UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Exploring Complex Networks: Matrix-based and Multiscale Approaches for Pattern Recognition

Mariane Barros Neiva

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

ICMC USP SÃO CARLOS

Mariane Barros Neiva

Exploring Complex Networks: Matrix-based and Multiscale Approaches for Pattern Recognition

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Odemir Martinez Bruno

USP – São Carlos
October 2023

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

B277e

Barros Neiva, Mariane
Exploring Complex Networks: Matrix-based and Multiscale Approaches for Pattern Recognition / Mariane Barros Neiva; orientador Odemir Martinez Bruno. -- São Carlos, 2023.
180 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Redes Complexas. 2. Decomposição de grafos. . 3. Classificação. . 4. Modelagem de grafos. . 5. Matrizes de Grafos.. I. Martinez Bruno, Odemir, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Mariane Barros Neiva

Explorando Redes Complexas: Abordagens Baseadas em Matrizes e Multiescala para Reconhecimento de Padrões

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutora em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Odemir Martinez Bruno

USP – São Carlos
Outubro de 2023

Esse trabalho é dedicado a todas as mulheres que enfrentaram com coragem uma área (ainda) dominada por homens como a tecnologia.

O mundo é nosso!

ACKNOWLEDGEMENTS

Primeiro quero agradecer a minha família que sempre me apoiou em todas as minhas decisões, mesmo sem nem entender o que eu fazia! Um agradecimento especial à meu pai Josué, minha irmã Fabiane, e minha mãe^2 Valderis. Obrigada por cuidarem de mim em dias tão dificeís como foi a minha cirurgia em 2019/2020. Também quero agradecer minha tia Adriana por ser a melhor tia do mundo e um exemplo de mulher forte pra mim! Também sou imensamente grata aos meus avôs Josélio e Marlene por vibrarem e torcerem por mim e terem guardado toda matéria/lista que sai com meu nome no jornal. E como não poderia deixar, agradeço minha mãe por ter esperado tantos dias na escola/kumon/inglês para nos levar sã e salvas pra casa durante alguns anos!

Agredeço meu companheiro Etevaldo que durante esses últimos anos de doutorado, enfrentou o dia-a-dia de uma pandemia, com muitas incertezas, medos e ansiedade. Eu nem tenho palavras para agradecer tamanha paciência diária no confinamento! À minhas amigas-irmãs, meus apoios durante anos! Tenho certeza que posso contar com vocês sempre e pra sempre, obrigada pelos risos, pelas lágrimas e pela parceria de anos!

Agradecimentos especiais são direcionados ao Instituto de Ciências Matemáticas e Computação (ICMC), ao Instituto de Física de São Carlos (IFSC), aos meus parceiros de laboratório do SCG, aos parceiros do projeto RARAS (FMRP-USP) especialmente o professor Dr. Domingos Alves, também ao professor Dr. Antoine Vacavant e, principalmente, ao meu orientador Odemir por toda a motivação e paciência durante esses anos de parceria. Á todos, obrigada.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

"...O crédito pertence ao homem que encontra-se na arena, cuja face está manchada de poeira, suor e sangue;
aquele que esforça-se bravamente;
que erra, que se depara com um revés após o outro, pois não há esforço sem erros e falhas;
aquele que esforça-se para lograr suas ações, que conhece grande entusiasmo, grandes devoções, que se entrega à uma causa nobre;
que, no melhor dos casos, conhece no fim o triunfo da realização grandiosa,
e quem, que no pior dos casos, se falhar, ao menos falha ousando grandemente,
para que seu lugar jamais seja com aquelas frias e tímidas almas que não conhecem vitória ou fracasso."
(Citizenship in a Republic - Theodore Roosevelt)

RESUMO

NEIVA, M. B. **Explorando Redes Complexas: Abordagens Baseadas em Matrizes e Multiescala para Reconhecimento de Padrões**. 2023. 180 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Redes complexas são ferramentas essenciais para compreender sistemas interconectados em diversos domínios. Esta tese concentra-se na análise, classificação e modelagem de redes complexas, com o objetivo de extrair insights significativos usando metodologias inovadoras. O estudo explora a classificação de redes complexas, com um enfoque secundário na modelagem de fenômenos reais na área da saúde e análise de formas. O objetivo da pesquisa é desenvolver metodologias inovadoras que superem as técnicas existentes de classificação de redes. Duas principais abordagens são investigadas: a utilização da matriz de adjacência para análise de redes e a aplicação de técnicas multiescala para análise de grafos. A investigação das matrizes de grafos revela resultados promissores, com a ordenação baseada na centralidade dos nós e na similaridade dos nós aprimorando a representação para análise de imagens. Análises quantitativas em diversos conjuntos de dados, incluindo sistemas reais, demonstram acurácias satisfatórias na classificação com baixa parametrização. Além disso, técnicas inspiradas em visão computacional, como a decomposição k-core e a transformada de distância, aprimoram a classificação de grafos e formas. A conclusão deste doutorado em redes complexas também explora a rede ICD-ORPHA do Ministério da Saúde do Brasil. Para lidar com as limitações do sistema ICD-10 para doenças raras, é empregada uma terminologia médica especializada conhecida como ORPHA, que fornece uma nomenclatura abrangente especificamente projetada para doenças raras. Essa pesquisa expande o conhecimento sobre a modelagem de redes complexas e sua aplicação na área da saúde por meio de um sistema web interativo. Além disso, durante a pandemia de COVID-19, um modelo proposto baseado no modelo SIR avalia a dinâmica populacional e melhora a compreensão da evolução da pandemia. Essas metodologias oferecem ferramentas valiosas para insights em saúde pública e melhoria no desempenho de classificação. Em conclusão, esta pesquisa avança a análise, classificação e modelagem de redes complexas com metodologias inovadoras. Os resultados têm amplas aplicações em diversos domínios, incluindo redes sintéticas e reais, dados de saúde e análise de formas. Os resultados da pesquisa oferecem soluções práticas para a compreensão de sistemas interconectados e contribuem para o avanço da análise de redes complexas.

Palavras-chave: Redes Complexas, Classificação, Decomposição de grafos, Modelagem, Matrizes de Grafos.

ABSTRACT

NEIVA, M. B. **Exploring Complex Networks: Matrix-based and Multiscale Approaches for Pattern Recognition**. 2023. 180 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Complex networks are essential tools for understanding interconnected systems across various domains. This thesis focuses on the analysis, classification, and modeling of complex networks, aiming to extract meaningful insights using innovative methodologies. The study explores the complex network classification, with a secondary focus on modeling real phenomena in health science and shape analysis. The research objective is to develop novel methodologies surpassing existing network classification techniques. Two key components are investigated: utilizing the adjacency matrix for network analysis and applying multiscale techniques for graph analysis. The investigation of the graph matrices reveals promising results, with node centrality-based ordination and node similarity enhancing image analysis representation. Quantitative analysis on diverse datasets, including real systems, demonstrates satisfactory classification accuracies with low parametrization. Also, computer vision-inspired techniques, such as k-core decomposition and distance transform enhance graph and shape classification. The completion of this PhD in complex networks also explores the ICD-ORPHA network from the Brazilian Ministry of Health. To address the limitations of the ICD-10 system for rare diseases, a specialized medical terminology known as ORPHA is employed, providing a comprehensive nomenclature specifically designed for rare diseases. This research expands the understanding of complex network modeling and its application in the healthcare domain through an interactive web-app system. Furthermore, during the COVID-19 pandemic, a proposed SIR-based model evaluates population dynamics and enhances understanding of the evolution of the pandemic. These methodologies offer valuable tools for public health insights and classification performance improvement. In conclusion, this research advances complex network analysis, classification, and modeling with innovative methodologies. Findings have broad applications across domains, including synthetic and real networks, health data, and shape analysis. The research outcomes offer practical solutions for understanding interconnected systems and contribute to the advancement of complex network analysis.

Keywords: Complex Networks, Classification, Graph decomposition, Graph matrices, Modeling

.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

Over time, scientists have recognized the crucial role of modeling natural phenomena as a means to comprehend, classify, and forecast patterns. This recognition has led to the emergence of data science as a critical field. Representing a convergence of diverse disciplines, data science amalgamates methodologies and techniques from statistics, computer science, and specialized domain knowledge. This synthesis enables the extraction of meaningful insights in the context of the Big Data era. Not only does this field address the complexities of large-scale data, but it also leverages the potential of such information to drive decision-making processes, thus highlighting the transformative power of data science in our modern world.

Recent technological advancements have equipped researchers with the capacity to explore deeper into the underlying structures and dynamics of real data. In this scenario, the study of complex networks has emerged as a critical tool for understanding various interlinked systems. These systems encompass a wide range of examples, such as the Web, social networks, and other interconnected structures. The versatility and wide applicability of complex networks across multiple scientific and sociological disciplines — including biology, epidemiology, marketing, social relationships, computer vision, and engineering — have drawn significant research interest. Essentially, complex networks can portray any system where connections among data exist.

From constructing networks from datasets to their statistical analysis, researchers employ various methodologies when working with complex networks. In the modeling stage, networks can be constructed from a consistent set of entities based on different relationships, such as familial, social, professional, or sexual connections. For network evaluation, developers analyses information dissemination in marketing, where researchers can study the network structure to develop more effective advertising strategies by identifying nodes that efficiently disseminate information [Yang *et al.* 2010]. Generally, in modeling, it is imperative to ensure the alignment of data relationships within the structure with the research objective, to facilitate the extraction of pertinent characteristics from the graph.

Although the study of networks has gained prominence by transcending the constraints of traditional reductionist approaches and incorporating more comprehensive information to tackle problems [Miranda, Machicao and Bruno 2016], its origins trace back to the 1960s. Notable pioneers Erdös and Rényi introduced the concept of random networks, the most fundamental existing model [Erdős, Rényi *et al.* 1960]. Subsequently, other researchers explored the methodology, leading to discoveries such as Barabási and Albert's scale-free networks [Barabási and Albert 1999] and Watts and Strogatz's small-world networks [Watts and Strogatz 1998], which exhibited similarities to real-world systems and facilitated practical applications. These models exhibited remarkable resemblances to real-world systems, thereby enhancing their practical applications. Small-world networks, as seen in social networks, feature short paths between any two entities in the graph, while scale-free networks, exemplified by the Web, are characterized by the presence of hubs—nodes with a significantly higher number of connections compared to other nodes in the network.

1.1 Motivation

As previously mentioned, complex networks have numerous applications, with patterns extracted from networks that can be modeled from various data sources, such as images, time series, texts, or other systems where elements interact with each other. The highly interdisciplinary nature of this approach has led to an enthusiastic scientific community that is keen on developing methodologies for analyzing these systems for a range of applications, such as clustering and classification [Costa *et al.* 2011].

To generate a network from a dataset, it is crucial to establish a method for relating the data. For instance, in images, neighborhood relationships can form edges in networks, while in medicine, synapses can serve as relationships for analyzing dynamic brain processes. Essentially, any application with a spatial, temporal, or abstract structure, such as friendship, can benefit from network modeling. Complex network analysis is particularly suitable for nonlinear natural phenomena found in various image databases and brain systems, given the possibility of a comprehensive observation of the elements.

In the context of pattern recognition, the main goal of this work is to develop novel methodologies for analyzing complex networks and extracting essential system attributes to label networks (holistically, without categorizing nodes) of various natures. The goal is to surpass existing network classification techniques in the literature. The research group to which the candidate belongs, the Scientific Computing Group (SCG) of the São Carlos Institute of Physics (IFSC - USP), mentored by Professor Dr. Odemir Martinez Bruno, is a pioneer in utilizing complex network analysis for recognizing patterns in images [Chalumeau *et al.* 2008, Backes, Casanova and Bruno 2009, Gonçalves *et al.* 2012, Backes and Bruno 2013, Casanova 2018]. The group has frequently analyzed the use of texture with complex networks, as seen in [Chalumeau

et al. 2008], which has inspired members to apply this approach to different types of data, such as contours, shapes, dynamic textures, and even non-image-related information [Miranda, Machicao and Bruno 2016]. Among the SCG's work involving this methodology, the primary approaches examined include degree-based measures [Backes and Bruno 2013, Gonçalves, Machado and Bruno 2015] and deterministic and random walks [Gonçalves *et al.* 2012, Gonçalves *et al.* 2012]. Many of these strategies have produced superior results compared to other traditional texture and shape analysis methodologies.

Furthermore, our background in image classification and computer science is evident from our previous projects. With dissertation and other studies in texture analysis and classification [Neiva, Vacavant and Bruno 2018, Neiva, Guidotti and Bruno 2017, Neiva 2016], have significantly bolstered our expertise in this field. This proficiency has also been instrumental in our pioneering approach to combine images and graphs, deviating from the conventional research methodologies by transforming images into complex networks.

By leveraging the benefits of graph modeling, which has been proven effective in portraying a diverse range of phenomena, along with the knowledge of SCG, we have cultivated a keen interest in exploring innovative methodologies. As previously mention, the main goal of this exploration is to enhance the overall classification efficiency of complex networks. This unique blend of our previous experiences and academic skill has instigated our search to push the boundaries in this domain.

Moreover, health data was an important subject over the development of this thesis. First, we faced a huge pandemic process which moved several researchers to focus on the pandemic data analysis in order to mitigate the bad effects of virus transmission. In addition, the project RARAS network available at https:/raras.org.br and funded by the Ministry of Health of Brazil and cnPQ is also a health-domain project that seeks to collect, analyse and generate insights over the rare diseases in the country. The author participate since the beginning of the project and also was able to develop a work combining rare disease codification and complex network modeling.

1.2 Objectives

The primary objective of this project is to investigate and devise innovative methodologies mainly within the domain of computer vision and graph theory. The central goal is to engineer efficient techniques for network feature extraction, with effectiveness assessed via classification accuracy. It is important to underscore that our suggested approaches should proficiently identify patterns across a wide range of applications, irrespective of the original data format. This is predicated on the understanding that any system characterized by interconnected elements can be modeled as a complex network. We are introducing two main proposals in this project: the employment of the adjacency matrix as a pivotal element of graph analysis, and the application of multiscale techniques for graph analysis.

The first goal of this thesis is to evaluate the power of the one-to-one representation of the graphs: the adjacency matrix. Typically, networks are visualized as an ensemble of nodes, symbolized by circles, and links, denoted by lines connecting them. Contrary to the conventional approach in global graph classification from images, our study explores the two-dimensional representation in an effort to distinguish patterns across different classes in the datasets.

Our proposed methodology comprises two versions: the binary and the colored. For both versions, the matrix must be rearranged for proper analysis. Subsequently, we incorporate our approach with numerous established image classification methodologies such as convolutional neural networks. The results demonstrate that, particularly for the binary version of the approach, the classification accuracy surpasses existing literature results in eight out of twelve datasets, including both real and synthetic models. Drawing inspiration from computer vision techniques, we aim to determine if the patterns within the models can be accentuated by exploring the various granularities of the graph. To create these 'scales' in the graph, we implement the k-core methodology to decompose the original dataset into numerous subgraphs, each exhibiting distinct characteristics. Then, for each sub graph, we can build the features of the network from its two-dimensional representation, the matrix, from classical measures and/or using other methodologies such as [Zielinski *et al.* 2022]. This particular later approach has shown remarkable results in comparison to other global network classification techniques.

Furthermore, the multiscale approach is also evident in our result entitled *Distance transform technique to classify shapes* [Ribas, Neiva and Bruno 2019]. In this study, we delve into the modeling of complex networks and employ the distance transform technique to effectively classify shapes. This research represents a convergence of the SCG group's expertise in both complex networks and computer vision. A draft version of the paper can be found in Annex C, providing additional details and insights into our methodology and findings.

Moreover, during the work course of this thesis, we undertook various projects both with and without the utilization of complex networks. One significant example is our response to the challenging circumstances presented by the **COVID-19** pandemic. We delved deeply into the Portal COVID-19 project, a volunteer organization focused on daily monitoring of the COVID-19 pandemic in Brazil. Leveraging our expertise in complex networks, the SCG group conducted a comprehensive study on the evolution of the pandemic, incorporating multiple constraints and developing a SIR-based model adding several characteristics of the disease. A version of this work is presented in Annex B, while the complete study can be found in [Scabini *et al.* 2021].

In addition, the completion of this PhD in complex networks also provided an opportunity to delve into the modeling of the ICD-ORPHA network from the Brazilian Ministry of Health, as mentioned before. The ICD-10, known as the International Classification of Diseases, 10th Revision, is a globally recognized system developed by the World Health Organization for classifying and coding diseases, medical conditions, and related health information. It is extensively utilized in the Brazilian health system for disease codification [Organization 2022].

However, the ICD-10's general nature makes it inadequate for rare diseases (RD), which affect a small number of individuals within a population. For the classification of RD, a specialized medical terminology known as ORPHA is more suitable [Weinreich *et al.* 2008]. ORPHA provides a comprehensive nomenclature specifically designed for rare diseases.

In the work, detailed in Annex A, we developed a web-app iterative system to model the complex network between both terminologies within the context of the Brazilian Ministry of Health. Additionally, we explored the characteristics and statistics of the proposed complex network for each component of the model. The research has been accepted for presentation at the HCist - International Conference on Health and Social Care Information Systems and Technologies in 2023. The developed web-app system serves as a valuable tool for bridging the gap between the ICD-10 and ORPHA in the Brazilian context, enabling a more comprehensive understanding and representation of rare diseases within the Brazilian healthcare system.

In summary, the goals of this thesis are:

- Conduct a thorough investigation and analysis of the pattern recognition methods based on complex networks that are relevant to the scope of this work.
- Introduce new methodologies for characterizing complex networks, utilizing both two-dimensional and multiscale representations from the graphs.
- Propose new techniques for characterizing shapes, utilizing multiscale through distance transform modeling
- Carefully assess and compare the effectiveness of the proposed methods with the existing approaches found in the relevant literature evaluating the accuracy of the models.
- Apply the developed methods in synthetic real-world datasets.
- Contribute to the knowledge and efforts aimed at mitigating the transmission of COVID-19 by utilizing complex network modeling and studying social dynamics.
- Explore and model the cross-reference graph between the ICD-10 codification and ORPHA terminology for Brazilian Rare Diseases.

1.3 Contributions

It is important to highlight that this work does not narrowly focus on a specific application, but rather endeavors to develop novel methodologies that can enhance the classification accuracy of complex networks. The main contributions of this work are:

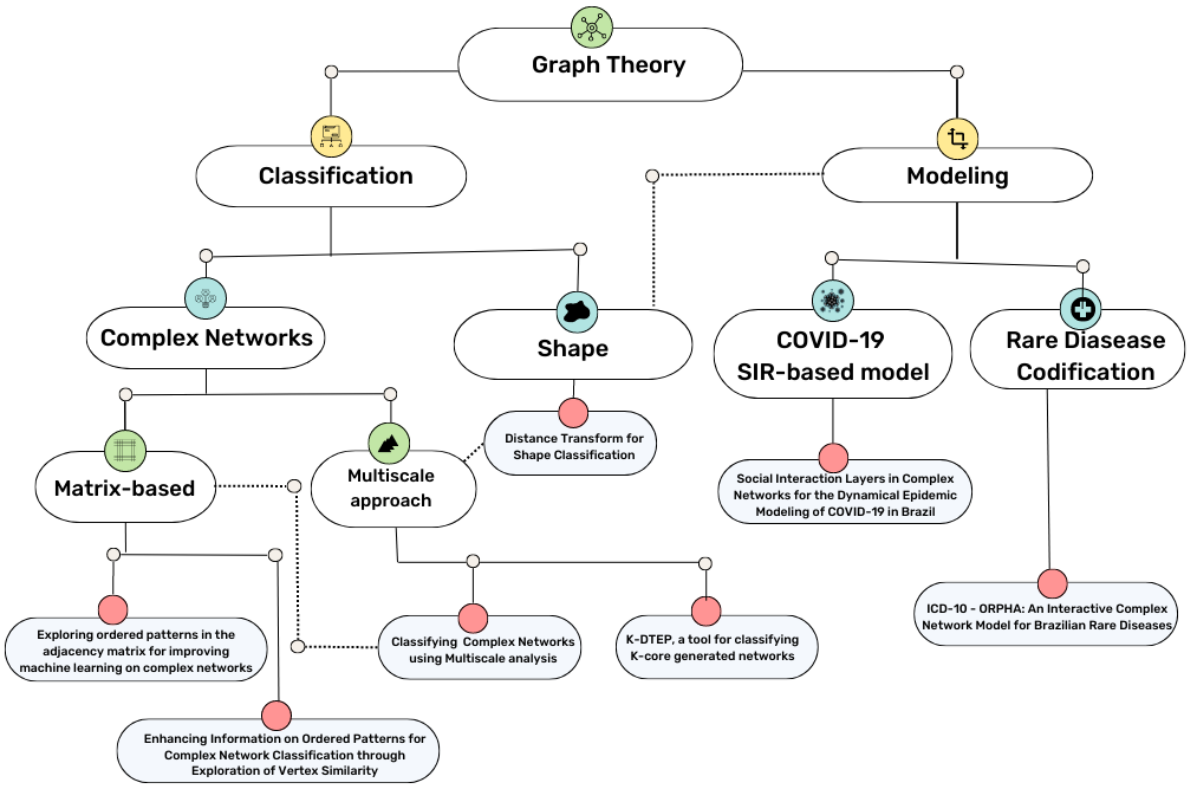
- Methodologies that surpass the existing methods in the literature for characterizing complex networks and shapes, yielding, in general, superior results.

- Present proposals with a minimal level of parameterization when compared to the compared methodology, reducing the complexity of the approach for complex networks classification.
- Ability the potential for combining the proposed approaches with various image exploration techniques, including convolutional neural networks, enhancing the overall results.
- Visual and quantitative analysis of complex networks is made possible through a transformative change in perspective, as proposed in the utilization and reorganization of adjacency matrices.
- The outcomes from these methodologies, implemented on both synthetic and real networks, underscore the significant of this work. The real networks that can be assessed encompass a range of applications, including classification of social networks (Google+ and Twitter) and metabolic among others, which will be presented in the course of the project.
- The application of CN knowledge was demonstrated in two distinct contexts within health data: rare diseases and the COVID-19 pandemic.
  - This was utilized to model and elucidate the complexity of codifying Brazilian rare diseases.
  - Additionally, it showcased the potential of using social dynamics to mitigate virus transmission, underscoring the value of this approach in managing public health crises.

1.4 Thesis Structure

Figure 1 shows the overall structure of the thesis project. This document is organized as follows: Chapter 2 introduces the theoretical concepts associated with complex networks, such as their definition, main models, and some metrics found in the literature. Additionally, the chapter provides an overview of some techniques used in the. Chapter 3 conducts a literature review on the topic of complex network classification and the connections between methodologies to be used and the data model. Chapters 5, 6, 7 and 8 are presented in paper-like format and provides the core of the thesis development in complex network classification. They present the usage of adjacency matrix, Jaccard Index-based matrix, k-core decomposition and the combination of D-TEP and decomposition, respectively. Finally, Chapter 9 presents the final conclusion of the doctoral research project. Then, we chose to add published papers, specially those regarding modeling, in the annex to keep the CN classification techniques in the main course of the text. Thus, Annex C presents the use of complex networks modeling for shape classification and Annex A and B presents the use of CN for health data, where the first uses CN for ICD-ORPHA modeling while the later proposes a susceptible-infected-recovered inspired model over social dynamic for COVID-19 pandemic.

Figure 1 – Thesis structure. From graph theory area, the primarly goal of this thesis is to classify complex networks which encorporates proposals with matrix-based and multiscale analysis. Moreover, shape classification with multiscale analysis by the use of distance transform is also proposed. Furthermore, modeling was also explored during the thesis. Two approaches are presented: ICD-ORPHA graph construction and the SIR-based model for COVID-19 transmission.



