 50, n. 5, p. 645–651, 2018. Citation on page 73.

BABUR, Ö.; GÖNEN, M.; AKSOY, B. A.; SCHULTZ, N.; CIRIELLO, G.; SANDER, C.; DEMIR, E. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. **Genome biology**, Springer, v. 16, n. 1, p. 1–10, 2015. Citation on page 78.

BARABÁSI, A.-L. **Network science**. Cambridge: Cambridge University Press, 2015. Citations on pages 23, 24, 27, 28, 29, 30, 37, 39, 42, 52, 53, 54, 55, 56, and 63.

BARABASI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. **Nature reviews genetics**, Nature Publishing Group, v. 5, n. 2, p. 101–113, 2004. Citation on page 62.

CANISIUS, S.; MARTENS, J. W.; WESSELS, L. F. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. **Genome biology**, BioMed Central, v. 17, n. 1, p. 1–17, 2016. Citation on page 62.

CIRIELLO, G.; GATZA, M. L.; BECK, A. H.; WILKERSON, M. D.; RHIE, S. K.; PASTORE, A.; ZHANG, H.; MCLELLAN, M.; YAU, C.; KANDOTH, C. *et al.* Comprehensive molecular portraits of invasive lobular breast cancer. **Cell**, Elsevier, v. 163, n. 2, p. 506–519, 2015. Citations on pages 61, 62, 63, 71, 73, and 79.

CISOWSKI, J.; BERGO, M. O. What makes oncogenes mutually exclusive? **Small GTPases**, Taylor & Francis, v. 8, n. 3, p. 187–192, 2017. Citations on pages 19 and 77.

CLAUSET, A.; NEWMAN, M. E.; MOORE, C. Finding community structure in very large networks. **Physical review E**, APS, v. 70, n. 6, p. 066111, 2004. Citations on pages 42 and 55.

CUTIGI, J. F. **Computational approaches for the discovery of significant genes in cancer**. Phd Thesis (PhD Thesis) — Universidade de São Paulo, 2021. Citations on pages 76 and 77.

CUTIGI, J. F.; EVANGELISTA, A. F.; REIS, R. M.; SIMAO, A. A computational approach for the discovery of significant cancer genes by weighted mutation and asymmetric spreading strength in networks. **Scientific Reports**, Nature Publishing Group, v. 11, n. 1, p. 1–10, 2021. Citation on page 89.

CUTIGI, J. F.; EVANGELISTA, A. F.; SIMAO, A. Approaches for the identification of driver mutations in cancer: a tutorial from a computational perspective. **Journal of Bioinformatics and Computational Biology**, World Scientific, v. 18, n. 03, p. 2050016, 2020. Citation on page 76.

CUTIGI, J. F.; EVANGELISTA, R. F.; RAMOS, R. H.; FERREIRA, C. d. O. L.; EVANGELISTA, A. F.; CARVALHO, A. C. de; SIMAO, A. Combining mutation and gene network data in a machine learning approach for false-positive cancer driver gene discovery. In: SPRINGER. **Brazilian Symposium on Bioinformatics**. [S.l.], 2020. p. 81–92. Citation on page 91.

DAUM, H.; PERETZ, T.; LAUFER, N. Brca mutations and reproduction. **Fertility and sterility**, Elsevier, v. 109, n. 1, p. 33–38, 2018. Citation on page 65.

DEMKOW, U.; PLOSKI, R. **Clinical applications for next-generation sequencing**. [S.l.]: Academic Press, 2015. Citations on pages 19 and 71.

DENG, Y.; LUO, S.; DENG, C.; LUO, T.; YIN, W.; ZHANG, H.; ZHANG, Y.; ZHANG, X.; LAN, Y.; PING, Y. *et al.* Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. **Briefings in Bioinformatics**, Oxford University Press, v. 20, n. 1, p. 254–266, 2019. Citations on pages 20, 71, 78, 79, and 82.