Source: Developed by the author

CHAPTER

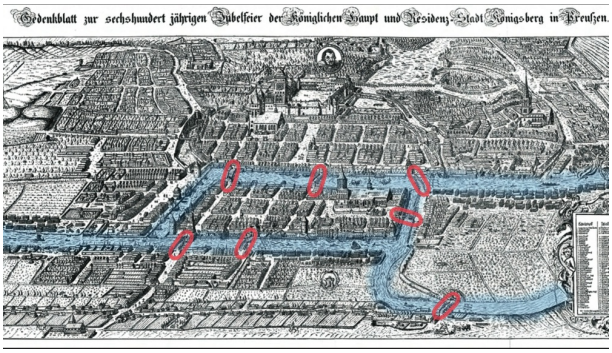
2

THEORETICAL FUNDAMENTALS

For a comprehensive understanding of this project, several concepts related to the proposal and the subject of study are delineated in this section. Here, we present definitions associated with modeling, complex networks, classical metrics, adjacency matrix, among others.

2.1 The History of Graph Theory

Figure 2 – Map of Königsberg in Russia presenting the seven brigdes.



Source: Developed by the author

During the 18th century in Königsberg, Russia, there were seven bridges that connected two islands and the mainland, spanning the Pregel River (see Figure 2). The mathematician Carl Gottlieb Ehler became intrigued by the question of whether it was possible to cross each bridge exactly once.

Initially dismissed by Euler, the problem sparked his interest as he realized that the necessary geometric principles had not yet been established. As a result, Euler introduced a new branch of geometry known as the *geometry of position*. He simplified the Königsberg bridges into a problem involving four nodes (the land masses) and edges (the bridges connecting these lands). This shift in perspective allowed Euler to approach the situation as a mathematical problem,

focusing on the degree of each node, which represented the number of bridges connected to each land mass.

Euler's conclusion was that for each bridge to be crossed exactly once, the degree of each land mass had to be even, which was not the case in the Königsberg bridge arrangement. The only exception would be if there were a path where the starting and ending nodes had odd degrees. If the challenge requires that each node can only be crossed once and only once, it means that for each land, the number of degree should be even, which was not the case of Konigsberg problem. This introduced the concept of an Eulerian cycle: a path in a graph that starts and finishes at the same vertex and traverses each edge exactly once.

During their investigation of the problem, Ehler and Euler corresponded approximately 20 times [Sachs, Stiebitz and Wilson 1988], culminating in the publication of their findings [Euler 1741]. Today, the city of Königsberg and its seven bridges no longer exist in their original form. Some of the bridges were destroyed during World War II by Soviet airstrikes, and the area is now part of the city of Kaliningrad in Russia.

Despite Euler's contribution, the term *graph* gained popularity in 1891 through the work of mathematician Julius Petersen, as documented in his article [Petersen 1891]. Since then, numerous researchers have proposed various methods, theories, and applications using graph theory. For example, the usage of CN representation, such as adjacency matrices, was introduced by KONIG in 1931. Additionally, DIJKSTRA made significant contributions to the field, notably through the development of Dijkstra's algorithm for finding the shortest path in a graph. This algorithm has found applications in network routing, GPS navigation, and other domains.

With the advancement of technology and the rise of Big Data and Data Science, graph theory has further evolved. CNs have emerged as powerful models for understanding and analyzing real-world phenomena. CNs have been successfully applied in diverse fields such as biology, sociology, finance, marketing, and medicine, to name a few. While some researchers focus on generating networks from datasets, your work specifically emphasizes the analysis of graphs.

The analysis of graphs provides valuable insights into the structural properties, dynamics, and behavior of complex systems. By leveraging the rich mathematical and analytical tools of graph theory, this research aims to contribute to the classification of complex networks and its applications in various domains.

2.2 Complex Networks

In this section, we provide an overview of the theory underlying complex networks, including their fundamental properties and relevant metrics. Complex networks have been extensively studied, revealing significant features that establish connections between these

networks and real-world systems. Notable discoveries include the emergence of scale-free properties [Barabási and Albert 1999], the presence of small-world phenomena [Watts and Strogatz 1998], and the development of community detection techniques [Girvan and Newman 2002].

By combining modeling techniques and statistical methods, complex networks aim to gain insights into the connections, patterns, behavior, and evolution of interconnected systems. For instance, the study of small-world networks demonstrates that any two individuals in the world can be connected through at most six intermediaries, highlighting the prevalence of short distances between nodes [Watts 2003]. This intriguing finding serves as one motivation for exploring the potential of complex networks. In this section, we delve into the formation processes and fundamental characteristics of these networks.

2.2.1 Definition

A complex network (or network) is represented by a graph G, consisting of a set of nodes V connected by edges E, where $E \subseteq (u,v)|u,v \in V$. These edges signify relationships between pairs of vertices based on the specific modeling approach and can be either directed (digraphs) or undirected. Furthermore, edges can contain additional information, such as costs, which often play a crucial role in network analysis. In the case of weighted networks, edge weights, denoted as $w_{u,v}$, represent the cost associated with the edge (u,v). Such weights can represent distances, differences in intensity or capacity, similarities, and other relevant factors.

There exists a debate within the scientific community regarding the distinction between a graph and a complex network. While every complex network can be regarded as a graph, some argue that the reverse is not necessarily true. Advocates of this viewpoint propose that a graph should only be considered a complex network when it exhibits specific topological properties as outlined in Section 2.2.2 [Costa *et al.* 2007]. However, in this work, both terminologies will be treated as synonymous, particularly when used in conjunction with neural networks, to prevent confusion between the terms.

In addition, a graph can contain multiple components, each of which represents a subset of vertices and edges. Within a component, every vertex can be reached from any other vertex by traversing the edges of the subgraph. Essentially, a component is a connected subgraph in which all vertices are mutually reachable. Analyzing the components within a graph can provide valuable insights, and it is often of interest to evaluate each component individually or focus on the largest component in particular.

2.2.2 Main Metrics and Related Concepts

After modeling the system as a graph, some terms and metrics are used and considered classic for representing the network. Some of these metrics and concepts are presented here:

- **Complete Graph**: a network where all nodes are connected to each other, that is, for any pair of vertices in G there exists an edge e connecting them. In formal mode, a complete graph is represented by K_n, where n is the number of nodes in the graph. Also, due to its property, the total number of edges is given by E = n(n-1)/2 and every node has n-1 connections.

- **Regular Graph**: a graph where the number of connections of each node/vertex is the same. Both the complete graph and the regular graph do not present relevant topological structures and are rare to be found in real systems. However, some interesting aspects are noted in some regular graphs:

  - If a all nodes in a regular graph have degree one, the graph is composed of n/2 components.
  - A cycle or a set of disjoint cycles are exhibit in regular graphs with degree equals to 2.

- **Path**: a path is a sequence of nodes connected by edges representing a route between two elements in the graph. A path C is represented by a finite ordered sequence of edges of the form C = $\{(v_1, v_2), (v_2, v_3), ..., (v_{n-1}, v_n)\}$. The size of the path is all defined by the number of edges, i.e. n-1 vertices. The concept appears in several metrics of graph characterization and can be used in application such as network routing and GPS navigation [Dijkstra 2022].

- **Diameter**: The metric represents the maximum distance between two nodes in the graph. It is important to highlight that to compute the diameter, we evaluate all shortest paths between all pair of nodes, it means, the paths between the pair that requires the minimum number of edges. The diameter is able to demonstrate the overall size of the graph and provides insights regarding the efficiency of the information transmitted withing the network. Furthermore, if the size of the graph is high but the diameter is low, it also give us a clue about the high connectivity of the model. Application such as routing and social analysis can use this metric for network characterization. The diameter D is computed by:

$$D = max\{sp(u,v)\}, \forall (u,v) \in V, \tag{2.1}$$

  where u,v are nodes in the graph and sp output the length of the shortest path between pair (u,v).

- **Average shortest path**: To compute the average shortest path, one can compute the shortest path between every pair of vertices and then find the average of those path lengths. Likely diameter, the average shortest is an important metric to evaluate the connectivity of the network as well as the ability of information routing. To compute the average shortest

path in an undirected node:

$$s\bar{p}_G = \frac{\sum sp(u,v)}{(n(n-1)))} \tag{2.2}$$

- **Degree**: One of the main measurements of the graph, the degree represents the number of edges connected to a certain node u. When a graph is directed, there is a differentiation between the number of incoming and outgoing edges of the node (in-degree and out-degree). Many extraction methods use this measure to compose their feature vector which, despite being simple, can bring much representativeness to the structure of the graph and a hint of its most important nodes. It will be presented here in this study by the k_i, i.e., the degree of node i.

- **Average degree**: The metric refers to the average number of edges connected to each vertex in the network and is computed by:

$$\bar{k} = \frac{\sum_{\forall i \in V} k_i}{n} \tag{2.3}$$

- **Degree distribution**: Like the histogram in images, the degree distribution is a probability function that shows the frequency of each degree. The degree distribution of undirected graphs can be represented by a cumulative function P_k, where k is a degree present in the network:

$$P_k = \sum_{k'=k}^{\infty} p_{k'} \tag{2.4}$$

  In random networks, the degree distribution follows the Poisson law and for many real networks, the distribution is of the power law type $p_{k'} \sim k^{-\alpha}$.

- **Clustering coefficient**: The global metric offers the idea that if a node u is connected to a vertex v and v is connected to j, then the probability of u being connected to j is higher. Simplistically, the clustering coefficient relates to the number of triangles formed in the network. The calculation is done according to Equation 2.5, where #Δ represents the number of triangles formed in the network and #v the number of vertices that belong to triangles.

$$CA = 3\frac{\#\Delta}{\#v} \tag{2.5}$$

- **Degree assortativity** (Pearson correlation): It quantifies the level of degree correlation among nodes in the network. It ranges between -1 and 1, indicating that nodes with similar degrees are more likely to be connected to each other, while a value close to -1 indicates that nodes with dissimilar degrees are more likely to be connected. The pearson correlation is computed by:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2.6}$$

where ρ_{XY} represents the Pearson correlation coefficient between the degree sequences X and Y of the graph, X_i and Y_i are the degrees of the ith vertex in the respective sequences. $\bar{X}$ and $\bar{Y}$ denote the means of the degree sequences X and Y.

- **Closeness centrality** (cl): Represented by a vector cl of size n, where for each position i, it represents the inverse of the sum of all short distances between i and all other nodes in the graph, the closeness. The closeness if computed by:

$$cl_u = \frac{\sum_{\forall v \in V} sp(u,v)}{n-1} \tag{2.7}$$

  For a global measurement, one can compute the average closeness centrality of all nodes in the graph.

- **Eccentricity** (ecc): In contrast with closeness, the eccentricity of a node is the highest geodesic distance between a node i and any other node in the network.

$$ecc_u = max\{sp(u,v)\}, \forall v \in V \tag{2.8}$$

  Vertices with small eccentricity are considered well centered or well-connected in the network. Likely in other centrality measures, to compute the global eccentricity, simply average $ecc_u \; \forall$ u $\in$ V

- **Betweenness** (bet): Another measurement based on shortest paths, the betweenness counts how many times a particular node i is on all shortest paths of a graph. It quantifies the influence of the vertex in the graph and how it serves as a bridge to connect other pair of nodes. The metric is computed as:

$$bet(u) = \sum_{i \neq v \neq j} \frac{\sigma_{i,j}(v)}{\sigma_{i,j}} \tag{2.9}$$

  where $\sigma_{i,j}(v)$ is the total number of shortest paths between vertices i and j that cross v and $\sigma_{i,j}$ is the amount of shortest paths between i and j. A high betweenness indicates that the node plays a critical element for information transmission.

- **Average Hierarchical Degree (level 2)** ($\bar{Hk}_2$): For a given node v, the $Hk_2(v)$ is the the sum of the degrees of its neighboring nodes. In other words, it represents the total number of edges connecting the neighbors of node i. Consequently, the average hierarchical degree of level 2 can be determined by dividing the sum of $Hk_2(v)$ values for all nodes in the graph by the total number of nodes.

- **Average Hierarchical Degree (level 3)**: For a given node v, the $Hk_3(v)$ is the the sum of the degrees of the neighbors of its neighboring nodes. Consequently, the average

hierarchical degree of level 3 can be determined by dividing the sum of $Hk_3(v)$ values for all nodes in the graph by the total number of nodes.

- **Average Path Length (l)**: The metric represents the average distance or number of edges required to travel between any two vertices in a graph. It provides insights into the overall efficiency and connectivity of the network. It is computed by:

$$l = \frac{1}{n(n-1)} \sum_{i,j} \sigma i, j \tag{2.10}$$

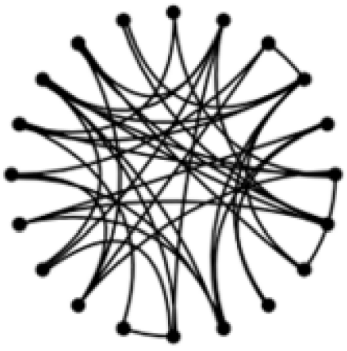
2.2.3 Complex Networks Models

Over the years of studying complex networks, three main types of categories have been classified and are currently widely cited. They are:

2.2.4 Random Networks

A noticeable property in random networks is that the degree distribution follows the Poisson distribution. Proposed by [Erdős, Rényi *et al.* 1960], this is the simplest network model. The network has two parameters to be created, n and p, where n is the total number of vertices and p is the probability of an edge being included in the network. The expected degree of a random network is $p(n-1)$. According to ERDŐS; RÉNYI *et al.*], the network is said to be random due to the fact that vertices are randomly connected, without a priority of connection. Furthermore, since each edge can either be present or absent in the network, a total of $\binom{n}{2}$ edges are possible in the network. Figure 3 shows an example of a random network.

Figure 3 – Example of a random network.

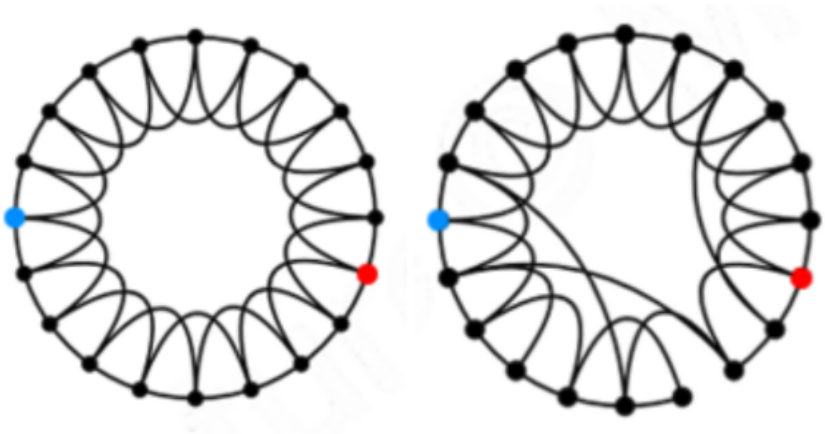


Source: Adapted from [Watts and Strogatz 1998]

2.2.5 Small-world Networks

In small-world networks [Watts and Strogatz 1998], the clustering coefficient is high due to the low number of edges in relation to the high number of triangles formed in the network. In these networks, nodes are highly connected, with short paths between them.

Figure 4 – A small-world network can be easily constructed from a regular network by rewiring some edges.



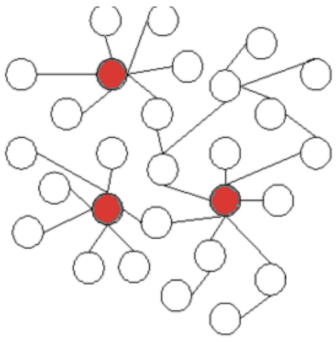
(a) Regular network (b) Small world network

Source: Adapted from [Watts and Strogatz 1998]

An example of a small-world network is social networks, where there is a low number of people connecting two users. To create such a network, one can generate a regular graph (all nodes with the same degree) and choose an edge from the initial k connections of each node to be rewired with probability p. Figure 4 shows an example of a regular network and a small-world network. In the figure, the small-world network was generated from the network where all vertices have the same degree.

2.2.6 Scalefree Networks

Figure 5 – Example of a scalefree network.



Source: Developed by the author

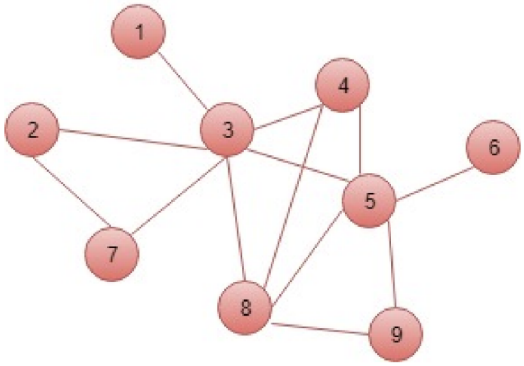
In this model, BARABÁSI; ALBERT observed that some networks have an irregular tendency of connection where some vertices present many connections (hubs) in comparison with others. In this category the degree distribution follow a power law $P(k) \sim k_{-y}$ where y is the exponent scale. To generate a network of this type, a higher probability of connection is assigned to the hubs. The probability is then calculated proportionally according to the value of the degree of each node according to Equation 2.11, where k_j is the degree of vertex j and $\sum_u k_u$ is the sum of the degrees of the other vertices in the network. Figure 5 shows an example of a scale-free network.

$$P(i,j) = \frac{k_j}{\sum_u k_u} \tag{2.11}$$

The following section is dedicated to overcome special aspects of this project regarding graph theory and other related concepts that are important for the development of this thesis.

2.3 Matrices of the Graph

Figure 6 – Example of a graph and its adjacency matrix



nó	1	2	3	4	5	6	7	8	9
1	0	0	1	0	0	0	0	0	0
2	0	0	1	0	0	0	1	0	0
3	1	1	0	1	1	0	1	1	0
4	0	0	1	0	1	0	0	0	0
5	0	0	1	1	0	1	0	1	1
6	0	0	0	0	1	0	0	0	0
7	0	1	1	0	0	0	0	0	0
8	0	0	1	1	1	0	0	0	1
9	0	0	0	0	1	0	0	1	0

(a) Graph (b) Adjacency Matrix

Source: Developed by the author

Usually, graphs are represented by circles representing the nodes and lines between representing edges as seen in Figure 6a. However, other alternatives can be used. One of the main one-to-correspondence to the classical circles and lines representation of the network is the adjacency matrix (Figure 6b). The matrix is a square matrix that express the connections or relationships between vertices in a graph. If there is an edge between vertices i and j, the corresponding entry in the matrix will be a non-zero value, indicating the presence of the edge. If there is no edge between vertices i and j, the entry will be zero. The adjacency matrix is formally

defined by:

$$A(i,j) = \begin{cases} 1, & \text{if i and j are connected} \\ 0, & \text{otherwise.} \end{cases} \tag{2.12}$$

Furthermore, other characteristic of the network shown is the the matrix is the size of the graph G. In the two-dimensional approach, the number of rows or columns is equal two n, the number of nodes in G. In addition, if edges have no direction, i.e., if node i is linked to j, j is also linked to i, the matrix is symmetric. Also, the matrix is binary is graph has no weight in the edges. However, for weighted graphs, if i and j are connected in matrix A with weight f, $A(i,j) = f$.

From this representation, several metrics can be easily computed such as degrees, number of nodes and path size. Moreover, several matrices are derived from the original adjacency matrix such as:

- **Degree Matrix (D)**: The degree matrix sets all diagonal values to the degree of referenced node in ith line. For instance, value D(2,2) = k_2, where k_2 is the degree of node 2 in the graph. Formally, the matrix can be computed as:

$$D(i,j) = \begin{cases} k_i, & \text{if i=i} \\ 0, & \text{otherwise.} \end{cases} \tag{2.13}$$

- **Laplacian Matrix (L)**: The third matrix uses the difference between the degree matrix (D) and the adjacency matrix (A), i.e., L = D - A. The matrix presents several properties such as: symmetry, non-negative eigenvalues, and the multiplicity of the zero eigenvalue equals the number of connected components in the graph. In addition, each element represents the degree of a vertex and its connections to other vertices. The matrix can also be used to understand important topological properties of the system.

  In general, the graph's matrices are used representation in graph theory and allows for efficient manipulation and analysis of graph structures. It can provides a concise and compact way to store information about the connections between vertices in a graph.

2.4 K-core Decomposition

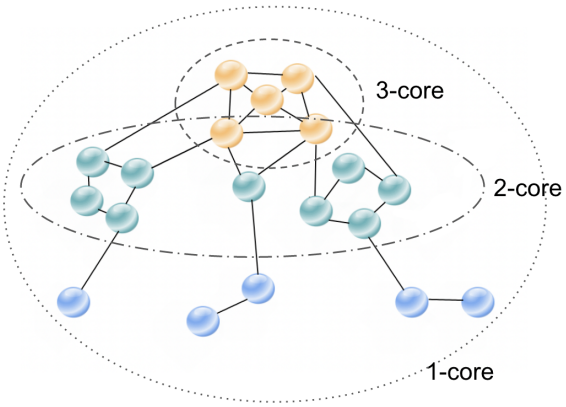
Away from the graph representation, other subject relevant to our study is the k-core decomposition. In this section, one will understand the main concept and definitions of this topic. A k-core a is the largest subgraph in which all vertices have at least degree k. It is used in complex networks to simply and analyze the structure of the model and provide a layered or 'multiscale' understanding of the graph.

The algorithm for identifying a specified k-core is delineated as follows [Batagelj and Zaveršnik 2002]:

1. Initiate the algorithm by removing all nodes with a degree less than k.
2. Ascertain if the resultant graph has all remaining nodes with a degree of at least k.
   - If nodes with degrees less than k persist, implying the existence of further edges, revert to step 1 and reapply the procedure.
   - Conversely, if all nodes exhibit a degree of at least k, the resultant subgraph can be identified as the k-core.

Note that, with the increase of k, the subgraph is more connected and nodes with higher centrality such as degree metrics remains due to the nature of the algorithm [Kitsak *et al.* 2010]. The analysis, done by coloring the nodes with different colors according to certain characteristics, is widely used in community detection and clustering due to the strong connection that a k-core subgraph presents and it is used as an important tool for visualization. The k-core decomposition can also be analyzed as a multiscale generation of a larger graph and is very useful for visualization [Alvarez-Hamelin *et al.* 2005]. An example of $k$core decomposition is shown in Figure 7.

Figure 7 – Example of graph with k={1,2,3}-core.



Source: Developed by the author

2.5 Image Feature Extraction Methods

As mentioned before, this work is inspired by computer vision techniques to classify networks. Thus, some feature extraction methods are further used in the development of this project to evaluate the proposed algorithms. In this section, three techniques are detailed:

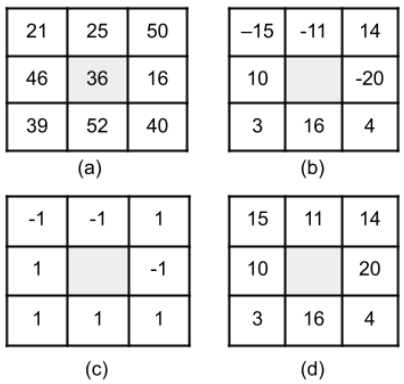
2.5.1 Complete Local Binary Pattern

The first handcraft method is the Complete Local Binary Pattern, CLBP [Guo, Zhang and Zhang 2010], a method extended from the Local Binary Pattern, LBP [Ojala, Pietikäinen and Harwood 1996], that evaluates the distribution of pixels locally. The LBP is a simple and very efficient texture operator which labels the pixels of an image by thresholding the neighborhood of each pixel and considers the result as a binary number. However, the Complete Local Binary was proposed to overcome the limitation of the traditional LBP in representing image local structures.

While traditional LBP only encodes the binary differences between the central pixel and its neighbors, CLBP considers three complementary components of local image structures: central pixel (CLBP_C), local differences (CLBP_S), and magnitude of local differences (CLBP_-M).

First, the CLBP_C component represents the gray level of the central pixel. The feature is construct by the binary code of the intensity level in the central pixel. Then, the CLBP_S component corresponds to the traditional LBP, which represents the sign of the local difference. Finally, the CLBP_M encodes the magnitude of the local difference. Both CLBP_S and CLBP_M are calculated by comparing each pixel in the neighborhood with the central pixel. According to [Guo, Zhang and Zhang 2010], the component CLBP_S is able to preserve local structures compared to the magnitude map, CLBP_M. Figure 8 represents the three matrices from a 8x8 window in the image.

Figure 8 – Example of CLBP algorithm. (a) represents a 3x3 window. (b) local difference (c) signal difference (d) magnitude difference



Source: Developed by the author

However, the concatenation of these three components allows CLBP to provide a more robust and comprehensive representation of local image structures, which improves the performance of tasks such as texture classification and face recognition.

2.5.2 Hu Moments

Another crucial method for extracting texture features is the use of Hu Moments. This measure yields a set of seven numbers, which encapsulate the image's moments and enable the depiction of shapes present in the image for computer vision activities [Hu 1962].

The idea of moments in the context of image processing refers to the assessment of intensity (pixel brightness) dispersion around an image's center or a specific region. Image moments aid in pinpointing particular features of an image, such as centroid, area, orientation, and so forth.

Hu Moments expand on this concept and are especially recognized for their invariance attributes. They remain unchanged with respect to translation (position shift), scale (size alteration), and rotation, making them highly valuable in image recognition, as they can identify a shape regardless of its location, dimensions, or orientation in the image.

These seven moments are derived from the image's normalized central moments. These seven moments are a mixture of the image's second and third order moments. Here's the mathematical representation for Hu Moments where the seven moments (ϕ) are computed by normalized central moments (η) in the following manner:

$$
\begin{aligned}
\phi_1 &= \eta_{20} + \eta_{02} \\
\phi_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
\phi_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
\phi_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
\phi_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
\phi_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
\phi_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] - \\
&\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned}
\tag{2.14}
$$

2.5.3 Neural Networks

Artificial neural networks, inspired by the neural networks in living organisms, are computational models aimed at learning from distinct types of data. In the brain, neurons are echoed by processing units in artificial constructs. For cognitive processes to occur, a process analogous to synapses takes place in artificial networks, wherein unit A transmits information to unit B. Depending on various factors, this transmission can increase or decrease the activation of unit B [McCulloch and Pitts 1943].

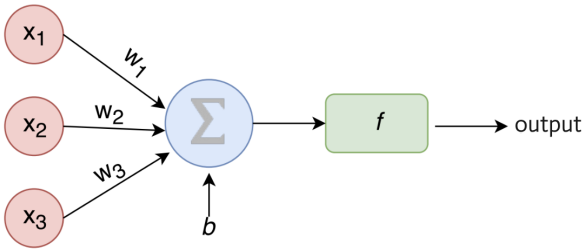
The study of neural networks dates back to the mid-twentieth century where the model

has been the subject of extensive debate. Pioneering researchers in this field include [McCulloch and Pitts 1943, Graziano 2006, Rosenblatt 1958]. Rosenblatt's perceptron model, introduced in 1958 [Rosenblatt 1958], was among the earliest proposed neural networks. This model represents a simple linear classification network, with the binary classifier expressed by the equation:

$$f(x) = \begin{cases} 1 & \text{if } wx + b \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.15}$$

In Equation 2.15, w refers to the weight of the connection, x denotes the input vector to be classified, and b represents the bias term which adjusts the threshold of the activation function [Rosenblatt 1961]. The model is also illustrated in Figure 9 However, as a linear classifier, the perceptron was incapable of solving problems like the XOR. This limitation disillusioned the scientific community, who had initially been optimistic about neural networks as the future of science. Consequently, researchers lost interest in this field until the advent of the multilayer perceptron (MLP) in 1961, which introduced a nonlinear activation function capable of classifying non-linearly separable data [Rosenblatt 1961].

Figure 9 – Example of perceptron with 3 channels



Source: Developed by the author

With the return of credibility to neural networks, researchers' interest in further studying this model, so inspired by human nature, has grown once again. The inclusion of more than one layer of neurons has made neural networks a significant tool for data classification used by large corporations and banks. Learning takes place by updating the network weights, which, in the standard mode, occurs after each training set example is presented.

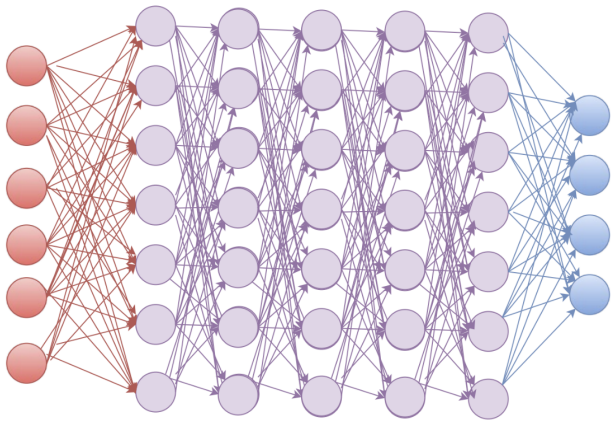
In the batch model, all training examples are presented to the network, and the average error is calculated. Based on this error, the necessary adjustments are made to the network weights, and the calibration process is carried out again until a minimum average error or maximum number of algorithm iterations is achieved. Furthermore, for both cases, the speed of weight modification is determined by the developer.

This study used deep neural network to retrieve features and they have proven to be powerful in image recognition applications and therefore have significant potential for studying graph feature extraction according to our approach.

2.5.4 Deep Neural Networks

Deep neural networks have gained substantial fame due to their extraordinary pattern recognition abilities and their use by corporations such as Google and Microsoft, among other multinational companies. However, there remain questions regarding the definition of this type of network and what constitutes a deep neural network. According to [Deng and Yu 2014], deep learning involves machine learning techniques that utilize many layers for supervised and unsupervised learning, as well as pattern analysis. The technique's renown is certainly associated with its use in Computer Vision [Krizhevsky, Sutskever and Hinton 2017, Farabet *et al.* 2012, Tompson *et al.* 2014, Szegedy *et al.* 2015]. The hugh number of layers characteristic of neural networks allow image features to be learned hierarchically at each stage. While the initial layers extract information such as points and edges, subsequent layers combine this information to extract patterns such as borders and shapes.

Figure 10 – Example of a deep neural network with five fully connected hidden layers shown in purple. The red and blue layers represent the input and output layers, respectively.



Source: Developed by the author

There is no universal agreement on the minimum number of layers needed to construct a neural network for it to be considered "deep". Some believe that a count greater than two layers suffices to form a deep network, while others require a minimum of ten layers, reaching up to 150. An example of a deep neural network is presented in Figure 10. Regardless, some models have gained prominence in the literature due to their recognition power and good performance in competitions related to the field of machine learning, such as GoogleNet [Szegedy *et al.* 2015], VGG-16 [Simonyan and Zisserman 2014], ResNet [He *et al.* 2016], and ImageNet [Krizhevsky, Sutskever and Hinton 2017]. Each of these models, and others developed by researchers in the field, have their unique design elements, including the specific number of layers, the number of neurons in each layer, the filters contained in the network, and other model characteristics.

It is crucial to note that the development of deep networks was made possible thanks to technological evolution and the rise of GPUs. Given the size of the networks and the volume of data, the time required for weight adjustment for network training was significantly reduced.

Once trained, these networks can be quickly utilized by users. This is the reason why we have many mobile applications today that do not require supercomputers to run algorithms that use this technique. However, many scientists assert that for optimal weight adjustment, a large number of examples is required. This works well in the "Big Data" era, where data generation is rapid and needs to generate useful information, but not in all fields like biology, where data generation and availability may be limited. Therefore, there are three alternatives for using deep learning networks in pattern recognition:

- **Network Training:** Use a deep neural network model and calibrate the neural network weights using new training data. It is essential to remember that given the size of the networks and the slow convergence of the gradient function (weight-error), this use of the technique may require a long training time and a large amount of data [Bengio 2012].
- **Transfer Learning:** Most applications currently using deep neural networks employ this approach. By choosing a proposed network model such as GoogleNet [Szegedy *et al.* 2015], VGG-16 [Simonyan and Zisserman 2014], ResNet [He *et al.* 2016], and ImageNet [Krizhevsky, Sutskever and Hinton 2017], these can be modified, with layers added or removed, but the weights of the remaining neurons are kept the same. This way, the training phase is reduced, and the classification of new data is much faster. This use is also advantageous when there is not enough data for network training.
- **Feature Extraction:** In the third approach, the neural network is not utilized to its full classification potential. During feature extraction, the developer can choose to extract the data representation using the output of one layer as input for another artificial intelligence method such as Naive Bayes or Support Vector Machines (SVMs). This last methodology is not widely used in the literature, but it can prove powerful given the fact that the neural network extracts hierarchical filters from the data [Bengio, Courville and Vincent 2013].

Over the years, various models have been created, among which are the convolutional models. In these networks, 2D convolutional filters are part of the neural network layer composition, transforming the layer's input. This is beneficial in various computer vision problems since the filters function as a means of pre-processing the image. In addition, pooling layers are used to resize the data from one layer to serve as input for a subsequent layer. The combination of convolutional layers, resizing layers, and neurons enables the robust learning of data characteristics, particularly when a large number of examples are presented to the model. However, some researchers criticize the technique precisely because it deviates from traditional machine learning methodologies where the way in which extracted features are known and not done as a black box as occurs in the methodology in question [Goodfellow, Bengio and Courville 2016].

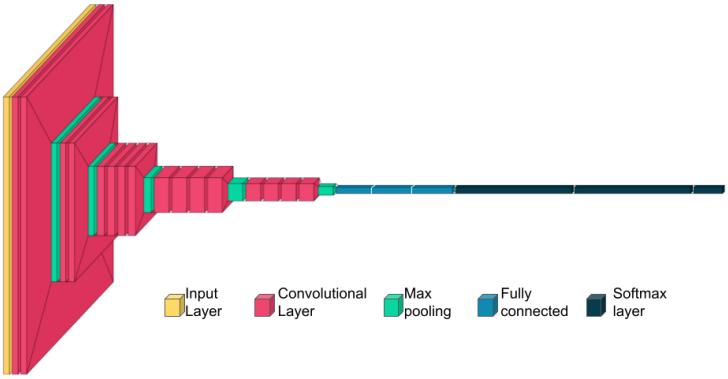
This section does not aim to detail the composition of each one of the existing neural networks in the literature. However, the following section explains of the de CNN (convolutional

neural network) used in this study. Furthermore, various development platforms like Python and Matlab provide support for implementing applications with deep neural networks using libraries such as TensorFlow [Abadi *et al.* 2016], Theano [Bergstra *et al.* 2011], Keras [Chollet 2015], Torch [Collobert, Bengio and Mariéthoz 2002], Caffe ([Jia *et al.* 2014]) found in Matlab 2017(a or b), which will be used in the experiment.

2.5.5 Very Deep Convolutional Network - VGG-19

The VGG-19 is a convolutional neural network model that contains 19 layers, 16 convolutional and 3 fully connected [Simonyan and Zisserman 2014]. Proposed by K. Simonyan and A. Zisserman, the network starts with 64 channels in the first layers and increases by a factor of 2 after each max-pooling layers. The whole structure is represented in Figure 11.

Figure 11 – Very Deep Convolutional Network - VGG-19 layers and connections



Source: Developed by the author

Furthermore, each convolutional layer in the VGG-19 model utilizes a 3x3 receptive field with a stride of 1, and padding is used to preserve the spatial resolution of the previous layer. The max pooling layers have a 2x2 filter size with a stride of 2. Finally, the first two fully connected layers contain 4096 channels each, and the third fully connected layer is used to output the predicted class scores, having a size equal to the number of classes in the dataset (1000 for ImageNet, the dataset it was originally trained on).

The strength of the VGG-19 model lies in its simplicity, achieved by using only 3x3 filters in the convolutional layers and 2x2 pooling kernels in the pooling layers, as well as its depth. However, the model is quite large in terms of trainable parameters and memory usage due to its fully connected layers, which can be a disadvantage in certain applications.

2.6 Classification Models and Validation

Since we used the deep neural network as a feature extraction method in the project and our goal is to compare the label of complex networks models, in this section we aim to explain

the classification models and validation techniques to evaluate the proposed methods.

2.6.1 K-Nearest Neighbors - KNN

The first model is simple yet powerful, the K-Nearest Neighbors (KNN) is a lazy classification model that is able to classify labels or be used in regression tasks [Cover and Hart 1967]. In classification tasks, the model tries to classify each sample according to the majority class of the k_{knn} nearest neighbors. This is why the technique is lazy, for each new sample, the method has to compute the distance (usually the Euclidean one) to all other samples in the training set. The KNN method assumes that items that are alike are found in proximity to each other. The structured process is defined by the algorithm:

1. Compute the distance between the instance being tested and all instances in the training set. The distance is often determined using the Euclidean distance, but other measures such as Manhattan distance or Minkowski distance can also be employed.
2. Sort the calculated distances.
3. Choose the k_{knn} instances from the training set that are nearest to the test instance.
4. For classification, return the most frequently occurring class label among these k_{knn} instances.
5. Selecting an appropriate value for k_{knn} is an essential step in the process.

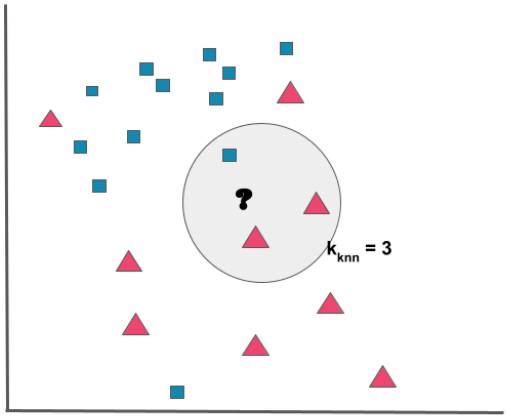
To illustrate, Figure 12 represents the process of a KNN method with $k_{knn} = 3$. Regarding the parameter k_{knn}, a smaller value may result in sensitivity to outliers, while a larger value may include too many points from distinct labels. Due to the nature of the method, it it can be computationally demanding particularly with larger datasets. However, it is intuitively understandable and frequently produces satisfactory results.

2.6.2 Support Vector Machine - SVM

Our second classification model is the Support Vector Machine (SVM) [Sain 1996]. Likely KNN, the well known and widely used model is applied for regression and classification tasks.

In the most basic form, namely binary classification, the SVM transform training data into points within a space, aiming to broaden the margin between two distinct classes as much as possible, i.e., aims to identify a "hyperplane" (in two-dimensional space, this is merely a line) that separates the data into two distinct categories to the greatest extent possible. More specifically, it discovers the hyperplane with the largest margin, that is, the biggest distance between data points from both categories. The hyperplane is shown in Figure 13.

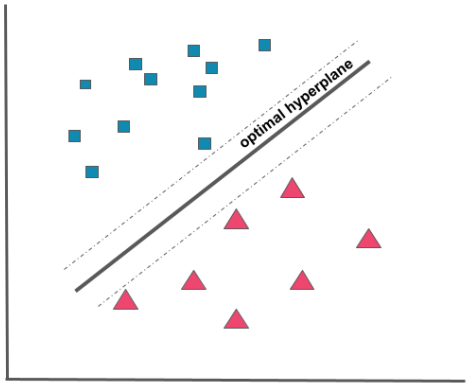
Figure 12 – Example of KNN model with k_{KNN} = 3. In this case, the new example is labeled as a triangle due to the nature of the algorithm.



Source: Developed by the author

Furthermore, different from KNN, an important characteristic of SVMs is their capability to manage high-dimensional data, as well as their adaptability in modeling various types of data. This process is facilitated by a method known as the "kernel trick". A kernel is a function utilized to measure the similarity between two observations. The "kernel trick" leverages these functions to transform complex high-dimensional data into a space that simplifies the prediction process. There's a variety of kernel types, but some of the most commonly used ones include linear, polynomial, and radial basis function (RBF). The kernel choice can impact the final accuracy.

Figure 13 – An example of SVM model illustrating the search for an optimal hyperplane that maximizes the margin between two labels for effective classification.



Source: Developed by the author

2.6.3 Cross-Validation

Finally, to evaluate the proposal we applied K-fold cross validation technique [Kohavi *et al.* 1995]. The K-fold cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. Likely other methods, the technique has a parameter k_{fold} that refers to how many times the model will be evaluated. The algorithm works as follows:

1. Shuffle the dataset randomly.
2. Split the dataset into k_{fold} groups or folds.
3. For each unique group:
    - Take the group as a test data set.
    - Take the remaining groups as a training data set.
    - Fit a model on the training set and evaluate it on the test set.
    - Retain the evaluation score and discard the model.

The result of k-fold cross-validation is often given as the mean of the model skill scores. Furthermore, the standard deviation can be retrieved to understand the variance over each group in the training/test sets. The technique is used to test all proposed methods in this thesis.

Furthermore, as one may have noticed the multiple instances of the parameter k employed across various methods within this project. Indeed, such usage can create ambiguity within the text. To mitigate this confusion and enhance readability, we have elected to append the 'k' notation with the corresponding method's name.

CHAPTER

3

RELATED LITERATURE

To further motivate this work, it is necessary to conduct a literature review of studies that relate to the primary objectives of the work: classification of complex networks.

There are several objectives for pattern recognition techniques in complex networks. While some researchers seek nodes that in some way behave similarly to generate groups (clustering), others focus on the search for metrics and ways to represent the network for comparison. The work in question focuses on the latter technique. Also, here, the emphasis is to analyze the whole graph and not the label each node or groups of nodes; the goal is to label networks (not vertices). Therefore, this section presents works related to the classification of complex networks previously proposed.

As shown in [Costa *et al.* 2011], most real systems can be transformed into complex networks. From this modeling, many researchers have focused their efforts on creating metrics that analyze statistical and dynamic characteristics of the networks [Costa *et al.* 2007]. However, little has been done in the scope of pattern classification to classify the whole complex networks. In addition, most of the works that exist for classification are focused on a specific application. Some of these and their applications are described as follows:

Numerous studies found in the literature are devoted to image classification [Costa 2004, Backes, Casanova and Bruno 2009, Backes, Martinez and Bruno 2011, Backes and Bruno 2013, Ribas, Scabini and Bruno 2022, Scabini *et al.* 2017]. The volume of work related to vision is due to initial studies that proposed ways to model images as complex networks [Costa 2004]. From this, various studies emerged for the classification of static and dynamic textures and shapes using network metrics.

The work of BACKES; MARTINEZ; BRUNO (2011) for example, uses partial deterministic walks to explore the image and generate a network capable of showing important texture features. On the other hand, BACKES; CASANOVA; BRUNO (2009a) models the contour of an object as a complex network and applies dynamic network transformations to generate

small-world characteristics and extract degree measures to compose a feature vector. The method can recognize shapes even in the presence of noise, rotation variance, scale, and degradation.

Similarly, BACKES; BRUNO (2013) connect all the texture pixels with each other, transforming the network into a regular graph. A regular network does not bring relevant information for system characterization and for this reason, cuts are made in the edges to highlight important texture characteristics. The same author proposed another methodology for recognizing contours in 2014 [Backes, Casanova and Bruno 2014]. In this new study, by applying dynamic transformations in the network of the contour, a polygon approximation is computed that retains important shape information from a small number of segments and calculates classic metrics such as the geodesic path to form a vector representation.

A recent work for shape recognition was proposed by SCABINI *et al.* (2017) and after linking the contour points to form a graph, uses the angles formed between the points for the generation of the feature vector. This work was latter applied to classify fish species in [Ribas, Scabini and Bruno 2022]

For dynamic textures, where a sequence of textures produces a video, the work of GONÇALVES; MACHADO; BRUNO (2015) is applied. This models the frames and connects them through a graph model to then extract degree measures, considered simple, but achieves good results in the characterization of the images.

Given that complex networks are capable of modeling most real systems, the need to present independent classification techniques from the origin of the data has arisen. Thus, given networks of different labels, we would be able to classify them with a high success rate looking only at their structure. In the literature, only two works were found with this approach and are most similar to the proposal of this doctorate: [Gonçalves, Martinez and Bruno 2012] and [Miranda, Machicao and Bruno 2016].

The first [Gonçalves, Martinez and Bruno 2012], uses 4000 synthetic networks from 4 different classic models in the literature to test the classification methodology of the proposed method. As most of the classic metrics are related to each other and may not adequately represent a system, GONÇALVES; MARTINEZ; BRUNO (2012) proposed the classification of complex networks based on partial walks with good results in relation to separate classic metrics.

The second [Miranda, Machicao and Bruno 2016], has the advantage of application in real networks such as metabolic, stomatal, and social. This is very important given the proof of the generalization of the method. In addition, experiments were conducted with synthetic networks of various didactic models separating not only the structure but also the average degree of the network. In the work, network features were extracted by transforming them into cellular automata, initially giving each network node a status of alive or dead. The automaton then evolves according to the statuses of the neighboring nodes of each cell and a Life-like rule. This temporal evolution of the network is able to expose dynamic phenomena that can be characterized by

entropy statistics, word size, and Lempel-Ziv. This work is of great importance to the field, but also has a significant limitation: the number of parameters is very high. Despite the good results, there are several parameters necessary for the execution of the algorithm such as: definition of the evolution time of the automaton, definition of the Life-like rule, and the percentage of living nodes at the beginning of the program.

In addition, both works suffer from the lack of comparison methods. Since most of the methods mentioned earlier are focused on a single application and have specific issues such as modeling techniques and definition of edge weights for dynamic evolution among other aspects, the more general methods of complex network classification do not have much alternative for comparison. Therefore, both methods [Miranda, Machicao and Bruno 2016] and [Gonçalves, Martinez and Bruno 2012] use classic metrics to compare the results obtained by feature extraction. From these works, and when necessary from other more applied works, comparisons through accuracy rates classification of data with the proposals made here will be analyzed, aiming for a higher rate in the use of the new methodologies.

The use of complex networks in the literature primarily aims to analyze the structure and formation of groups in a single network. In this work, the objective is to start from several networks that can be labeled and find an efficient and robust way to do so. Moreover, the focus is not on network modeling, a crucial factor in application-focused methods like the ones mentioned above. Some of the works, like [Scabini *et al.* 2017], for instance, would not be valid for networks where the position of the nodes is not relevant, as the calculation of the angle between the nodes would not make sense. Also, the works on computational vision in texture [Backes and Bruno 2013, Gonçalves, Machado and Bruno 2015, Gonçalves, Martinez and Bruno 2012] work with a weight matrix for edge cutting generated by the difference in intensity of the connected pixels, a particularity of the application.

Of the main works related to complex network classification shown so far, it is noticeable that only two of them are strictly related to a general classification of networks: [Gonçalves *et al.* 2012] and [Miranda, Machicao and Bruno 2016]. These are the greatest inspirations for the work, as they use the abstraction power of networks for application in a myriad of areas as shown by [Costa *et al.* 2011]. However, these works have flaws. The proposal of [Gonçalves *et al.* 2012] analyzes only synthetic networks, and the results of combinations of classic metrics with the proposed method are quite similar given that the classic measures were heavily studied over theoretical models.

On the other hand, the work of [Miranda, Machicao and Bruno 2016] analyzed both real and synthetic networks for issues of method generalization. The article [Machicao *et al.* 2018] is an extension of the method applied to text, showing a versatility of general network classification techniques. However, the number of parameters and the time for analyzing these is very large. For each set of networks, an infinity of CA evolution rules were tested, in addition to the parameters of time and the percentage of living nodes.

The proposal aims to create new methods of feature extraction that improve the success rate in the classification of complex networks, not necessarily aimed at an application, having this factor as a consequence, and that is evaluated both in synthetic networks and in real networks. Additionally, alternatives are sought that are not as costly as in the work of [Miranda, Machicao and Bruno 2016] and that do not require so many calibration parameters.