DING, W.; FENG, G.; HU, Y.; CHEN, G.; SHI, T. Co-occurrence and mutual exclusivity analysis of dna methylation reveals distinct subtypes in multiple cancers. **Frontiers in cell and developmental biology**, Frontiers, v. 8, p. 20, 2020. Citations on pages 19, 20, and 77.

FAYSAL, M. A. M.; ARIFUZZAMAN, S. A comparative analysis of large-scale network visualization tools. In: IEEE. **2018 IEEE International Conference on Big Data (Big Data)**. [S.l.], 2018. p. 4837–4843. Citations on pages 46 and 47.

GAO, J.; AKSOY, B. A.; DOGRUSOZ, U.; DRESDNER, G.; GROSS, B.; SUMER, S. O.; SUN, Y.; JACOBSEN, A.; SINHA, R.; LARSSON, E. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. **Science signaling**, American Association for the Advancement of Science, v. 6, n. 269, p. pl1–pl1, 2013. Citation on page 79.

GARCÍA-CAMPOS, M. A.; ESPINAL-ENRÍQUEZ, J.; HERNÁNDEZ-LEMUS, E. Pathway analysis: state of the art. **Frontiers in physiology**, Frontiers, v. 6, p. 383, 2015. Citations on pages 59, 60, 61, 62, 63, and 64.

GOODRICH, M. T.; TAMASSIA, R.; GOLDWASSER, M. H. **Data structures and algorithms in Java**. [S.l.]: John Wiley & Sons, 2014. Citations on pages 25 and 27.

HANAHAN, D.; WEINBERG, R. A. The hallmarks of cancer. **cell**, Elsevier, v. 100, n. 1, p. 57–70, 2000. Citations on pages 19 and 71.

JASSAL, B.; MATTHEWS, L.; VITERI, G.; GONG, C.; LORENTE, P.; FABREGAT, A.; SIDIROPOULOS, K.; COOK, J.; GILLESPIE, M.; HAW, R. *et al.* The reactome pathway knowledgebase. **Nucleic acids research**, Oxford University Press, v. 48, n. D1, p. D498–D503, 2020. Citations on pages 59 and 64.

JIN, M. H.; OH, D.-Y. Atm in dna repair in cancer. **Pharmacology & therapeutics**, Elsevier, v. 203, p. 107391, 2019. Citation on page 65.

KHATRI, P.; SIROTA, M.; BUTTE, A. J. Ten years of pathway analysis: current approaches and outstanding challenges. **PLoS Comput Biol**, Public Library of Science, v. 8, n. 2, p. e1002375, 2012. Citations on pages 59, 61, 62, 63, and 64.

KOH, G. C.; PORRAS, P.; ARANDA, B.; HERMJAKOB, H.; ORCHARD, S. E. Analyzing protein–protein interaction networks. **Journal of proteome research**, ACS Publications, v. 11, n. 4, p. 2014–2031, 2012. Citation on page 49.

LAGE, K. Protein–protein interactions and genetic diseases: the interactome. **Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease**, Elsevier, v. 1842, n. 10, p. 1971–1980, 2014. Citations on pages 49 and 50.

LEISERSON, M. D.; REYNA, M. A.; RAPHAEL, B. J. A weighted exact test for mutually exclusive mutations in cancer. **Bioinformatics**, Oxford University Press, v. 32, n. 17, p. i736–i745, 2016. Citations on pages 62 and 78.

MANESCU, D.; KEICH, U. A symmetric length-aware enrichment test. **Journal of Computational Biology**, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 23, n. 6, p. 508–525, 2016. Citation on page 78.

MARTINEZ-JIMENEZ, F.; MUINOS, F.; SENTIS, I.; DEU-PONS, J.; REYES-SALAZAR, I.; ARNEDO-PAC, C.; MULARONI, L.; PICH, O.; BONET, J.; KRANAS, H. *et al.* A compendium of mutational cancer driver genes. **Nature Reviews Cancer**, Nature Publishing Group, v. 20, n. 10, p. 555–572, 2020. Citation on page 76.