Other literature references will be added along the development of this thesis. This is due the fact that each proposal for network feature extraction has its principles and, therefore, its own inspirations.

In general, two main approaches were considered to compare the proposed method with results in the literature: the LLNA (Life-Like Network Automata) [Miranda, Machicao and Bruno 2016], BP-LLNA [Ribas, Machicao and Bruno 2020], and D-TEP [Zielinski *et al.* 2022], as well as the performance of classical structural metrics analysis.The recent publications have taken the advantage of complex networks and cellular automata-based statistics [Miranda, Machicao and Bruno 2016]. The method converts the original network to cellular automata. The three compared methods use the Life-Like Network Automata rules [Gardner 1970] to evolve the model in order to analyze the network's behavior. While [Miranda, Machicao and Bruno 2016] used entropy measures over patterns to compute the network feature in the original LLNA, [Ribas, Machicao and Bruno 2020] analyzes binary patterns in BP-LLNA. Additionally, [Zielinski *et al.* 2022] uses the density of alive/dead nodes to enhance information on the dynamics of the graph in D-TEP (explained in details in Chapter 8). We compared the results from this proposal with the best results from each LLNA-based approach.

CHAPTER

4

SYNTHETIC, REAL DATASETS AND FEATURE EXTRACTION METHODS

In this study, we evalued the use of application of the proposed methodologies in twelve datasets, in general. Therefore, this chapter describes the synthetic and real networks used for the classification task. In the following explanation, $\bar{k}$ is the average degree, and N is the number of nodes. The datasets were also applied in other literature methods [Miranda, Machicao and Bruno 2016, Ribas, Machicao and Bruno 2020, Zielinski *et al.* 2022].

- **Synthetic Dataset**: In the complex networks literature, some models are well-known and help us to encapsulate many real phenomena. Therefore, this dataset is composed of four different classes of graphs: Random [Erdos and Rényi 1960] (connection probability between two nodes of p= $\bar{k}$ /n), Small-world [Watts and Strogatz 1998] (rewiring probability of p= 0.1;), Scale-free [Barabási and Albert 1999] (with linear and non-linear preferential attachments), and Geographical [Waxman 1988] . Each model is filled with networks with different average degrees: $\bar{k}$ = {4, 6, 8, 10, 12, 14, 16} and different sizes N = {500, 1000, 1500, 2000}. For each model combination, average degree, and size, 100 networks are created, totaling 11200 files.

- **Scalefree Networks**: Among the scalefree models, several researchers have proposed techniques to build networks with power-law degree distributions. This database comprises scale-free synthetic models generated according to the models proposed by Barabási & Albert [Barabási and Albert 1999] and Dorogovtsev & Mendes [Dorogovtsev and Mendes 2013]. The Barabási & Albert model consist of four different classes based on the power-law parameter α = {0.5, 1.0, 1.5, 2.0}, representing linear and non-linear networks. Meanwhile, the Dorogovtsev & Mendes model adds another class to the dataset. The result is a dataset containing 500 networks with a size of N = 1000 and an average degree of $\bar{k}$ = 8.

- **Noisy-synthetic dataset**: By adding or removing edges, we can test if the proposed method is robust to noise in the scalefree dataset. Thus, this dataset modifies the graph's edges according to $\sigma = \{10\%, 20\%, 30\%\}$. For instance, for a σ value of 10%, 5% of the edges are added and 5% are removed. The resulting database is composed of 8 classes, each containing 100 networks.

- **Social Networks**: In an attempt to distinguish the structures of social networks on Google+ and Twitter, 65 graphs of each class were extracted from the SNAP (Stanford Network Analysis Project) platform [Leskovec, Krevl and Datasets 2014] were extracted. Each network was created by examining the social relationships of various users. The goal is to determine whether a simple method can differentiate between the two structures.

- **Metabolic Dataset**: The last six datasets consist of metabolic networks from substrate-product models [Zhao *et al.* 2006]. The models considers metabolites as vertices and their chemical reactions as edges, based on the Kyoto Encyclopedia of Genes and Genomes [Kanehisa *et al.* 2016]. The six datasets are:

  1. **Kingdom dataset**: This set has 160 networks in four classes of the *eukaryota* domain: *animal, plant, fungi, and protist*.
  2. **Animal dataset**: The second dataset contains 14 networks from the following labels: *mammal, bird, fish, and insect.*
  3. **Fungi dataset**: With 15 graphs, the third model has four classes: *saccharomycetes, sordariomycetes, eurotiomycetes, and basidiomycetes*.
  4. **Plant dataset**: This dataset, with three classes (*monocots, green algae, and eudicots*), contains 27 samples.
  5. **Firmicutes-Bacilis-database**: The dataset contains four labels: *Bacillus* (122 samples), *Staphilococcus* (76 graphs), *Streptococcus* (133 samples), and *Lactobacillus* (83 CNs).
  6. **Actinobacteria-database:**: This last set contains three classes: *Mycobacterium, Corynebacterium, and Streptomyces*, with 60, 86, and 53 graphs, respectively.

4.1 Image Feature Extraction Techniques

The local and global methods are analyzed and discussed in this section, from the simplest to the more complex levels of exploration of the images. In summary, different experiments are performed:

1. **Projection**: This simple conversion of the 2D image to a 1D representation is performed by summing the values of each column in the x axis. Therefore, since the nodes with

higher values are located in the first columns, the feature vectors present a descending characteristic. Without any quantitative analysis, one could think that it could be helpful to distinguish random and scalefree graphs due to the difference in the angular coefficient of the vector. Since networks contain different sizes, the projection vector is set to 2500, and when networks are smaller than this size, the remaining values are set to zero. We understand that adding zeros does not influence the results since it means degrees equal to zero.

2. **CLBP**: The Complete Local Binary Pattern method [Guo, Zhang and Zhang 2010] is an extension of the classical LBP [Ojala, Pietikainen and Maenpaa 2002]. Unlike its predecessor, which only analyzes the binary difference among neighbors, the CLBP also uses the difference magnitude and signal information to compute the features. The addition of this information has proven suitable for many applications such as texture recognition [Guo, Zhang and Zhang 2010]. In the experiments, the window among a central pixel used to evaluate the neighbors is 3x3.

3. **Hu Moments**: Hu image moments are comprehended as the weighted average of pixels intensities of a particular region or a function of previous moments. The methods have been used to describe areas of the image, area, and centroid. However, some statistics are known to be invariant to scale, translation, and rotation. Proposed in HU (1962), the Hu moments are a set of seven features constructed based on a function of other moments and encapsulate information such as inertia around the image's centroid and can distinguish mirrored images. The set of seven features is used as input for KNN and SVM analysis.

4. **VGG-19**: important due to their ability to classify images, the deep neural networks are a powerful source of features for classification. Therefore, in this experiment, we use the VGG-19 model with pre-trained weights from the Imagenet dataset [Simonyan and Zisserman 2014]. The weights output in the last polling layer of the model are used as input for SVM in the experiments. The methodology for texture analysis in [Condori and Bruno 2021] has been promising.

CLBP, Hu Moments and VGG-19 are also explained in details in Chapter 2.

CHAPTER

5

EXPLORING ORDERED PATTERNS IN THE ADJACENCY MATRIX FOR IMPROVING MACHINE LEARNING ON COMPLEX NETWORKS

This section draws heavily from the paper submitted to the Physica A. The paper can be accessed through the reference provided: NEIVA; BRUNO (2023).

To ensure clarity and avoid redundant information within the thesis, certain components of the original paper, including the dataset and comparison techniques, have been removed. These modifications aim to enhance the overall readability of the thesis while maintaining the integrity of the research findings.

5.1 Introduction

The emergence of Big Data has sparked interest in structuring data as closely as possible to reality and evaluating it to extract knowledge. Traditional data analysis often reduces complex phenomena to simplified objects. However, technological advances and the ability to gather, process, and store larger amounts of data allow us to explore information from various viewpoints. The capability to create a system with elements and relationships shifts the reductionist approach to an integrative one.

In addition, pattern recognition has been a prominent branch of data science in understanding the world through the perspective of technology. If one were to think of a method that could combine the benefits of artificial intelligence techniques and integrative data analysis, complex networks would be a natural choice. Pattern recognition in complex networks includes a range of algorithms, such as classification and clustering. Although clustering plays a significant

role in the field, classification enables us to recognize diseases, species, structures, and cities, among others. This task is crucial nowadays due to the large amount of data generated that would be too time-consuming and costly to analyze manually. Furthermore, complex networks have a significant advantage for pattern recognition in the era of Big Data, as they can be used to model a wide variety of data, from images to biological systems. It has been demonstrated over the years that most real systems exhibit characteristics of small-world and scale-free networks. The former refers to structures in which elements are connected, on average, by short minimum paths, similar to what occurs in social networks where there is a high probability that a person's friend is also a friend of the person in question. The latter refers to the fact that there are frequently reached elements in a network, such as prominent researchers in a field or influential articles in a text. The latter example illustrates the importance of using graphs to analyze the patterns and structures of a given organization. Therefore, this work's quantitative analysis focuses on classifying synthetic and real networks.

Based on the advantages mentioned above, researchers have successfully used the model for pattern recognition in various applications, such as the classification of static and dynamic textures [Gonçalves, Machado and Bruno 2015], shapes [Ribas, Neiva and Bruno 2019], authorship [Machicao *et al.* 2018], and others. Recently, some works, such as the use of cellular automata in [Miranda, Machicao and Bruno 2016, Zielinski *et al.* 2022, Ribas, Machicao and Bruno 2020], the construction of multidimensional and deep embeddings from networks in [Scabini *et al.* 2022], and the analysis of angles formed in the graph of shapes in [Scabini *et al.* 2017, Ribas, Scabini and Bruno 2022], have distinguished themselves from traditional analysis that uses classical statistical metrics such as degree and clustering coefficient to create a one-dimensional graph representation. The recent efforts of some researchers to find novel techniques that overcome the redundant information found in the composition of descriptors based on classical metrics have produced good results in graph classification. As shown in [Costa 2004], the concatenation of some correlated metrics is sometimes not helpful for pattern recognition. However, have we exhausted all the simplest analyses on complex networks? This study aims to partially answer this question by investigating a simple alternative to represent the graph: the adjacency matrix.

The adjacency matrix of a graph has a one-to-one correspondence with the graph itself. This representation allows for the quick computation of metrics such as degree and co-citation. The matrix has been applied in graph visualization [Behrisch, Schreck and Pfister 2019] or visualizing the temporal evolution of contact networks in [Linhares *et al.* 2021]. However, a visual inspection of the adjacency matrix provides little information for classification, as permutations of the rows or columns do not change the underlying graph. To address this issue, we propose an ordination of the rows of the matrix such that patterns within the matrix become consistent and allow for the distinction of global network labels in synthetic and real networks. In addition, we evaluate various feature extraction methods applied to the sorted adjacency matrix for classification purposes, including data projection, deep learning feature extraction, CLBP analysis, Hu moments, and classical measurements. Our results on synthetic models, metabolic

networks, and social networks demonstrate that our approach can classify networks with over 90% accuracy, outperforming the accuracy rates of compared works [Miranda, Machicao and Bruno 2016, Ribas, Machicao and Bruno 2020, Zielinski *et al.* 2022].

5.2 Proposed Approach

This section presents an alternative representation of the graph, the adjacency matrix, and describes some complex network metrics.

A complex network is a set G = {V,E} where V represents the nodes (or vertices) of the system, V = $\{v_1, v_2, v_i, v_j, v_N\}$, and E represents the edges relating vertices according to an established condition, E = $\{e_{ij} = (v_i, v_j)|$ if v_i and v_j are connected $\}$. As mentioned before, complex networks are derived from graph theory; therefore, the classical definition of the structure is the same as its precursor. However, even though some researchers believe that complex networks are more than just graphs due to some topological characteristics capable of incorporating complexity into the model, we can (and will) consider both theories the same. Thus, a matrix form represents a graph, named adjacency matrix. For an undirected network G with N vertices, the matrix A is computed by:

$$A(v_i, v_j) = \begin{cases} 1, \text{ if } v_i \text{ and } v_j \text{ are connected} \\ 0, \text{ otherwise} \end{cases} \tag{5.1}$$

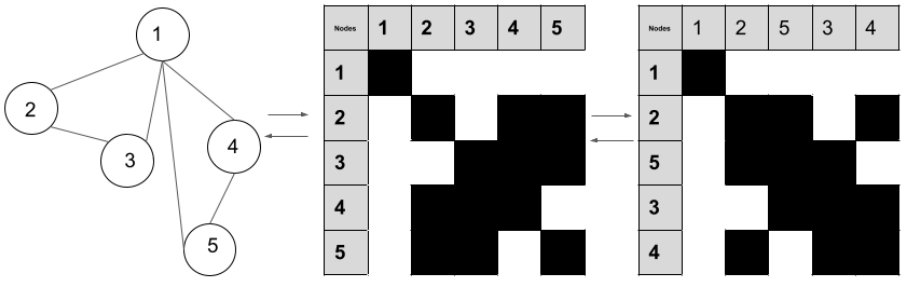
This means that every edge e_{ij} and all nodes v_i are represented in A. Furthermore, the adjacency matrix is symmetric for an undirected network, and its diagonal contains only zeros if it has no self-loops. Matrix A can compute a series of statistics such as degree, co-citation ($C = AA^T$), and bibliographic coupling ($B = A^T A$). Co-citation and bibliographic coupling are especially used for directed networks to create a projection and transform the non-symmetric, which are important metrics for the analysis of references and, for instance, to identify essential nodes in a network.

However, it can be observed that in a graph representation, the nodes usually have no order. Therefore, two adjacency matrices can represent the same network, as shown in Figure 14. Regarding the lack of spatial order in the matrix, this work proposes a sorting technique for the matrix to find patterns in a graph. Nevertheless, the following section aims to explain this transformation on A and the reason for its choice.

5.2.1 Adjacency Matrix Based Signature for Complex Networks

One of the basic characteristics that can be analyzed in a graph is the degree of a vertex i, k_i. The metric is related to the number of connections a vertex has with other nodes in the graph. Also, a straightforward analysis considers the highest-degree node as the most important in the

Figure 14 – A single graph can have several correct representations. While the first image represents the vertices in circles connected by a line, the other two are in matrix form. As one can see, the visual patterns are different even though both matrices correspond to the same information; exploiting the matrix could be difficult. However, a sorting algorithm would allow machine learning methods to classify the patterns.



Source: Developed by the author

system. Therefore, the first sorting of matrix A is to order all nodes in decreasing order based on their degrees. It means that the node with a higher degree will populate the first row and column of A.

However, several nodes have the same degree, and the goal of creating a unique network representation will fall apart. Thus, we propose the following rule: order the rows according to a descending sort of the degrees of the nodes. If two nodes at rows i and j contain the same degree, we use the betweenness, clustering, closeness and eigenvector (in this order) as tie-breakers.

This simple ordination will ensure that nodes with more connections occupy the first positions in the new A' matrix. Also, due to the second rule, for a single network, matrix A' will be unique for a given input. Since we are analyzing statistics dependent on the matrix arrangement, we need to ensure that the matrix image is always the same, regardless of how A was first initialized.

A', the ordered version of an adjacency matrix A, will be evaluated in this study in two ways: visually and quantitatively. Finally, the final signature is evaluated based on several descriptor methods described in Section 5.4.2. The model's simplicity, in contrast with the results, shows that there is still room to explore basic representations of systems, such as the adjacency matrix.

5.3 Experiments

To assess the proposed approach, we applied the methodology to twelve datasets, as described in Chapter 4 These datasets encompass a variety of synthetic and real complex network models, providing valuable insights into the effectiveness of the sorting method.

Additionally, the experiments conducted in this study compare the results obtained from the proposed approach with those of other methods, LLNA, BP-LLNA, D-TEP, and classical

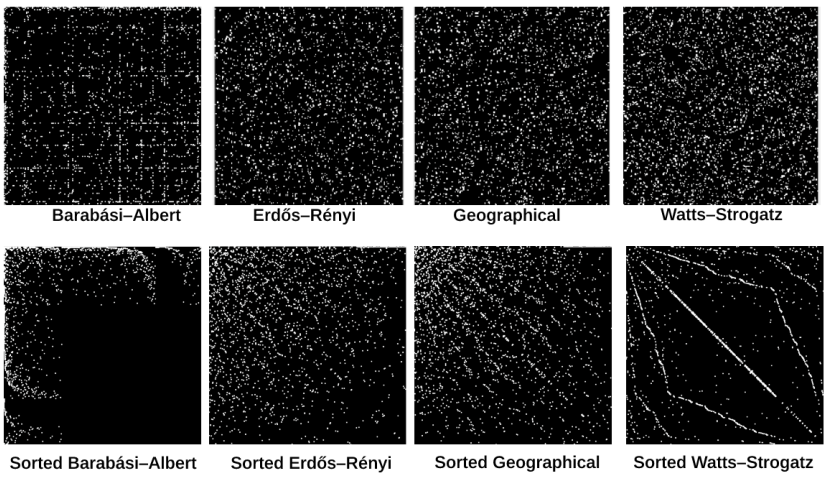
methods. The feature vector used for classical metrics comprises a concatenation of degree assortativity, diameter, average closeness, average eccentricity, average betweenness, average degree, and average clustering, all of which are defined in Chapter 2

5.4 Results and Discussion

5.4.1 Visual Analysis of the Ordered Adjacency Matrix of Synthetic Networks

First, we compute the adjacency matrix in all networks of a given dataset and order them to produce matrices where hubs are encountered at the beginning of the graph. The ordination ensures that the final A' will remain independent of the original arrangement of vertices in the original A. The primary analysis performed in this study is a visual inspection of A'. By analyzing the distribution of white pixels (connections in A'), one can determine if the characteristics of synthetic models are enhanced with visual inspection. Figure 15 shows the effect of the methodology on the synthetic dataset (all images are dilated by a small amount to improve visualization).

Figure 15 – The first row shows the unsorted adjacency matrices, while the second shows the result of the proposed approach. The application enhances the features of each model compared to unsorted matrices. White points represent edges, and image pixels are dilated by a small amount to improve visualization.

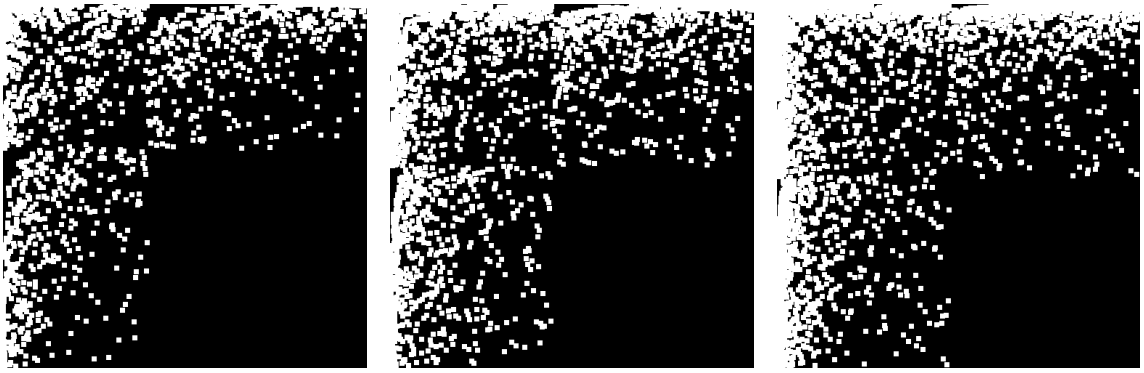


Source: Developed by the author

Figures 16, 17, 18 and **??** show some samples from each synthetic class, ordered as suggested in Section 5.2.1. From visual inspection, we can observe some of the theoretical characteristics of these four methods represented in the ordination.

First, Figure 16 shows examples of Barabási–Albert networks which is a method for producing scalefree graphs [Barabási and Albert 1999]. Therefore, to understand the output image, we must understand the model. The scalefree networks proposed by Barabási–Albert

Figure 16 – After the application of the proposed method, the patterns of Barabási–Albert model are highlighted. All networks contain 500 vertices.



(a) Barabási–Albert, $\bar{k} = 4$ (b) Barabási–Albert, $\bar{k} = 6$ (c) Barabási–Albert, $\bar{k} = 8$

Source: Developed by the author

create graphs by starting with an arbitrary network of N_0 nodes. Then, new nodes are added and linked to c existing nodes, where $c \leq N_0$. Furthermore, the probability of connecting to a specific node is proportional to its degree, meaning a vertex with a higher degree has a greater chance of being connected to the new node.

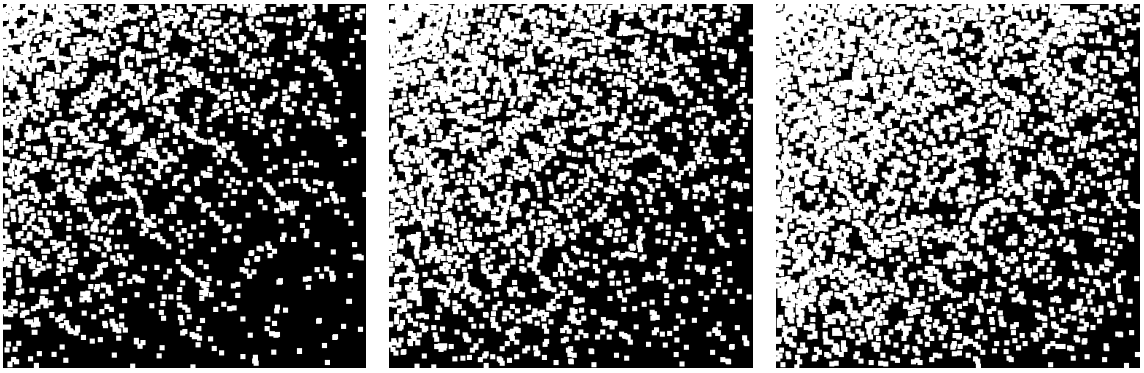
Consequently, it can be said that there is preferential attachment between new vertices and those that already have many connections. This results in scale-free networks that exhibit a degree distribution following a power law of the form $P(k) \propto k^{-3}$. An important example of this network is the World Wide Web, where heavily connected vertices are known as "hubs" and are crucial for understanding the system and extracting statistics and information about its dynamics.

With this in mind, we can now return to the visual inspection in Figure 16. The first aspect to notice is the high density of points in the upper left corner of the image, where each white point represents an edge. Furthermore, the ordination guarantees that nodes with more links are located in the first rows and columns. The lower right corner is more sparse, indicating that nodes with a lower degree are the majority of the graph, which is consistent with the power law degree distribution of the model.

Our second model is the Erdős–Rényi network [Erdos and Rényi 1960]. To create an Erdős–Rényi model, N nodes are connected according to a probability p of an edge existing between any pair. As p is a parameter of the method, graphs with a lower p are weakly connected, while a higher p results in a graph with many edges. The connection between two nodes is performed with a probability independent of the nodes themselves, which is different from what occurs in the previous method. These characteristics are visible in Figure 17. Furthermore, the points are spread all over the image, indicating a level of uniformity compared to other synthetic models. As expected, the upper left corner contains more points due to the ordination, but the difference is smaller in random networks compared to scalefree networks.