MAYAKONDA, A.; KOEFFLER, H. P. Maftools: Efficient analysis, visualization and summarization of maf files from large-scale cohort based cancer studies. **BioRxiv**, Cold Spring Harbor Laboratory, p. 052662, 2016. Citations on pages 71 and 72.

NUSSINOV, R.; JANG, H.; TSAI, C.-J.; CHENG, F. Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers. **PLoS computational biology**, Public Library of Science San Francisco, CA USA, v. 15, n. 3, p. e1006658, 2019. Citation on page 76.

OLDHAM, S.; FULCHER, B.; PARKES, L.; ARNATKEVICIŪTĖ, A.; SUO, C.; FORNITO, A. Consistency and differences between centrality measures across distinct classes of networks. **PloS one**, Public Library of Science San Francisco, CA USA, v. 14, n. 7, p. e0220061, 2019. Citations on pages 28, 32, 36, 39, 64, and 67.

PACZKOWSKA, M.; BARENBOIM, J.; SINTUPISUT, N.; FOX, N. S.; ZHU, H.; ABD-RABBO, D.; MEE, M. W.; BOUTROS, P. C.; REIMAND, J. Integrative pathway enrichment analysis of multivariate omics data. **Nature communications**, Nature Publishing Group, v. 11, n. 1, p. 1–16, 2020. Citations on pages 61 and 62.

PAVLOPOULOS, G. A.; PAEZ-ESPINO, D.; KYRPIDES, N. C.; ILIOPOULOS, I. Empirical comparison of visualization tools for larger-scale network analysis. **Advances in bioinformatics**, Hindawi, v. 2017, 2017. Citations on pages 46 and 47.

PIOVESAN, A.; ANTONAROS, F.; VITALE, L.; STRIPPOLI, P.; PELLERI, M. C.; CARACAUSI, M. Human protein-coding genes and gene feature statistics in 2019. **BMC research notes**, BioMed Central, v. 12, n. 1, p. 1–5, 2019. Citation on page 61.

POIREL, C. L.; OWENS, C. C.; MURALI, T. Network-based functional enrichment. In: SPRINGER. **BMC bioinformatics**. [S.l.], 2011. v. 12, n. 13, p. 1–13. Citation on page 61.

PORRAS, P. Network analysis of protein interaction data: an introduction. European Bioinformatics Institute (EMBL-EBI), 2016. Available: <https://doi.org/10.6019/TOL.Networks_t.2016.00001.1>. Citations on pages 23, 49, and 66.

PULIDO-TAMAYO, S.; WEYTJENS, B.; MAEYER, D. D.; MARCHAL, K. Ssa-me detection of cancer driver genes using mutual exclusivity by small subnetwork analysis. **Scientific reports**, Nature Publishing Group, v. 6, n. 1, p. 1–12, 2016. Citation on page 76.

RADICCHI, F.; CASTELLANO, C.; CECCONI, F.; LORETO, V.; PARISI, D. Defining and identifying communities in networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 101, n. 9, p. 2658–2663, 2004. Citation on page 42.

RAMOS, R.; CUTIGI, J.; FERREIRA, C.; EVANGELISTA, A.; SIMAO, A. Analyzing different cancer mutation data sets from breast invasive carcinoma (brca), lung adenocarcinoma (luad), and prostate adenocarcinoma (prad). In: SBC. **Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde**. [S.l.], 2020. p. 37–48. Citations on pages 62, 72, 73, 88, and 91.

RAMOS, R. H.; CUTIGI, J. F.; FERREIRA, C. d. O. L.; SIMAO, A. Topological characterization of cancer driver genes using reactome super pathways networks. In: SPRINGER. **Brazilian Symposium on Bioinformatics**. [S.l.], 2021. p. 26–37. Citations on pages 56, 64, 68, 76, 87, 88, 89, 90, and 91.

REACTOME. **Reproduction**. 2006. Available: <https://reactome.org/content/detail/R-HSA-1474165>. Citation on page 65.