A synthetic model to simulate this category was proposed by WATTS; STROGATZ in

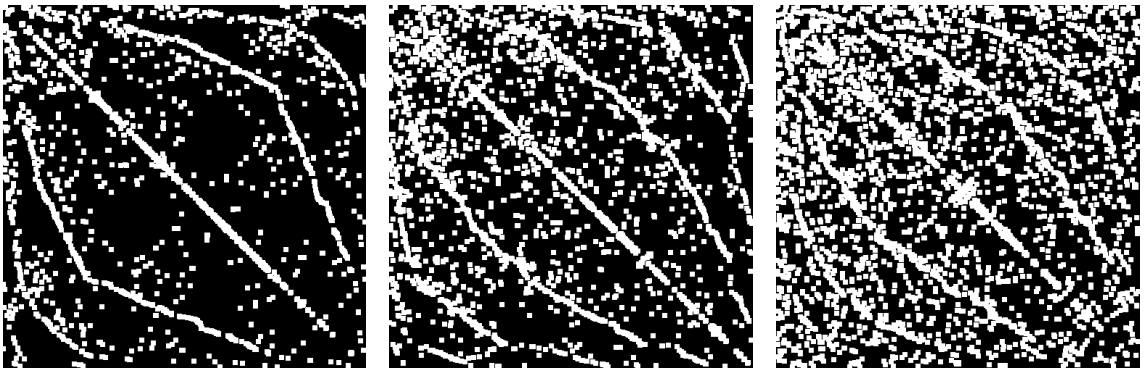
Figure 17 – The Erdős–Rényi model images show a wide distribution of points along all the images strengthening the characteristics of the randomness of the network. All networks contain 500 vertices.



(a) Erdős–Rényi, $\bar{k} = 4$ (b) Erdős–Rényi, $\bar{k} = 6$ (c) Erdős–Rényi, $\bar{k} = 8$

Source: Developed by the author

Figure 18 – It is interesting to notice how patterns of hubs are represented in Watts-Strogatz model images. Even with the increase in the number of nodes, the basic pattern remains. All networks contain 500 vertices.

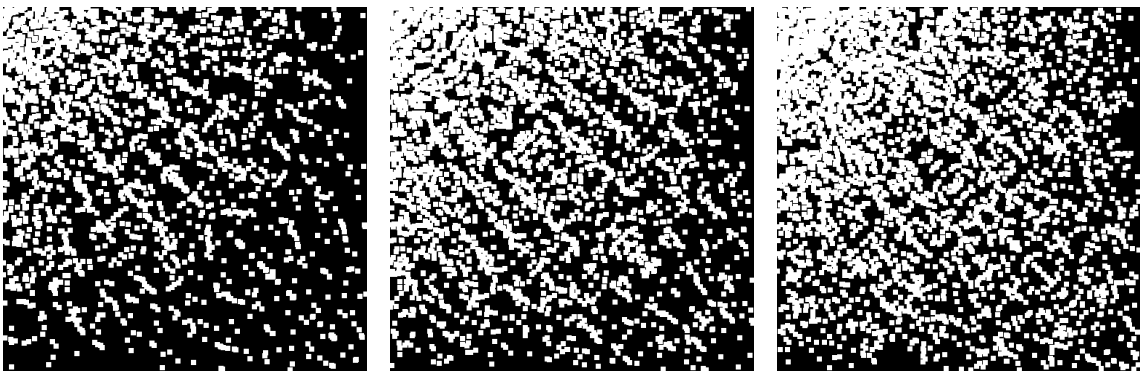


(a) Watts-Strogatz, $\bar{k} = 4$ (b) Watts-Strogatz, $\bar{k} = 6$ (c) Watts-Strogatz, $\bar{k} = 8$

Source: Developed by the author

1998. It initiates the graph as a regular network with N nodes and average degree of k, where first, a node is linked to $k/2$ neighbors on each side. Then, each edge on the right side of a node is rewired with probability $0 \leq \beta \leq 1$ (β is a parameter of the method). In the end, the algorithm delivers connected nodes with high clustering coefficient and low diameter of the metric graph. In practice, it means a shortest-path linking any two vertices in the system. Finally, let us analyze the output images in Figures 18, which represent the ordered adjacency matrices of three model samples. The images show an interesting pattern for the model; it is possible to see a diamond shape that can be explained by the rewiring method applied to the $k/2$ neighbors of each node and the probability used ($\beta = 0.1$). The pattern shape also allows us to understand the small-world characteristic. If one node cannot reach another in a straight line in one step, connecting them will require a small path. This characteristic is visible by the linear line in the pattern and the knowledge that the initial setup of the network is a regular ring.

Figure 19 – The Geographical model. In the images, one could check that the patterns within the model are maintained. All networks contain 500 vertices.



(a) Geographical, $\bar{k}$ =4 (b) Geographical, $\bar{k}$ =6 (c) Geographical, $\bar{k}$ =8

Source: Developed by the author

Finally, the last synthetic model does not have a solid statistical characteristic, so we cannot assume much about the patterns presented in Figure 19. Geographical networks are constructed based on the spatial properties of the system. Although there are no strict rules, the patterns within the model remain. Additionally, one can notice a pattern similar to the random networks in Figure 17. This similarity may indicate difficulty in distinguishing both classes, which can be evaluated by statistical methods.

5.4.2 Quantitative Analysis Based on Adjacency Matrix Signature

The visual inspection showed that the approach proposed in Section 5.2.1 was feasible for network characterization. However, numerical analysis was also performed using signatures described in this section. In this context, we not only evaluate synthetic networks but also real and scalefree models. We compare the results obtained by this simple approach with results available in LLNA-based methods [Miranda, Machicao and Bruno 2016, Ribas, Machicao and Bruno 2020, Zielinski *et al.* 2022] and classical structural metrics. Tables 1 and 2 show the AUC (accuracy) using SVM. Additionally, cross-validation using k-fold (k = 10), is used to analyze generalization. It is important to notice that we evaluate several features extracted from images produced by the ordination of the adjacency matrix of each complex network.

5.4.2.1 Results and Discussion

Before discussing the results presented in Tables 1 and 2, it is important to revisit the objective of the proposal. Although the feature extraction methods are not the study's novelty, the subject of analysis is. The projection method, which is one of the most basic approaches, captures the degree structure of the network by sorting nodes based on their centrality values. Classical metrics, widely used in the literature to represent nodes, essentially analyze the graph's structure and how the elements are connected. It is worth noting that most of the measures used in the

Table 1 – Classification results comparing the proposed sorting method with the application of feature extraction techniques over the unsorted adjacency matrix.

	VGG-19		PROJECTION		CLBP		HUMOMENTS	
	SORTED	UNSORTED	SORTED	UNSORTED	SORTED	UNSORTED	SORTED	UNSORTED
Synthetic	**99.86 ± 0.12**	99.7 ± 0.15	**96.44 ± 9.56**	83.65 ± 0.38	**82.00 ± 0.76**	80.73 ± 1.09	**53.10 ± 1.15**	43.85 ± 1.50
Scalefree	**100.00 ± 0.00**	98.6 ± 3.2	**100.00 ± 0.00**	94.80 ± 2.12	**98.74 ± 1.60**	97.30 ± 0.4	**88.08 ± 7.85**	74.40 ± 4.53
Noise 10%	**100.00 ± 0.00**	99.3 ± 2.7	**96.67 ± 1.99**	78.40 ± 3.65	**99.26 ± 0.64**	97.78 ± 0.49	**97.78 ± 1.91**	54.08 ± 1.72
Noise 20%	**100.00 ± 0.00**	99.0 ± 0.90	**97.67 ± 1.42**	79.00 ± 1.47	**99.33 ± 0.63**	98.13 ± 0.48	**93.63 ± 2.66**	60.00 ± 3.49
Noise 30%	**99.9 ± 0.76**	98.7 ± 1.5	**97.41 ± 1.29**	78.15 ± 3.97	**99.01 ± 0.00**	97.90 ± 0.24	**92.22 ± 1.88**	62.85 ± 4.25
Social Network	91.47 ± 9.21	**96.8 ± 11.2**	87.69 ± 7.6	**89.91 ± 3.95**	**92.31 ± 8.88**	78.34 ± 9.01	**54.62 ± 11.72**	51.92 ± 2.29
Kingdom	**99.6 ± 3.47**	97.5 ± 5.6	**100.00 ± 0.00**	89.38 ± 2.39	**90.00 ± 2.89**	78.13 ± 6.25	**91.88 ± 2.39**	87.50 ± 4.56
Animal	**93.20 ± 4.22**	86.9 ± 8.9	**95.00 ± 7.64**	90.38 ± 2.56	85.71 ± 11.66	**87.50 ± 6.84**	**91.07 ± 3.57**	71.43 ± 14.29
Fungi	**81.95 ± 2.38**	72.25 ± 6.58	**88.33 ± 13.02**	45.00 ± 11.39	**51.67 ± 6.38**	51.67 ± 10.00	**53.33 ± 10.89**	35.00 ± 10.00
Plant	79.2 ± 4.13	**82.91 ± 7.67**	**95.00 ± 15.00**	67.26 ± 17.21	**85.71 ± 11.66**	74.40 ± 13.10	**67.86 ± 21.43**	66.67 ± 6.73
Firmicutes-Baccilis	**99.9 ± 0.27**	94.56 ± 2.64	**96.14 ± 1.91**	83.61 ± 1.57	**89.40 ± 2.59**	84.59 ± 2.31	**80.23 ± 3.39**	73.49 ± 2.96
Actinobacteria	**99.5 ± 0.87**	97.02 ± 3.17	**99.00 ± 2.00**	92.97 ± 2.57	**93.46 ± 3.44**	92.47 ± 1.88	**94.98 ± 1.99**	90.97 ± 2.53

Source: Developed by the author

Table 2 – Classification results comparing the proposed method with literature methodologies.

	PROPOSED				LITERATURE			
	VGG-19	PROJECTION	CLBP	HUMOMENTS	LLNA	BP-LLNA	D-TEP	STRUCTURAL MEASURES
Synthetic	99.86 ± 0.12	96.44 ± 9.56	82.00 ±0.76	53.10 ± 1.15	99.99 ± 0.00	**100.0 ± 0.00**	**100.0 ± 0.00**	**100.00 ± 0.00**
Scalefree	**100.00 ± 0.00**	**100.00 ± 0.00**	98.74 ± 1.60	88.08 ± 7.85	98.3 ± 0.07	99.52 ± 0.19	**100.0 ± 0.00**	**100.00 ± 0.00**
Noise 10%	**100.00 ± 0.00**	96.67 ± 1.99	99.26(0.64)	97.78 ± 1.91	99.98 ± 0.00	**100.0.0 ± 0.00**	**100.0 ± 0.00**	**100.0.0 ± 0.00**
Noise 20%	**100.00 ± 0.00**	97.67 ± 1.42	99.38 ± 0.63	93.63 ± 2.66	99.97 ± 0.01	99.98 ± 0.00	**100.0 ± 0.00**	**100.0.0 ± 0.00**
Noise 30%	99.9 ± 0.76	97.41 ± 1.29	99.01 ± 0.00	92.22 ± 1.88	99.95 ± 0.01	99.99 ± 0.00	99.75 ± 0.00	**100.0.0 ± 0.00**
Social Network	91.47 ± 9.21	87.69 ± 7.6	92.31 ± 8.88	54.62 ± 11.72	92.00 ± 1.00	**93.40 ± 0.92**	92.5 ± 0.50	92.31 ± 5.13
Kingdom	**99.6 ± 3.47**	**100.00 ± 0.00**	90.00 ± 2.89	91.88 ± 2.39	93.10 ± 5.38	97.44 ± 3.98	96.24 ± 0.35	96.61 ± 4.33
Animal	93.20 ± 4.22	95.00 ± 7.64	85.71 ± 11.66	91.07(3.57)	77.25 ± 16.29	84.87 ± 15.25	**100.0 ± 0.00**	83.71 ± 15.29
Fungi	81.95 ± 2.38	**88.33 ± 13.02**	51.67 ± 6.38	53.33 ± 10.89	54.58 ± 19.38	76.17 ± 17.45	81.00 ± 4.38	54.90 ± 15.39
Plant	79.2 ± 4.13	**95.00 ± 15.00**	85.71 ± 11.66	67.86 ± 21.43	69.70 ± 4.67	74.81 ± 5.64	79.58 ± 2.12	54.19 ± 9.17
Firmicutes-Baccilis	**99.9 ± 0.27**	96.14 ± 1.91	89.40 ± 2.59	80.23 ± 3.39	84.63 ± 2.00	98.30 ± 1.17	95.73 ± 0.34	95.67 ± 0.59
Actinobacteria	**99.5 ± 0.87**	**99.00 ± 2.00**	93.46 ± 3.44	94.98 ±1.99	91.48 ± 1.60	95.13 ± 1.22	97.65 ± 0.29	93.16 ± 0.70

Source: Developed by the author

experiments rely on the study of shortest paths, which can lead to redundancy in the combined analysis [Costa 2004]. In addition, the CLBP method, which is a texture extraction technique, analyzes the image locally and can extract essential features as the exploration moves away from the upper left of the image, where there is a usually a concentration of hubs. The Hu moments method, which is invariant to scale, rotation, and translation, analyzes the statistics of regions in the image, such as inertia and mirroring. It yields significant results based on its ability to quantify the global dispersion of the pixels in the proposed adjacency matrix. Finally, VGG-19, a convolutional neural network, builds hierarchical features and can create robust representations of the input images, resulting in high classification rates, as seen in the results.

First, Table 1 shows that the proposal demonstrates substantial enhancements in classification rates when comparing the same experiment over unsorted adjacency matrices across various datasets, highlighting its potential to classify networks. When applied to the CLBP method, the classification accuracy of social networks increased from 78.34% to 92.31%. This notable improvement highlights the efficacy of the ordination method in increasing the performance of existing classification techniques.

In addition to the CLBP method, the ordination methodś impact on neural network classification is present in Table 1. The sorting proposal, an integral part of the proposed method, was able to enhance the classification scores in 10 out of 12 datasets. For instance, in the Fungi dataset, the accuracy increased from 72.25% to 81.95% when the ordination method was

implemented.

Furthermore, the ordination method's influence extends beyond neural network classification and CLBP, as evidenced by its impact on the projection feature extraction process. In the Plant dataset, the application of the ordination method led to a 28.73% increase in the classification rate. These results allow us to understand the advantages of sorting rows and columns beyond visual inspection.

Regarding the results in Table 2, the five synthetic datasets are constructed based on theoretical constraints, which are well captured by structural measures. This knowledge is shown by the high classification rate obtained by structural statistics and projection over synthetic and scalefree datasets. Moreover, structural measures have been shown to be robust to noise. All synthetic models are classified by structural measures with 100% accuracy. However, the proposed method with VGG-19 also achieves 100% AUC in four out of five datasets, with the only exception being the 4-classes synthetic models, which is classified with 99.86% accuracy.

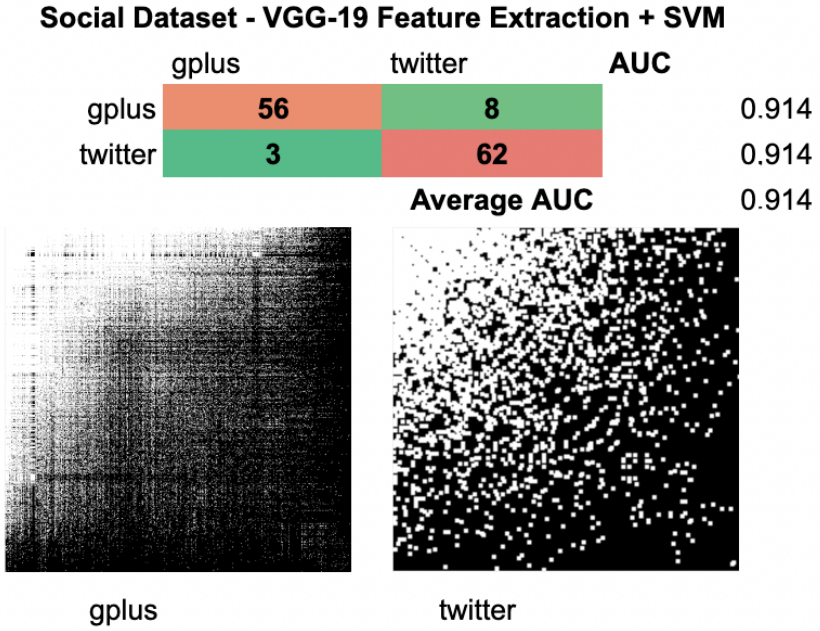
Ranging from 96.44% to 100.00% accuracy, the projection of A' in the x axis can also distinguish the synthetic models very well due to their degree distribution. The outcomes achieved using the CLBP method are not only significant for real-world network classification but also hold substantial importance for the classification of synthetic models. However, when considering the use of Hu moments for feature extraction, their performance is found to be less expressive when compared to other proposed methods and existing techniques in the literature.

It is crucial to highlight that all LLNA-based methods exhibit exceptional accuracy for synthetic models, even when subjected to the presence of noise. Nonetheless, a key advantage of the proposed ordination method of the adjacency matrix lies in its lack of parametrization. In contrast, the method works by transforming the original network into a cellular automata, but the selection of the best set of Life-Like rules introduces a significant computational cost and proves to be time-consuming for each new dataset tested [Miranda, Machicao and Bruno 2016, Ribas, Machicao and Bruno 2020, Zielinski *et al.* 2022]. This distinguishing feature of the proposed method offers a more efficient and streamlined approach to global network classification without compromising on accuracy. The approach evolves the tessellation based on a Life-like rule.

The proposal is able to classify patterns of social networks with 92.31% AUC with CLBP feature extraction. However, the result is slightly better when considering BP-LLNA. The latter method classifies Gplus and Twitter classes with 93.40% accuracy. In addition, confusion matrices show the exact mistakes of classification. Gplus and Twitter had been confused 11 times: eight Gplus networks were predicted as Twitter, and three were misclassified on the opposite side (as seen in Figure 20).

The proposed ordination method significantly improved the classification accuracy of metabolic networks, except for the Animal dataset, which achieved the highest accuracy with the D-TEP algorithm. For instance, the Kingdom dataset attained a 100% accuracy rate using the

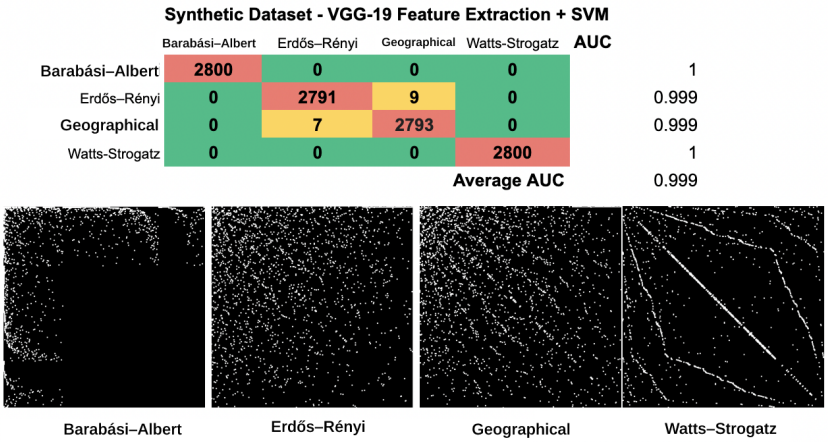
Figure 20 – Confusion matrix and AUC of social dataset for VGG-19 and SVM classification.



Source: Developed by the author

projection technique, and 99.6% accuracy using the VGG-19 convolutional neural network. The Firmicutes-bacillis and Actinobacteria datasets also achieved approximately 99% classification rate using the convolutional neural network approach, demonstrating the effectiveness of the proposed method compared to existing techniques in the literature.

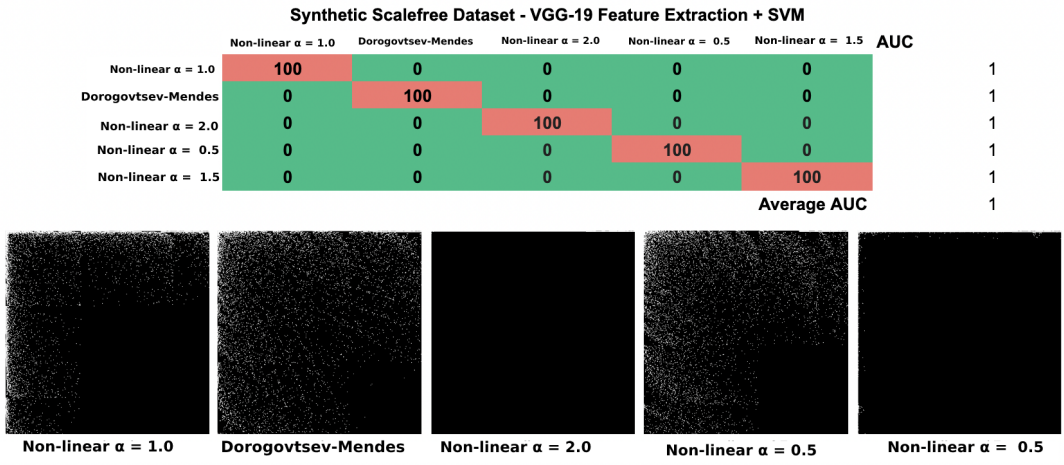
Figure 21 – Confusion matrix and area under curve(AUC) of synthetic dataset. The results are obtained by extracting features of VGG-19 neural network and classify then using SVM.



Source: Developed by the author

Regarding the confusion matrices of VGG-19 feature extraction and SVM classification of the proposal in Figures 21 and 22, the results highlight the need to understand the high accuracy of 99.86% and 100% obtained for the synthetic and scalefree datasets, respectively, as presented in Table 2. As expected from the visual inspection results, the Erdős–Rényi and geographical networks are mistaken due to their less distinct structural features. Erdős–Rényi

Figure 22 – Confusion matrix and AUC of scalefree dataset for VGG-19 and SVM classification.



Source: Developed by the author

and geographical networks were confused 16 times: seven geographical networks were predicted as Erdős–Rényi, and nine geographical were classified as random graphs. On the other hand, Barabasi-Albert and Watts-Strogatz maintain the maximum accuracy in the classification.

One limitation with simple hand-crafted descriptors such as projection and local pattern extractors is that the standard deviation is very high, especially for the Fungi and Plant datasets, ranging from 6.38 to 15.00 in these two datasets. This limitation is observed not only for the proposed approaches but also for structural and LLNA methods. In summary, all the results underscore the efficacy of the new approach in enhancing global network classification across a diverse range of datasets.

5.5 Conclusion

This paper investigate the use of a fundamental graph feature, the adjacency matrix, as the primary input for network analysis. Although the adjacency matrix provides a one-to-one representation of a network, a simple permutation of its rows retains all systems properties unordered. To address this issue, the paper proposes an ordination of the matrix based on the centrality of the nodes, allowing the generation of a single representation suitable for image analysis. Visual analysis demonstrates that the proposed data transformation accentuates the patterns within the models. Moreover, quantitative analysis is conducted on twelve different datasets, including seven real systems, and the results indicate that, in general, all proposed approaches and compared methods yield satisfactory classification accuracies.

Three aspects deserve further examination: correct classification rate, standard deviation, and method complexity. It is evident that for synthetic data, basic metrics are enough to represent the networks. However, for real datasets that typically exhibit high standard deviation due to the variability of the data, LLNA-based methods present the disadvantage of lengthy parameter

fitting times. Therefore, the efficacy of deep neural networks for image classification (with the adjacency matrix treated as an image in this experiment) emerges and demonstrates promising results, predominantly in metabolic and social network datasets.

In conclusion, the proposed approach of sorting and analyzing a basic network representation is well-suited for further investigation in practical domains, despite the limitations imposed by increasing graph size, which consequently enlarges the image and increases computational time. The findings suggest that a straightforward representation like the adjacency matrix serves as a valuable source of information for network classification. As future work, the development of new methodologies that incorporate additional information to the edges could potentially enhance the matrix's patterns for classification, thus further improving network analysis.

CHAPTER

6

ENHANCING INFORMATION ON ORDERED PATTERNS FOR COMPLEX NETWORK CLASSIFICATION THROUGH EXPLORATION OF VERTEX SIMILARITY

As the preceding chapter, this section adopts a paper-like format. As a result, some redundancy may be present in the text, emanating from the central aim of this thesis: the classification of complex networks. The chapter follows the idea of using a matrix-based form of the graph with additional information in each cell of the grid.