____. **Chromatin organization**. 2011. Available: <https://reactome.org/content/detail/R-HSA-4839726>. Citation on page 65.

REPANA, D.; NULSEN, J.; DRESSLER, L.; BORTOLOMEAZZI, M.; VENKATA, S. K.; TOURNA, A.; YAKOVLEVA, A.; PALMIERI, T.; CICCARELLI, F. D. The network of cancer genes (ncg): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. **Genome biology**, Springer, v. 20, n. 1, p. 1–12, 2019. Citation on page 76.

REYNA, M. A.; LEISERSON, M. D.; RAPHAEL, B. J. Hierarchical hotnet: identifying hierarchies of altered subnetworks. **Bioinformatics**, Oxford University Press, v. 34, n. 17, p. i972–i980, 2018. Citation on page 51.

SAFARI-ALIGHIARLOO, N.; TAGHIZADEH, M.; REZAEI-TAVIRANI, M.; GOLIAEI, B.; PEYVANDI, A. A. Protein-protein interaction networks (ppi) and complex diseases. **Gastroenterology and Hepatology from bed to bench**, Shahid Beheshti University of Medical Sciences, v. 7, n. 1, p. 17, 2014. Citation on page 49.

SAUL, N.; TRALIE, C. **Scikit-TDA: Topological Data Analysis for Python**. 2019. Available: <https://doi.org/10.5281/zenodo.2533369>. Citation on page 91.

SCHUSTER-BÖCKLER, B.; LEHNER, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. **nature**, Nature Publishing Group, v. 488, n. 7412, p. 504–507, 2012. Citation on page 65.

SHAFI, A. A.; KNUDSEN, K. E. Cancer and the circadian clock. **Cancer research**, AACR, v. 79, n. 15, p. 3806–3814, 2019. Citation on page 65.

STRATTON, M. R.; CAMPBELL, P. J.; FUTREAL, P. A. The cancer genome. **Nature**, Nature Publishing Group, v. 458, n. 7239, p. 719–724, 2009. Citations on pages 19 and 76.

SZKLARCZYK, D.; GABLE, A. L.; NASTOU, K. C.; LYON, D.; KIRSCH, R.; PYYSALO, S.; DONCHEVA, N. T.; LEGEAY, M.; FANG, T.; BORK, P. *et al.* The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. **Nucleic acids research**, Oxford University Press, v. 49, n. D1, p. D605–D612, 2021. Citation on page 51.

SZKLARCZYK, D.; JENSEN, L. J. Protein-protein interaction databases. In: **Protein-Protein Interactions**. [S.l.]: Springer, 2015. p. 39–56. Citation on page 50.

TAMBORERO, D.; GONZALEZ-PEREZ, A.; PEREZ-LLAMAS, C.; DEU-PONS, J.; KANDOTH, C.; REIMAND, J.; LAWRENCE, M. S.; GETZ, G.; BADER, G. D.; DING, L.; LOPEZ-BIGAS, N. Comprehensive identification of mutational cancer driver genes across 12 tumor types. **Scientific Reports**, The Author(s), v. 3, p. 2650–, Oct. 2013. Citation on page 74.

TURAJLIC, S.; SOTTORIVA, A.; GRAHAM, T.; SWANTON, C. Resolving genetic heterogeneity in cancer. **Nature Reviews Genetics**, Nature Publishing Group, v. 20, n. 7, p. 404–416, 2019. Citation on page 73.

VEEN, H. J. van; SAUL, N.; EARGLE, D.; MANGHAM, S. W. Kepler mapper: A flexible python implementation of the mapper algorithm. **Journal of Open Source Software**, The Open Journal, v. 4, n. 42, p. 1315, 2019. Available: <https://doi.org/10.21105/joss.01315>. Citation on page 73.

WU, G.; FENG, X.; STEIN, L. A human functional protein interaction network and its application to cancer data analysis. **Genome biology**, Springer, v. 11, n. 5, p. 1–23, 2010. Citations on pages 51 and 54.


ICMC
USP
SÃO CARLOS