6.1 Introduction

Driven by the enthusiasm of modeling data to closely mirror real-world situations and dissecting it for valuable insights, the evolution of hardware and software technologies has enriched the resources available for data storage and management. This advancement facilitates a transition from traditional basic entity analysis to a more intricate perspective, enabling the exploration of an interacting ensemble of elements, rather than single, isolated units.

Within this context, Complex Networks (CNs) emerge as a natural tool to examine these evolved systems. Particularly for pattern recognition methods, CNs become a key facet of data science, integral in deciphering our world from a technological standpoint. This model boasts an ability to represent an array of phenomena, from social networks to texture classifications [Costa *et al.* 2011].

Furthermore, the manifestation of small-world networks becomes apparent in scenarios where components interconnect through relatively short paths, much akin to social networks. Conversely, when networks comprise elements accessed frequently - such as pivotal researchers

within a discipline or impactful articles within a body of text - it suggests the networks are portraying scale-free characteristics. This emphasizes the importance of using graphs to probe an entity's patterns and structures. Therefore, this research zeroes in on the categorization of artificial and real networks through quantitative analysis.

Exploiting these advantages, researchers have successfully applied the model for various pattern recognition applications, from classifying static and dynamic textures to shapes, authorship, and others [Costa *et al.* 2007]. Recent works - incorporating cellular automata, constructing multi-dimensional and deep network embeddings, and analyzing angles formed in shape graphs - have diverged from traditional analyses that rely on classical statistical metrics for one-dimensional graph representation. The efforts of some researchers to innovate techniques that outdo the redundant data inherent in descriptors based on classical metrics have yielded positive results in graph classification. As depicted in several studies, the combination of certain correlated metrics may not always enhance pattern recognition.

However, a previous study [Neiva and Bruno 2023] investigates the adjacency matrix of the CNs, which maintains a direct correlation with the graph itself. The impressive outcome of this investigation with regards to complex network model classification, from synthetic to real models, outperforms several studies such as LLNA and BP-LLNA [Miranda, Machicao and Bruno 2016, Ribas, Machicao and Bruno 2020].

In essence, the previous proposed method systematically arranges the matrix's rows to maintain consistency within the matrix patterns, easing the differentiation of global network labels in artificial and real networks. However, the utilized matrix comprises 0's and 1's, due to the nature of the unweighted evaluated models.

As a continuation of the previous research, this proposal seeks to answer a remaining question: can vertex-vertex information enhance pattern recognition of complex networks? To this end, this study proposes incorporating vertex-simmilarity features to improve CN classification and applies image feature extraction methodologies to the colored sorted adjacency matrix. These methodologies comprise data projection, deep learning feature extraction, CLBP analysis, Hu moments, and classical measurements. Similar to the preceding methods, our results on synthetic models, metabolic networks, and social networks corroborate our approach, classifying networks with an accuracy rate exceeding 90%, surpassing the accuracy of compared studies.

6.2 Proposed Approach

As mentioned before, the goal is to extend the exploitation the adjacency matrix as the main input for complex network characterization. The adjacency matrix A is an one-to-one representation of the graph G where G = {V,E}, with V signifying the system's nodes (or vertices), V = $\{v_1, v_2, v_i, v_j, v_N\}$, and E denoting the edges that connect vertices based on a set rule, E = $\{e_{ij} = (v_i, v_j) |$ if v_i and v_j are interconnected $\}$. For simple models, the complex networks has no

weight information within is edges and it is characterized by matrix A where each element $A_{i,j}$ is equal to one with vertices i and j are connected and 0 otherwise.

However, as one might notice and supported by [Neiva and Bruno 2023], the position of nodes i and j can move along the matrix and the features of the networks such as diameter, betweenness, closeness, shortest paths among others, remains the same. Thus, without any further preprocessing method, it is hard to rely the image of the adjacency matrix as an input for feature extraction method. Any permutation of the network changes the texture distribution of the model.

However, other matrices can be explored from the original graph. Thus, this proposal, instead of using the classical binary adjacency matrix, construct a new matrix C based on Jaccard Index [Sorensen 1948]:

$$C(v_i, v_j) = j_{i,j} \tag{6.1}$$

where $j_{i,j}$ is the Jaccard Index between nodes i and j [Sorensen 1948]. This metric was chosen to create a weight for vertices relationships. Note that, it is not relied on the edges of networks. In the context of graph theory, the Jaccard Index is used as a measure of the similarity between two nodes. Usually, it is used in the context of finding groups of similar nodes in the graph, community detection, link prediction.

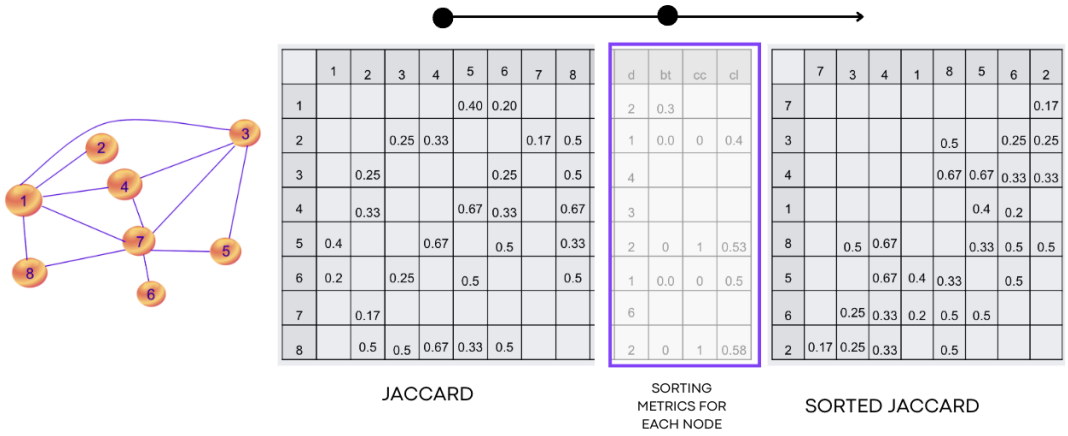
The index is computed as follows. From two nodes i and j, the metric is defined as the size of the intersection of these sets divided by the size of the union of these sets. In other words, it is the fraction of the total unique neighbors that both nodes share:

$$j_{i,j} = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|} \tag{6.2}$$

In Equation 6.2, function N(i) outputs the set of neighbors of node i. Likely in the previous adjacency matrix, C, for an undirected network, is symmetric, and its diagonal comprises solely zeros in the absence of self-loops.

Furthermore, after the construction of C for each network, the matrix must also be reordered as proposed in [Neiva and Bruno 2023]. Since a rudimentary analysis often identifies the node with the highest degree as the most pivotal within the system. The reordering method follows the following rule: sort the rows based on a descending order of the node degrees. In the event that two nodes at rows i and j possess identical degrees, we will employ betweenness, clustering, closeness, and eigenvector (in this sequence) as tie-breakers. In this way, we combine the power of complex networks models, aligned by known image feature extraction technique by the exploration of colored matrix C. Figure 23 shows the example of C matrix transformation.

Figure 23 – Proposed Approach: Firstly, the Jaccard Index matrix is computed for each pair of nodes. Next, the matrix is sorted based on the centrality measures of each node.



Source: Developed by the author

6.3 Experiments

We applied the established methodology to a selection of twelve datasets as discussed in Chapter 4 to evaluate the methods. The experiments carried out in this investigation contrast the results derived from the suggested method with outcomes from other techniques, such as LLNA, BP-LLNA, D-TEP, and traditional methodologies. Furthermore, as an extension of [Neiva and Bruno 2023], we compare the method with the previous approach.

Regarding the datasets, they consist of a blend of artificial and real-world complex network models, offering valuable data concerning the efficiency of the sorting technique. Moreover, the feature vector employed for classical metrics involves a fusion of degree assortativity, diameter, average closeness, average eccentricity, average betweenness, average degree, and average clustering, each explained comprehensively in Chapter 2.

6.4 Results and Discussion

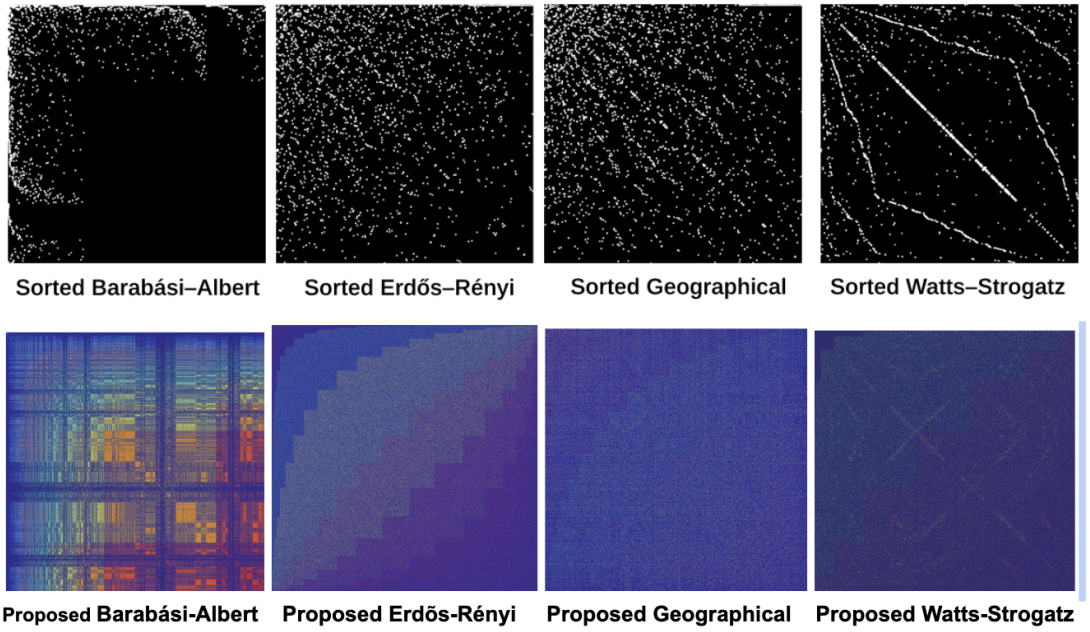
The experiment computed the C sorted matrix of all networks of dataset. It is important to highlight the although the Jaccard matrix does not contains the values of edges, C is sorted based on the centrality information of the vertices like in [Neiva and Bruno 2023]. C is arranged them in such a way that the network hubs appear at the start of the graph. This ordering guarantees that the final adjusted matrix, C', will not be influenced by the original arrangement of vertices in the initial matrix, C.

A previous experiments have indeed tested the combination of edge and Jaccard information, but results did not provide a good output in complex network classification.

6.4.1 Visual Analysis of the Colored Ordered Patterns

First, the main method of examination used in this study is a visual inspection of the C' matrix. Here, it is important to expose that all C' figures were colored with jet colormap to highlight for visualization purpose. The jet colormap applies cool colors to lowest values while warmest colors refers to highest values. Figure 24 illustrates the impact of this methodology on a synthetic dataset. From the images, one can notice that the patterns change completly for each class. However, within the approach, the pixels distribution is different for each class, allowing the use of several computer-vision techniques for classification. Figures 25, 26, 27, and 28 show some samples from each synthetic class. From visual inspection, we can observe some interesting pattern explored bellow.

Figure 24 – The first row shows the sorted A matrices from [Neiva and Bruno 2023], while the second shows the result of the novel proposed approach.

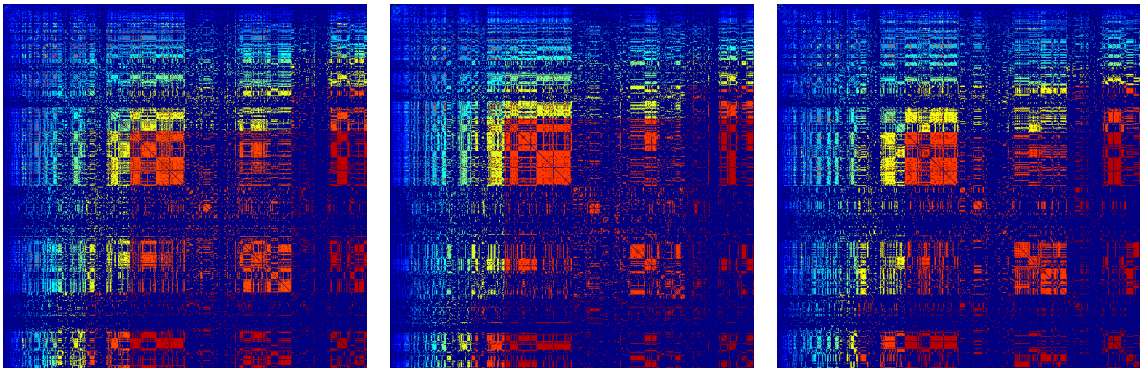


Source: Developed by the author

The main feature of Barabási-Albert networks is the presence of hubs, few number of nodes with high degree. The scalefree networks proposed by Barabási–Albert create graphs by starting with an arbitrary network of N_0 nodes. Then, new nodes are added and linked to c existing nodes, where $c \leq N_0$. Furthermore, the probability of connecting to a specific node is proportional to its degree, meaning a vertex with a higher degree has a greater chance of being connected to the new node.

Figure 25 shows examples of Barabási–Albert networks after the application of the proposed method. First, it is important to notice that none of the proposed models shows the edges itself of the models, creating an layer in the interpretation of the images. However, we know that the from the sorting method, hubs are located in the first columns and rows of the image. As presented, it is possible to notice that although the high centrality of the hubs, the

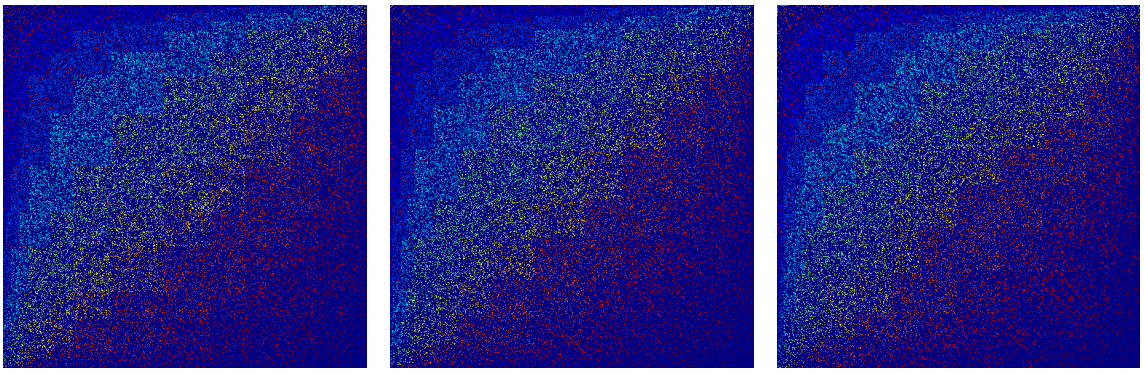
Figure 25 – After the application of the proposed method, the patterns of Barabási–Albert model are highlighted. All networks contain 500 vertices.



Source: Developed by the author

similarity with other nodes are low. This is easy to notice by the construction of the model, only few nodes contains these hub characteristic. Furthermore, the presence of blocks is an important visual feature of the pattern, with lower degree nodes with highest similarity with other nodes.

Figure 26 – The Erdős–Rényi model images show a wide distribution of points along all the images strengthening the characteristics of the randomness of the network. All networks contain 500 vertices.



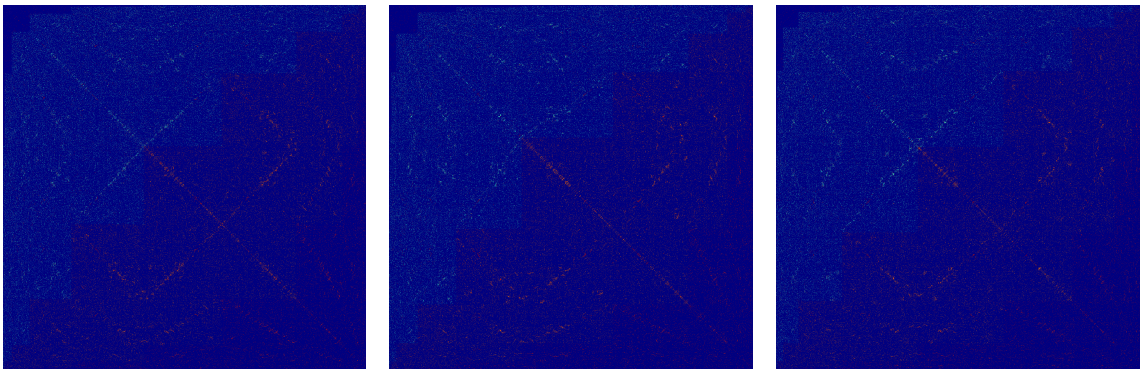
Source: Developed by the author

The Erdős–Rényi network model operates by connecting N nodes according to a probability p, without exhibiting any preference for linking [Erdos and Rényi 1960]. The parameter p plays a significant role in the structure of the resulting graph: a lower p generates a weakly connected graph, while a higher p produces a densely interconnected graph. This non-preferential attachment process gives rise to well-known random networks [Erdos and Rényi 1960]. However, distinct patterns can be visually discerned within this model, as shown in Figure 26.

The image's upper left corner features a mix of nodes, characterized by high centrality. With a varied of similarity values to other nodes in the matrix, Erdős–Rényi images presents sections of values according to the nunber of nodes. The nodes with the lowest degree tend to demonstrate a lower similarity with other nodes within the network. Interestingly, the patterns

produced through this method differ from those seen in the binary image from NEIVA; BRUNO study, as depicted in Figure 24. This contrast highlights the unique insights offered by our approach, showcasing its potential for unveiling novel patterns in network structures.

Figure 27 – It is interesting to notice how patterns of hubs are represented in Watts-Strogatz model images. Even with the increase in the number of nodes, the basic pattern remains. All networks contain 500 vertices.



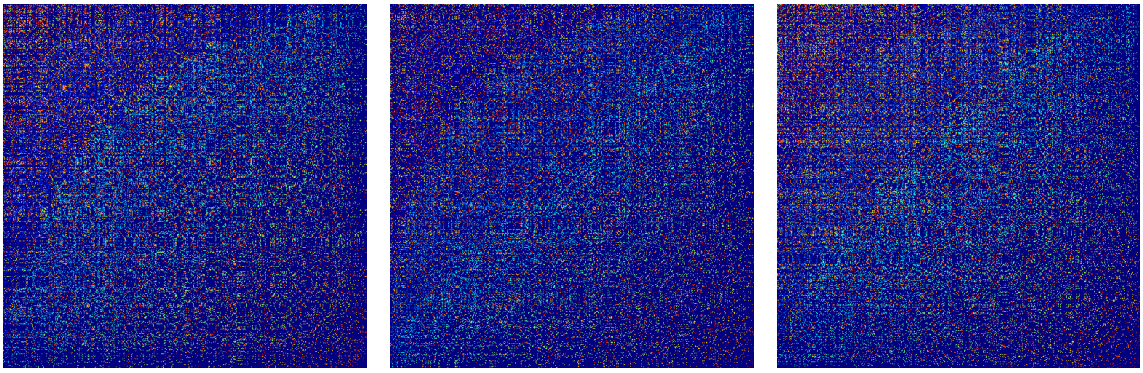
Source: Developed by the author

The third CN model is constructed over a k-regular graph with N nodes. Subsequently, each set of $k/2$ neighbors of each vertex is rewired with a probability $0 \leq \beta \leq 1$, where β is the model's parameter [Watts and Strogatz 1998]. This model's defining features include a high clustering coefficient of nodes, a low diameter, and short paths connecting any two nodes in the network, setting it apart from other models.

Examining the output in Figure 27, which is generated with $\beta = 0.1$, we can identify two main sections in the image. Nodes with the highest centrality exhibit low similarity between nodes, while nodes with lower centrality display higher similarity values. This pattern suggests that the model's low diameter and high node connectivity might reduce the variability of Jaccard indexes in the image. The discernible patterns underscore the model's distinct network structure, characterized by robust interconnectivity and short path lengths.

Lastly, the Geographical model (in Figure 28) demonstrates a pattern similar to the sections seen in Figure 26. However, in this case, there's a wider range of $j_{i,j}$ values within each section. Consequently, while the image highlights a discernible pattern, there's a noticeable increase in color pattern variability as centrality decreases. Geographical networks are constructed based on the system's spatial properties. Although the construction rules are not strictly defined, the patterns within this model are consistent with those identified in [Neiva and Bruno 2023]. The variance in $j_{i,j}$ values and centrality further emphasizes the richness and complexity of the network structure within the Geographical model.

Figure 28 – The Geographical model. In the images, one could check that the patterns within the model are maintained. All networks contain 500 vertices.



Source: Developed by the author

6.4.2 Quantitative Analysis of the Proposal for Complex Networks Classification

The second experiment transitions from visual inspection to a rigorous quantitative assessment of the proposed methodology. The promising results observed in Section 6.4.1 for texture classification within synthetic models served as the motivation for further exploration. Consequently, the evaluation proceeds with the application of several feature extraction techniques, including CLBP, Hu moments, and projection, as well as the transfer learning strategy derived from VGG-19.

The accuracy of these methods is evaluated using SVM alongside a 10-fold cross-validation process. These results are then compared with those of LLNA-based techniques [Miranda, Machicao and Bruno 2016, Zielinski *et al.* 2022, Ribas, Machicao and Bruno 2020], as well as classical structural metrics and previous methodologies as explored by NEIVA; BRUNO (2023).

The proposed methodology undergoes evaluation across both synthetic and real-world datasets, thereby facilitating a comprehensive exploration of its potential. The results of this comprehensive evaluation are detailed in Table 3. The evidence thus far suggests a promising avenue for further research, reinforcing the need for more nuanced and rigorous investigations into the power and potential of this methodology.

Firstly, concerning synthetic models, these are known for their strong characteristics linked to degree distribution and other traditional metrics. This is observed in the output of structural metrics from synthetic datasets, which show an accuracy rate of 100% for synthetic, scalefree, and noise datasets. In these cases, the previous methodology [Neiva and Bruno 2023] and D-TEP also yield high-quality results. It is crucial to note that to ensure fairness, the results showcased for each comparative method represent their best performance. A distinct advantage of our approach is the lack of a requirement for a parametrization step, unlike what is seen

Table 3 – Classification results comparing the proposed method with literature methodologies.

	PROPOSED				LITERATURE				
	VGG-19	PROJECTION	CLBP	HU MOMENTS	SORTED ADJ	LLNA	BP-LLNA	D-TEP	STRUCTURAL METRICS
Synthetic	**100.00% ± 0.00**	**100.00% ± 0.00**	**100.00% ± 0.00**	98.76% ± 1.21	99.86% ± 0.12	99.9% ± 0.00	**100.00% ± 0.00**	**100.00% ± 0.00**	**100.00% ± 0.00**
Scalefree	**100% ± 0.00**	99.60% ± 1.26	97.80% ± 1.75)	96.40% ± 2.07	**100.00% ±0.00**	98.3% ± 0.07	99.52% ± 0.19	**100.0% ± 0.00**	**100.00% ± 0.00**
Noise 10%	**100% ± 0.00**	99.63% ± 0.60	99.75% ± 0.78	99.26% 0.64	**100.00% ± 0.00**	99.98% ± 0.00	**100.0% ± 0.00**	**100.0% ± 0.00**	**100.0% ± 0.00**
Noise 20%	**100% ± 0.00**	99.75% ± 0.53	99.50% ± 0.87	94.38% ± 2.65	**100.00% ± 0.00**	99.97% ± 0.01	99.98% ± 0.00	**100.0% ± 0.00**	**100.0% ± 0.00**
Noise 30%	99.9% ± 0.52	99.88% ± 0.39	99.01% ± 0.78	87.28% ± 0.60	99.9% ± 0.76	99.95% ± 0.01	99.99% ± 0.00	99.75% ± 0.00	**100.0% ± 0.00**
Social Network	**93.5% ± 2.03**	91.99% ± 6.64	79.10 % ± 9.56	52.69 % ± 11.76	91.47% ± 9.21	92.00% ± 1.00	**93.40% ± 0.92**	92.5% ± 0.50	92.31% ± 5.13
Kingdom	**99.9% ± 2.17**	92.50% ± 5.74	96.88% ± 4.42	83.75% ± 5.27	**99.6% ± 3.47**	93.10% ± 5.38	97.44% ± 3.98	96.24% ± 0.35	96.61% ± 4.33
Animal	86.9% ± 4.84	91.33% ± 9.19	**93.00% ± 12.01**	62.33% ± 10.89	93.20% ± 4.22	77.25% ± 16.29	84.87% ± 15.25	**100.0% ± 0.00**	83.71% ±15.29
Fungi	78.0% ± 3.85	70.00% ± 20.49	71.67% ± 19.33	53.33% ± 25.82	**81.95% ± 2.38**	54.58% ± 19.38	76.17% ± 17.45	81.00% ± 4.38	54.90% ± 15.39
Plant	**81.9% ± 2.42**	80.00% ± 21.94	81.67% ± 19.95	73.33% ± 19.56	79.2% ± 4.13	69.70% ± 4.67	74.81% ±5.64	79.58% ± 2.12	54.19% ± 9.17
Firmicutes-Baccilis	**98.8% ± 0.57**	91.10% ± 2.71	87.01% ± 5.02	78.55% ± 4.88	**99.9% ± 0.27**	84.63% ± 2.00	98.30% ± 1.17	95.73% ± 0.34	95.67% ± 0.59
Actinobacteria	**99.5% ± 0.36**	95.47% ± 3.69	94.97% ± 5.77	84.34% ± 9.83	**99.5% ± 0.87**	91.48% ± 1.60	95.13% ± 1.22	97.65% ± 0.29	93.16% ± 0.70

Source: Developed by the author

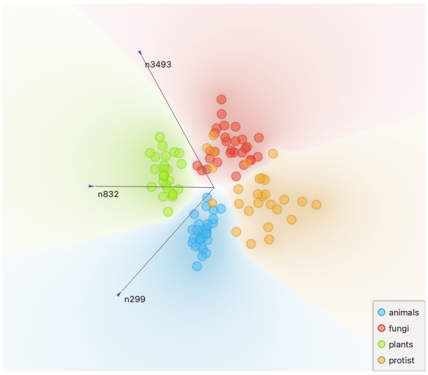
in LLNA-based models [Miranda, Machicao and Bruno 2016, Ribas, Machicao and Bruno 2020, Zielinski *et al.* 2022]. Furthermore, despite their simplicity, the projection technique and other image feature methodologies also produce results greater than 96% for synthetic and scalefree datasets.

The importance of the sorting approach cannot be overstated. Without the sorting step of the proposal, the Jaccard matrix rows and columns could be permuted while still maintaining node similarity information. This could result in a disorganized color mix in the CN-image, obscuring any evident pattern for texture feature extraction. The use of images also allows us to apply deep neural networks to identify the most significant features of the CN images.

For social networks, the C' matrix can accurately classify social network patterns with 93.5% accuracy using VGG-19 for feature extraction. This outcome is comparable to that of BP-LLNA, but with the added advantage of not requiring a parametrization process.

In the case of metabolic datasets, the new proposal's Kingdom and Plant networks classification results, using VGG-19 for feature extraction, surpass all previous accuracies and the application of CLBP with these datasets also plays a vital role in label distinction.

Figure 29 – Kingdom dataset - Principal components plot of the features



Source: Developed by the author

Regarding Kingdom dataset, Table 4 and Figure 29 summarize the performance of an SVM classifier that utilizes the Jaccard Sorting matrix and VGG-19 features for classifying biological organisms of different kingdoms. The matrix reveals high accuracy in predicting

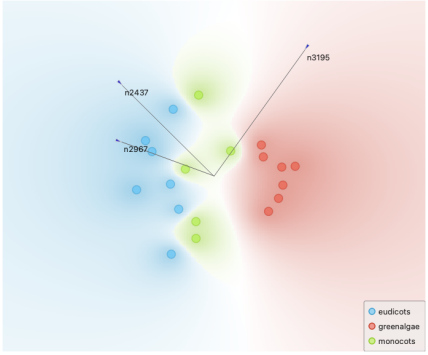
Table 4 – Kingdom dataset confusion matrix regarding VGG-19 results

KINGDOM	animals	fungi	plants	protist
animals	97.60%			
fungi		88.90%		
plants			100.00%	
protist	2.40%	11.10%		100.00%

Source: Developed by the author

animals, fungi, and plants, with accuracies of 97.60%, 88.90%, and 100.00% respectively. However, there are misclassifications between protists and other kingdoms, with 2.40% of protists being misclassified as animals and 11.10% as fungi. In terms of characteristics, fungi and protist shares some features such as chemoorganotrophic metabolism, meaning they obtain energy by breaking down organic compounds through various metabolic pathways and can obtain nutrients from the enviromnent, properties that can justify the errors. The misclassification can also be identified in principal component analysis figure on the right side of Figure 29 where Fungi and Protist classes are mixed in the space.

Figure 30 – Plant dataset - Principal components plot of the features



Source: Developed by the author

Table 5 – Plant dataset confusion matrix regarding VGG-19 results

PLANT	eudicots	greenalgae	monocots
eudicots	83.30%	10.00%	27.30%
greenalgae		90.00%	
monocots	16.70%		72.70%

Source: Developed by the author

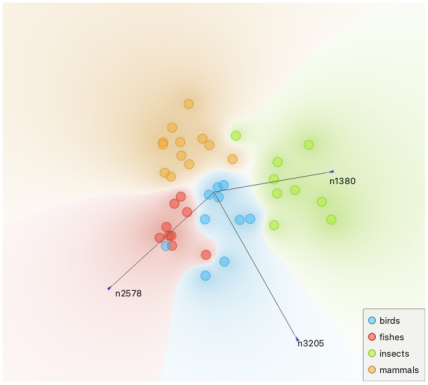
For classifying Plant species into three categories: eudicots, green algae, and monocots, SVM achieved high accuracy of 83.30% for eudicots and 90.00% for green algae. However, as shown in Table 5 and Figure 30, there were misclassifications observed, with some eudicots samples being wrongly predicted as green algae (10.00%) and monocots (27.30%). The classifier obtained an accuracy of 72.70% for monocots, but there were misclassifications as well, with 16.70% of the monocots samples being incorrectly predicted as eudicots. It is interesting to notice that the high misclassification of monocots label is also perceived in the subsampled data

plot after principal component analysis (PCA). The class, represented in the green region is located in the middle of both other classes which can deliver a uncorrected labeling process.

Regarding the Animal metabolic set, the performance of different classification methods was evaluated. The D-TEP method achieved perfect accuracy of 100% in classifying the classes. On the other hand, both the proposed CLBP method and the Sorted Adjacency technique achieved over 90% accuracy, demonstrating the potential of the proposed approach. The confusion matrix, depicted in Table 6, provides a breakdown of the classification results for each label.

The classifier demonstrated high accuracy for fishes (88.90%) and insects (81.80%). However, notable misclassifications were observed, particularly between birds and mammals. The confusion between these two classes may be attributed to their similar distributions in the feature space, as illustrated by the PCA plot in Figure 31. The sub-sampled data exhibited the distribution of the three main features in the space. It is evident from the graph that the region corresponding to the bird label falls in the middle of the other classes, particularly fishes and mammals. This proximity among the classes justifies the higher frequency of misclassifications between birds and mammals.

Figure 31 – Animal dataset - Principal components plot of the features



Source: Developed by the author

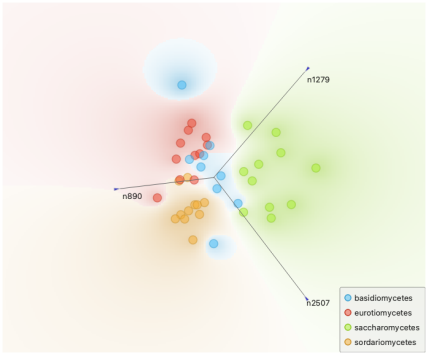
Table 6 – Animal dataset confusion matrix regarding VGG-19 results

ANIMAL	birds	fishes	insects	mammals
birds	46.20%		18.20%	26.10%
fishes	23.10%	88.90%		13.00%
insects	7.7%	11.10%	81.80%	13.00%
mammals	23.10%			47.80%

Source: Developed by the author

Overall, the results highlight the effectiveness of the proposed CLBP method and the Sorted Adjacency technique in achieving high accuracy in classifying the Animal metabolic set. The misclassifications observed between birds and mammals can be attributed to the overlap in their feature distributions regarding metabolic characteristics. These findings emphasize the importance of feature selection and the potential for further refinement in accurately distinguishing between these closely related animal classes.

Figure 32 – Fungi dataset - Principal components plot of the features



Source: Developed by the author

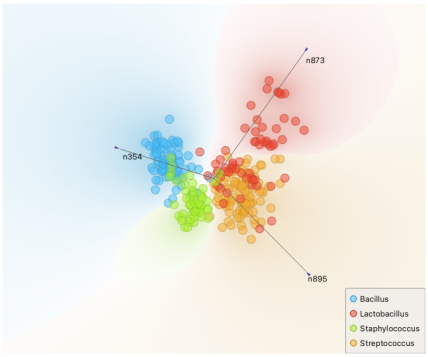
Table 7 – Fungi dataset confusion matrix regarding VGG-19 results

FUNGI	basidiomycetes	eurotiomycetes	saccharomycetes	sordariomycetes
basidiomycetes	31.20%		7.70%	42.90%
eurotiomycetes	25.00%	90.00%		9.50%
saccharomycetes	18.80%		92.30%	
sordariomycetes	25.00%	10.00%		47.60%

Source: Developed by the author

For Fungi, the best results are also obtained with the Sorted Adjacency proposal and D-TEP, with our approach slightly trailing by about 3% in accuracy with VGG-19 feature extraction. The confusion matrix in Table 7 shows that The classifier achieved high accuracy for eurotiomycetes (90.00%) and saccharomycetes (92.30%). However, there were notable misclassifications observed, particularly between basidiomycetes and sordariomycetes. The accuracy for basidiomycetes was 31.20%, while the accuracy for sordariomycetes was 47.60%. The misclassifications between basidiomycetes and other categories, as well as sordariomycetes and other categories, were relatively high which can also be perceived in the PCA plot in Figure 32. Basidiomycetes, including mushrooms, toadstools, puffballs, and bracket fungi, is the major group which can lead to misclassification due to its spatial distribution of features as seen in Figure 32.

Figure 33 – Firmicutes-bacillis dataset - Principal components plot of the features



Source: Developed by the author

The bacteria set Firmicutes-bacillis achieved 98.8% of accuracy with our approach, a

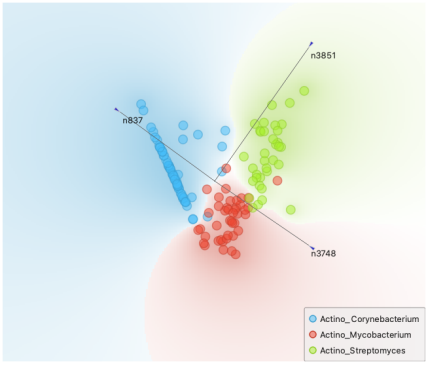
Table 8 – Firmicutes-bacillis dataset confusion matrix regarding VGG-19 results

FIRMICUTES-BACILLIS	Bacillus	Lactobacillus	Stephylococcus	Streptococcus
Bacillus	92.80%	3.40%	4.30%	
Lactobacillus	1.60%	89.80%		18.10%
Stephylococcus	5.60%	1.70%	96.70%	0.60%
Streptococcus		5.10%		81.20%

Source: Developed by the author

great classification rate. However, only our previous approach Sorted Adj obtained a better classification. BP-LLNA also provides a competitive result for this set. However, as mentioned before, our proposed method require zero parametrization process in comparison to LLNA-based methods. Furthermore, the confusion matrix provided in Table 8 summarizes the performance of the proposal for each one of the four labels in the dataset. The classifier achieved high accuracy for Bacillus (92.80%), Lactobacillus (96.70%), and Streptococcus (81.20%). However, there were misclassifications observed, particularly between Lactobacillus and Streptococcus. The information can be noticed by the mixing patterns presented in PCA plot (Figure 33). It is interesting to notice that Lactobacillus and Streptococcus share a spherical shape known as cocci. However, Bacillus has a rod-shaped structure known as bacilli, while Staphylococcus forms clusters of cocci.

Figure 34 – Actinobacteria dataset - Principal components plot of the features



Source: Developed by the author

Table 9 – Actinobacteria dataset confusion matrix regarding VGG-19 results

ACTINOBACTERIA	Corynebacterium	Mycobacterium	Streptomyces
Corynebacterium	92.20%	5.40%	
Mycobacterium	5.60%	94.60%	3.80%
Streptomyces	2.20%		96.20%

Source: Developed by the author

Finally, Actinobacteria matches the best results of all compared techniques, the sorted adjacency matrix. The 99.5% result with low standard deviation surpasses all LLNA-based models and structural measures. In summary, all results collectively highlight the effectiveness of our proposed approach in enhancing global network classification across a diverse range of datasets. The confusion matrix presented The classifier demonstrated excellent accuracy rates of 92.20% for Corynebacterium, 96.20% for Streptomyces, and 94.60% for Mycobacterium

as seen in Table 9. However, there were instances of misclassifications observed, specifically between Corynebacterium also noticed by PCA plot (Figure 34). However, the percentage of misclassifcation is low compared to the corrected prediction of the approach.

6.5 Conclusion

In conclusion, this study explores into the application of an image generation method that combines vertex similarity, computed using the Jaccard index, with the sorting ordination of matrices as previously proposed by NEIVA; BRUNO (2023). The results, as expounded in Section 6.4.1, demonstrate the efficacy of applying the jet colormap to these images, illuminating intriguing patterns and structures.

Moreover, the utilization of the sorting technique add confidence in the consistency and stability of these patterns, regardless of the input format. This consistency enables the creation of a unified and coherent representation suitable for image-based analysis. The evaluation conducted on twelve diverse datasets, encompassing seven real-world systems, underscores the potential of exploring novel approaches that further harness CN-image-based analysis techniques.

As avenues for future exploration, it is suggested to consider the incorporation of other network metrics, particularly measures of vertex relationships, to enhance the CN-image. The integration of these metrics has the potential to reveal new, intricate patterns, leading to more nuanced understandings of the underlying network structures. Additionally, a critical challenge lies in minimizing the standard deviation in classification for real datasets, such as those pertaining to Fungi and Animals. Addressing this challenge could significantly improve classification accuracy and bolster the methodology's robustness across diverse datasets. Therefore, the scope for refinement and advancement of this approach remains wide-ranging, signaling a promising future for CN-image-based network analysis.

CHAPTER

7

CLASSIFYING COMPLEX NETWORKS USING MULTISCALE ANALYSIS

The following chapter has a paper-like format. The chapter explores the use of multiscale analysis over in the attempt of enhancing graph features for complex networks classification.

7.1 Introduction

Despite the lack of adequate tools for analyzing dynamic objects, classical science has long confined natural phenomena to specific locations in the universe. However, with the advent of advanced technological tools, extensive data about specific structures can now be collected and stored.

The main goal in the field of data science is to explore, classify, and understand patterns. Complex networks (CNs) paired with an integrative methodology offer a fascinating perspective for this effort. CNs have become a critical tool, particularly in enabling computers to identify patterns and allowing analysis from shifting from a reductionist viewpoint to a more integrative mode. By taking advantage of the connections present in the data, CNs assist data scientists in discovering relationships and dependencies that could be missed when using conventional statistical methods.

Moreover, classification, a technique of pattern recognition, is particularly valuable in identifying various entities, such as social systems, diseases, images, and molecular structures [Costa *et al.* 2007]. Automating the analysis of such entities saves time and resources, while revealing the fundamental structure and organization of diverse systems, enhancing our understanding of their functionality and behavior.

CNs have been utilized in pattern recognition methodologies for a long time, employing graph metrics such as degree, betweenness, closeness, eigenvector centrality, and PageRank

[Costa *et al.* 2007]. However, alternative methodologies have emerged, including the use of cellular automata, multi-dimensional and deep embeddings, examination of angles formed within CNs, and randomized neural networks for feature extraction [Miranda, Machicao and Bruno 2016, Ribas, Machicao and Bruno 2020, Scabini *et al.* 2017, Zielinski *et al.* 2022, Scabini *et al.* 2022]. These innovations have expanded the scope of complex network analysis.

Despite these advancements, there is still room for further research in CNs, especially regarding improving classification accuracy through novel methodologies. Inspired by computer vision techniques, this study presents a method that employs k-core decomposition to generate different levels of detail within a graph. Conventional centrality metrics and exploitation of the adjacency matrix proposed in [Neiva and Bruno 2023] are then used to form a feature vector. The paper also addresses the challenge of varying CN sizes by devising a strategy to retain and maximize information content across all datasets. The findings demonstrate a significant increase in classification accuracy for both synthetic and real datasets, highlighting the potential advantages of this innovative approach in complex network analysis.

7.2 Proposed Approach

Complex networks, also known as graphs, consist of nodes and edges that exhibit non-trivial topological characteristics. These networks can be distinguished by attributes such as degree distribution, clustering coefficients, and community structures [Newman 2011]. For example, the classical synthetic Barabasi-Albert model [Barabási and Albert 1999] is constructed with a few nodes, known as hubs, having a high number of edges. This characteristic is evident in the power-law pattern observed in the degree distribution. Prominent examples of scale-free networks, following the Barabasi-Albert model, include the Internet [Faloutsos, Faloutsos and Faloutsos 1999], the World Wide Web [Barabási and Albert 1999], citation networks [Redner 1998], and certain social networks like Twitter [Huberman, Romero and Wu 2008].

To accurately describe the components of a graph, scientists commonly represent it as G = {V,E}, where V denotes the nodes (or vertices) as V = $\{v_1, v_2, v_i, v_j, v_N\}$, and E represents the links as E = $\{e_{ij} = (v_i, v_j) |$ if v_i and v_j are connected$\}$. Researchers extract features from the graph structure to create a summarized representation for various tasks, including classification, clustering, and others. In the context of classification, the entire structure is typically examined. However, decomposition techniques can be employed for analysis purposes or to enhance graph visualization. Analyzing motifs and community structures assists in achieving these goals and understanding the dynamics of a CN [Newman 2006, Girvan and Newman 2002].

One widely explored technique for visualization is the **k-core** decomposition [Alvarez-Hamelin *et al.* 2005]. The k-core refers to a maximal subgraph in which all nodes have a degree of at least k, meaning that each node is connected to a minimum of k other nodes within the subgraph. The process of extracting k-cores involves iteratively removing nodes with a

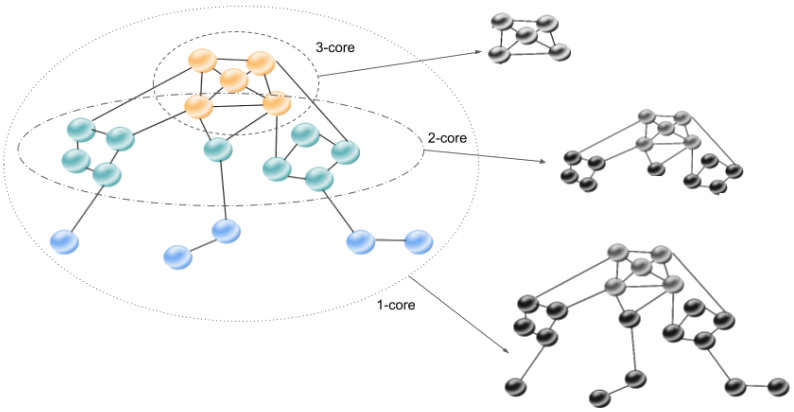
degree less than k until all remaining nodes have a degree of at least k [Seidman 1983]. This method allows researchers to identify cohesive substructures within the network and analyze their properties, such as connectivity, influence, hierarchical structure, and identification of central and peripheral nodes [Alvarez-Hamelin *et al.* 2005, Dorogovtsev, Goltsev and Mendes 2006]. Moreover, k-core decomposition has found extensive applications in various domains, including social network analysis, biological network analysis, and the study of the World Wide Web [Carmi *et al.* 2007].

To understand the method, Figure 35 demonstrates the three k-cores ($k = \{1,2,3\}$) of a small network for enhanced visualization. In a concise manner, the algorithm functions as described bellow:

1. Remove a vertex with a degree less than k (and their respective edges)

2. Repeat step 1 until all vertices to be removed is found.

Figure 35 – Example of the k-core decomposition of a graph.



Source: Developed by the author

The analysis of each k-core leads to the formation of *layers*, ranging from the most central (with high k values relative to the network) to the outermost ones. As the value of k increases, highly interconnected structures become more prominent. For instance, in social networks, nodes with high degrees may represent influential individuals, while in a collaborative network, a densely connected group of researchers can be observed [Dorogovtsev, Goltsev and Mendes 2006].

In conjunction with k-core decomposition, various techniques for feature description can be applied to extract essential characteristics of each layer of the graph. These extracted features can then be utilized as input for classification tasks, as elaborated in the subsequent sections.

7.3 Experiments

7.3.1 Parameter Selection and Decomposition

In addition, when using the k-core decomposition technique, it is essential to select an appropriate value for the parameter k. The number of maximum subgraphs obtained through k-core decomposition can vary depending on the specific graph and dataset. Instead of choosing a random value for k, we use a informed approach. A subset of the training set is utilized to estimate the average degree of the sampled data. This average degree is then used as the maximum value of k to decompose the original graph into subgraphs $G' = \{G_1, G_2, \ldots, G_{k^G_{\max}}\}$.

The decomposition into subgraphs enables the examination of each core as an individual graph. The final feature vector is constructed by concatenating the features from each subgraph. This approach introduces a novel perspective for graph analysis, allowing for a focused examination of patterns that might be otherwise overlooked when analyzing the entire system. Each core represents groups or communities of vertices with varying degrees, enabling the final feature vector to encompass a comprehensive set of attributes derived from multiscale analysis using the k-core technique.

7.3.2 Feature Extraction Methods

Concerning feature extraction, two methodologies are employed:

- **Traditional features extraction:** from each subgraph, classical features described in Chapter 2 are extracted. The features of each core are combined to form the input for the classification model. Specially for synthetic networks, these methodologies have been demonstrated effectiveness to distinguish CN patterns.

- **Sorted adjacency approach:** proposed in [Neiva and Bruno 2023], the method utilizes the sorted adjacency matrix proposed in Chapter 5 of each subgraph as input for image classification features extraction methods: Complete Local Binary Pattern (CLBP), projection and VGG-19 transfer learning technique explained in 4. Given that a permutation of the adjacency matrix's rows and columns represents an identical graph, the method initially sorts the rows and columns based on a vertex criterion to properly capture the class patterns.

The final feature vector for each G is the concatenation of the feature vector of each subgraph:

$$v_G = \{f(G_1), f(G_2), \ldots, f(G_{k^G_{\max}})\} \tag{7.1}$$

where f is the application one of the feature vector techniques evaluated in this study: VGG-19 transfer learning, CLBP, projection or classical features.

Furthermore, methods are evaluated in seven different datasets including. All datasets are detailed in Chapter 4. Regarding the comparative techniques of network classification, we compared the proposed methods with five tecnhiques. Two proposed in this thesis (Chapters 5 and 6), LLNA-based techniques [Miranda, Machicao and Bruno 2016, Ribas, Machicao and Bruno 2020, Zielinski *et al.* 2022] and classical measures from the whole dataset.

7.4 Results and Discussion

This section presents the results regarding the automatic parameter selection of each dataset and the classification of the proposed method in comparison to other results in the literature

7.4.1 Parameter Evaluation

Initially, as previously outlined, it is crucial to determine the maximum k-core value, k_{max}, to enable the decomposition of the initial dataset. The outcomes of this process are elucidated in Table 10. The importance of parameter selection lies in its capacity to counteract discrepancies in graph size, avoid parameter guessing, and accentuate the inherent characteristics of the dataset. Moreover, the parameter selection during decomposition facilitates the concatenation of features from each subgraph. This proves advantageous during the classification stage as all vectors maintain the same length, and each dimension corresponds to the same core level, preventing potential complications.

Table 10 – Results of parameter selection for each dataset evaluated in the proposed method.

Dataset	k_{Max}
Scalefree	7
Kingdom	4
Animal	4
Fungi	4
Plant	4
Firmicutes-Bacillis	4
Actinobacteria	5

Source: Developed by the author

From Table 10, metabolic networks demonstrate a mean k_{max} value of 4, a lower number of edges per vertex. This not only reveals the shortage of edges for each node but also underscores a characteristic feature inherent to such networks.

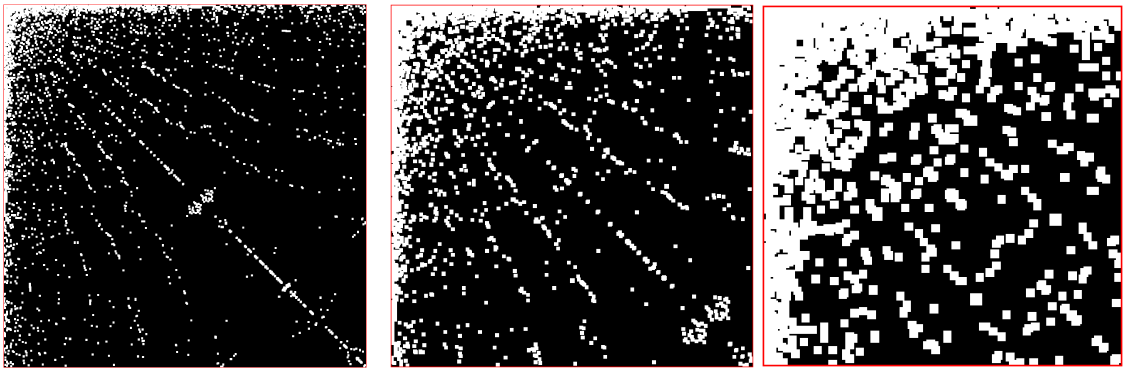
7.5 Visual Analysis of the Decomposition

The first analysis evaluates the impact of k-core decomposition in the patterns of the datasets. In this context, we aim to evaluate the differences regarding labels and the decomposition of a given graph. To highlight the patterns, all example images were dilated for the sake of visualization.

Figure 36 showcases the graph evolution from an example in Firmicutes-bacillis dataset. The subgraphs corresponding to k values of 2, 3, and 4 are presented, providing insights into the graph's structural changes at different levels. This application of the method enables a multiscale analysis, allowing for a closer examination of various regions within the graph.

As the k parameter increases, edges and vertices are progressively removed, leading to alterations in the ordering algorithm of the matrix by the changes in the centrality measures. This dynamic process contributes to a deeper understanding of the graph's characteristics and reveals how its connectivity evolves with different k values.

Figure 36 – Firmicutes-bacillis - decomposition of the same graph with k={2,3,4}



(a) k = 2 (b) k = 3 (c) k = 4

Source: Developed by the author

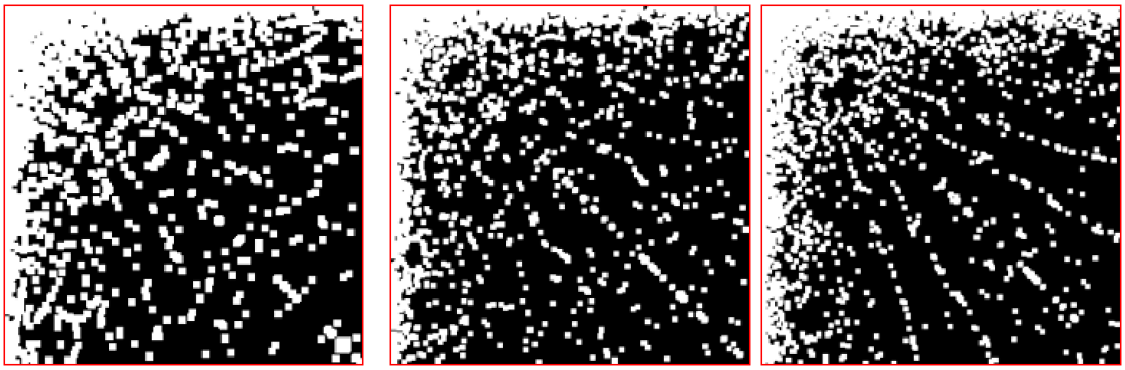
On the other hand, Figure 37 illustrates three image samples representing each label from the Actinobacteria dataset: Corynebacterium, Mycobacterium, and Streptomyces. All images depict the 4-core of the respective labels but exhibit distinct patterns of edge distribution.

Corynebacterium is characterized by a dense concentration of edges, indicating a high level of interconnectedness within the graph. In contrast, both Mycobacterium and Streptomyces exhibit the presence of 'lines' in the figure, particularly near the bottom right side of the images. Notably, the shape of Streptomyces closely resembles the small-world patterns observed in Watts-Strogatz graphs as documented in [Neiva and Bruno 2023], highlighting the presence of intricate connectivity patterns within the model.

By visually inspecting these image samples, one gains valuable insights into the varying edge distributions and structural characteristics of the different Actinobacteria labels. Such

observations provide a deeper understanding of the complex organization and connectivity patterns within the Actinobacteria dataset.

Figure 37 – Actinobacteria dataset sample. Example of 4-core three different labels: Corynebacterium, Mycobacterium and Streptomyces .



(a) Corynebacterium (b) Mycobacterium (c) Streptomyces

Source: Developed by the author

Regarding the Scalefree dataset, Figure 38 illustrates the 4-core of each label, showcasing the distinct patterns revealed by the proposed technique. Analyzing the Barabasi-Albert models, it is evident that increasing the α value leads to a concentration of edges, resulting in a limited number of highly connected nodes. This phenomenon arises from the preferential attachment mechanism, where well-connected nodes tend to attract connections.

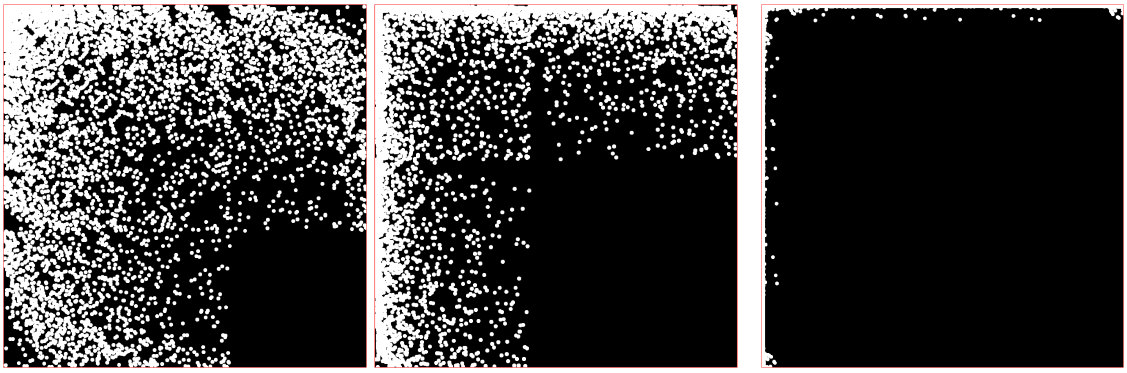
In contrast, the Dorogovtsev-Mendes model, while also producing scale-free networks, allows low-degree nodes to receive connections, albeit with a lower probability. This characteristic is exemplified by the white points within the black area in the bottom right region of the last label in the figure. This implies that even nodes with relatively low degrees have the potential to form connections and contribute to the network's structure.

7.6 Quantitative Results

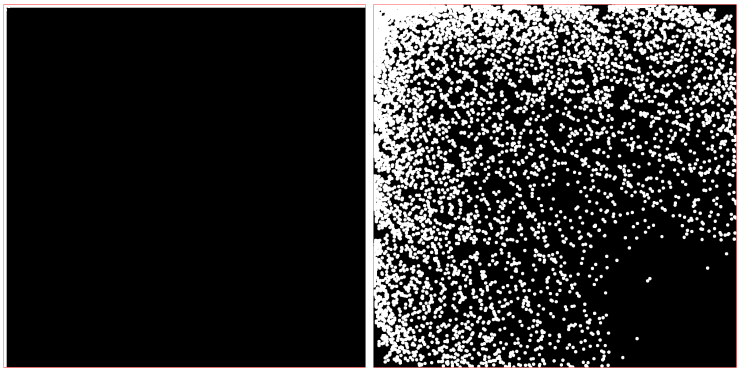
In addition, the result Table 11 and and 12 presents the results of the proposed methodologies and the results from the literature across real and synthetic datasets, respectively. In the context of classification accuracy, the best results are highlighted in bold for each domain. The values are presented in percentages, denoting the mean accuracy, and are accompanied by the standard deviation from SVM and cross validation classification.

Regarding the scalefree dataset, the proposed k-core aligned with the classical metrics, the proposed VGG-19, the Jaccard Matrix proposed from Chapter 6, the sorting method proposed in Chapter 5 and [Neiva and Bruno 2023], D-TEP from [Zielinski *et al.* 2022], and the concatenation of structural methods from the literature, showed perfect accuracy (100% $\pm$ 0.00). In contrast,

Figure 38 – Scalefree dataset example. Three 4-core subgraphs from the five different classes: Barabasi-Albert $\alpha = \{0.5, 1.0, 1.5, 2.0\}$, Mendes-Dorogovtsev-Mendes, respectively.



(a) Barabasi-Albert $\alpha = 0.5$ (b) Barabasi-Albert $\alpha = 1.0$ (c) Barabasi-Albert $\alpha = 1.5$



(d) Barabasi-Albert $\alpha = 2.0$ (e) Dorogovtsev-Mendes

Source: Developed by the author

Table 11 – Results from the proposed methods.

	VGG-19	PROJECTION	CLBP	K-CORE CLASSICAL
Scalefree	**100.00% ± 0.00**	88.20% ± 3.82	77.20% ± 4.13	**100.00% ± 0.00**
Kingdom	**98.8% ± 2.59**	95.63% ± 3.28	79.69% ± 4.42	96.9% ± 0.30
Animal	81.3% ± 3.5	**92.83% ± 8.06**	60.16% ± 12.50	87.7% ± 1.10
Fungi	78.5% ± 2.67	65.83% ± 9.38	46.25% ± 6.65	76.7% ± 1.01
Plant	**83.5% ± 2.28**	**86.18% ± 9.73**	62.91% ± 12.37	81.5% ± 1.00
Firmicutes-bacillis	98.6% ± 1.74	93.67% 2.07	76.63% ± 2.12	96.6% ± 0.07
Actinobacteria	97.7% ± 0.35	95.98% ± 1.65	83.42% ± 2.61	97.5% ± 0.19
Overall	90.2% ± 0.42	88.0% ± 0.35	69.45% ± 0.19	91.0% ± 0.21

Source: Developed by the author

projection approach and CLBP techniques fell behind, with the latter only achieving 77.20% ± 4.13.

In Kingdom, the proposed with VGG-19 transfer learning feature extraction achieved the highest accuracy (98.8% ± 2.59) among the proposed methods. However, it was marginally surpassed by the literature methods Jaccard and Sorted matrix-based methods, both scoring higher than 99.6%. For the Animal dataset, the best result regarding our proposal was output by

Table 12 – Results from the literature.

	JACCARD MATRIX	SORTED ADJ	LLNA	BP-LLNA	D-TEP	STRUCTURAL
Scalefree	**99.9% ± 0.00**	**99.9% ± 0.00**	96.7% ± 0.57	99.9% ± 0.00	99.9% ± 0.00	99.7% ± 0.57
Kingdom	**99.9% ± 2.17**	**99.6% ± 3.47**	93.10% ± 5.38	97.44% ± 3.98	96.24% ± 0.35	96.61% ± 4.33
Animal	86.9% ± 4.84	**93.20% ± 4.22**	77.25% ± 16.2	84.87% ± 15.25	**100.0% ± 0.00**	83.71% ± 15.29
Fungi	**78.0% ± 3.85**	81.95% ± 2.38	54.58% ± 19.38	76.17% ± 17.45	**81.00% ± 4.38**	54.90% ± 15.39
Plant	**81.9% ± 2.42**	79.2% ± 4.13	69.70% ± 4.67	74.81% ± 5.64	79.58% ± 2.12	54.19% ± 9.17
Firmicutes-bacillis	98.8% ± 0.57	**99.9% ± 0.27**	84.63% ± 2.00	98.30% ± 1.17	95.73% ± 0.34	95.67% ± 0.59
Actinobacteria	**99.5% ± 0.36**	**99.5% ± 0.87**	91.48% ± 1.60	95.13% ± 1.22	97.65% ± 0.29	93.16% ± 0.70
Overall	92.1% ± 0.42	93.0% ± 0.41	80.2% ± 0.58	89.4% ± 0.64	93.2% ± 0.36	83.9% ± 0.64

Source: Developed by the author

the projection method with accuracy of 92.83% (± 8.06). However, D-TEP from the literature methods overcomes all compared results with the accuracy of 100.00% ± 0.00.

In the Fungi dataset, none of the proposed methods outperformed the literature methods, with the highest accuracy achieved by the previous proposal, sorted adj (81.95% ± 2.42). However, for Plant, the subgraphs produced by k-core decomposition aligned with VGG-19 and projection were able to distinguish labels with 83.5% and 96.18%, respectively. It is important to notice that the decomposition also serves as data augmentation for classification, a common technique in deep neural networks [Shorten and Khoshgoftaar 2019]. In addition, for Firmicutes-bacillis, VGG-19 method from the proposed techniques showed high accuracy (98.6% ± 1.74), but it was slightly exceeded by the Sorted adjacency method from [Neiva and Bruno 2023] (99.9% ± 0.27). Finally, in the Actinobacteria domain, the proposed VGG-19 proposed technique, Jaccard Matrix and Sorted Adjacency all displayed similar high accuracy.

In general, the results demonstrate that the proposed methods can compete with, and in some cases exceed, the performance of established techniques from the literature, especially LLNA-based techniques [Ribas, Machicao and Bruno 2020, Miranda, Machicao and Bruno 2016] with lower parametrization process time and evaluation. However, the effectiveness of the methods varied across different domains, indicating that the selection of an optimal method may depend on the specific characteristics of the dataset under consideration.

7.7 Conclusion

This research evaluates various classification methodologies proposed in the study. The proposed approach involves decomposing the dataset into a set of subgraphs using k-core decomposition. This decomposition provides a multilayer perspective of the original dataset, allowing representation at different levels of detail by gradually removing vertices at each core.

The decomposed dataset is then subjected to feature extraction techniques, where either images (if sorted adjacency matrix is applied) or graphs (when classical metrics are utilized) are extracted. These extracted features are used for network class labeling. The performance of the proposed methodologies is compared against established techniques from the literature.

The evaluation is conducted on diverse real and synthetic datasets, each possessing unique characteristics. Classification accuracy serves as the performance measure, with cross-validation and SVM classification used to determine the mean and standard deviation of the results.

The research findings demonstrate the competitiveness of the proposed methodologies compared to existing methods in the literature. Notably, the proposed techniques show promising results when compared to LLNA-based techniques, offering advantages such as reduced parametrization time and improved accuracy. Moreover, it is important to note that the performance of the methodologies varies across different domains, emphasizing the need for careful parameter selection that aligns with the specific dataset characteristics. Future work could focus on addressing the limitation of dimensionality reduction and feature selection within the k-core decomposition approach. This would further enhance the effectiveness and applicability of the proposed methodologies.

CHAPTER

8

K-DTEP, A TOOL FOR CLASSIFYING K-CORE GENERATED NETWORKS

As with previous chapters, this text is structured similarly to a standalone academic paper within the thesis. Despite our best efforts to ensure conciseness, there may be certain passages that appear repetitive, which is a common occurrence given the context and conventions of the field. The chapter continues exploring the multiscale analysis for graph classification but combines the approach with previous proposal from ZIELINSKI *et al.* (2022)

8.1 Introduction

The capacity to build systems incorporating both elements and their interconnections facilitates a transition from viewing structures as mere points in space to a more integrative approach, capable of providing significant insights about the data analyzed. Traditional data analysis methods often reduce intricate phenomena to simplified forms. However, modern technological advancements permit a more comprehensive exploration of information.

Within this context, pattern recognition, through a technological lens, augments our understanding of the world. Complex networks, especially, emerge as a fascinating field for comprehensive analysis. They have been applied in diverse areas such as biology [Guye *et al.* 2010, Rain *et al.* 2001], social science [Scabini *et al.* 2021, Palla *et al.* 2005], and physical science [Carmi *et al.* 2009].

Furthermore, pattern recognition, combined with complex networks, plays a critical role in identifying entities like shapes [Backes and Bruno 2010], species [Ribas, Scabini and Bruno 2022], and authorship [Marinho, Hirst and Amancio 2016]. This underscores the importance of employing graphs to decipher the structures and patterns of any given organization.

Additionally, Cellular Automata (CA) have been used to investigate complex networks,

characterizing their intrinsic features for classification [Smith *et al.* 2011, Miranda, Machicao and Bruno 2016]. As a unique model category, CA excel at depicting complex systems and have become an essential tool for analyzing the complexity of spatio-temporal patterns [Wolfram 1983]. These patterns emerge from changes in a cell's state, determined by its neighboring conditions according to specified rules.

In this fusion, complex networks transform into a CA, creating a tessellation where each cell is symbolized by a node, and neighbors by adjacent cells. Subsequently, the time-evolution of patterns (TEP) is constructed by applying a set of rules to the states of a given cell and its neighbors' states. From this process, Miranda et al. [Miranda, Machicao and Bruno 2016] have demonstrated the substantial capability of TEPs to classify both synthetic and real models. The idea of applying Life-Like rules to evolve the automata was then expanded by Ribas et al. [Ribas, Machicao and Bruno 2020] in BP-LLNA, where the authors shifted from exploiting entropy metrics to analyzing binary patterns in BP-LLNA.

However, these prior methods focused on binary time evolution patterns, i.e., all cells representing values of zero or one. Hence, ZIELINSKI *et al.* (2022) proposed D-TEP, density time evolution patterns, combining previous LLNA-based techniques and density values computation from Miranda et al. [Miranda, Machicao and Bruno 2016], Machicao et al. [Machicao *et al.* 2018], and Ribas et al. [Ribas *et al.* 2018]. Beyond the promising results achieved by Zielinski et al. [Zielinski *et al.* 2022], this study introduces the combination of k-core decomposition of graphs into a set of subgraphs with varying levels of detail. The results surpass recent methodologies in literature for network classification in real datasets.

8.2 Proposed Method

As previously mentioned, Network Automata (NA) involves using a graph as a tessellation, where nodes correspond to cells. Formally, a Cellular Automata (CA) is defined as $C = (\tau, S, s_0, N, \phi)$, where:

- τ represents the tessellation, which is a collection of cells c_i.
- S is a finite set of states, and the notation $s(c_i, t) = 1$ denotes the state of cell c_i at time step t.
- N refers to the set of neighbors for each cell c_i.
- s_0 represents the initial condition of the CA, where a specific value is assigned to each cell at time $t = 0$.
- ϕ denotes the local transition function or rule, which controls the progression of the CA. This function determines how the state of each cell is modified, taking into account the states of its neighboring cells as well as its own current state.

The definition of Cellular Automata (CA) reveals two important aspects. Firstly, it is necessary to assign a set of states to each cell. Typically, cells are represented by binary values, where "0" signifies a "dead" cell and "1" represents an "alive" cell. Secondly, the local transition function ϕ governs the evolution of cell states in the CA. This function has the ability to generate diverse patterns and configurations as the dynamics unfold, potentially leading to vastly different outcomes compared to the initial state.

In the studies conducted by MIRANDA; MACHICAO; BRUNO (2016) and RIBAS; MACHICAO; BRUNO (2020), the evolution function in Network Automata is based on Life-Like rules. To ensure randomness and avoid bias, the methodology randomly assigns 50% of the cells as "alive" initially. The state of each cell at time t, denoted as $s(c_i,t)$, is then determined by the states of its neighbors at time $t-1$.

For example, following the rule B2/S3 or B3/S23, a cell transitions from state 0 to state 1 if it has three alive neighbors. Conversely, if a cell has 2 or 3 alive neighbors, it remains in state 1. However, it is important to keep in mind that differently from images, the number of neighbors (edges) in a network is not regular. Thus, the methodology defines the rules according to the density of neighbors alive or dead instead of considering the absolute value:

Then, the time-evolution pattern (TEP), sequential arrangement of the cell states in an automaton throughout the course of time, is retrieved. Considering [Miranda, Machicao and Bruno 2016] and [Ribas, Machicao and Bruno 2020], the TEP patterns are composed by a set of binary states. While [Miranda, Machicao and Bruno 2016] evaluates the patterns over the feature extraction based on Shannon entropy, the word length and the Lempel-Ziv complexity, [Ribas, Machicao and Bruno 2020] evaluates the local and global binary distribution of patterns.

However, [Zielinski *et al.* 2022] proposed a different approach for TEP construction. The D-TEP were proposed in order to move from a binary approach to a density one, where each state contains a value instead of only 0's and 1's. In the later case, each c_i's state at time t, $s(c_i,t)$, contains $\rho(c_i,t)$, i.e., the density of alive neighbors of a cell c_i at a timestep t, a continuous value between 0 and 1:

In Equation 8.1, A represents the adjacency matrix of G and A(i,j) is equal to zero if nodes i and j are not connected or one if connected. Finally, D-TEP matrix is represented by Equation 8.2.

$$\rho(c_i,t) = \frac{1}{k_i}\sum_j s(c_i,t)A(i,j) \qquad (8.1) \qquad \mathscr{D} = \begin{bmatrix} \rho(c_1,0) & \cdots & \rho(c_N,0) \\ \rho(c_1,1) & \cdots & \rho(c_N,1) \\ \vdots & \vdots & \vdots \\ \rho(c_1,T) & \cdots & \rho(c_N,T) \end{bmatrix} \qquad (8.2)$$

The D-TEP features can then be combined with traditional binary features, resulting in

SD-TEP, the state density-time evolution pattern:

$$S = \mathscr{D} \odot (2\mathscr{T} - 1) \tag{8.3}$$

In Equation 8.3, $\mathscr{T}$ represents the TEP binary matrix proposed in [Miranda, Machicao and Bruno 2016].

Then, a set of features are computed from the automata evolution. Each one of the histogram computed uses varying numbers of bins, denoted as L. This involved initializing a dictionary with L intervals and, for each node during every iteration in the D-TEPs, recording the frequency of values falling within a specific interval. The histogram are defined below:

- **Global histogram**: Computes the occurrences of various density intervals ranging from zero to one. The histogram contains data on the frequency of values within a particular **density interval**. However, as the overall occurrence of values varies depending on the network size, we normalize the histogram to make it independent of the network size. The features is computed according to Equations 8.4, 8.5 and 8.6

$$G_D^L(l) = \sum_{i=1}^{N} \sum_{t=0}^{T} f(D(i,t),l) \tag{8.4}$$

, where

$$f(D(i,t),l) = \begin{cases} 1, & if \frac{l}{L} \leq D(i,t) \leq \frac{l+1}{L} \\ 0, & \text{otherwise.} \end{cases} \tag{8.5}$$

The final global histogram is the vector for all levels of l, where $0 \leq l \leq L-1$:

$$\vec{\theta}_{\mathscr{D}}^{G}(L) = \begin{bmatrix} G_{\mathscr{D}}^{L}(0) \\ G_{\mathscr{D}}^{L}(2) \\ \vdots \\ G_{\mathscr{D}}^{L}(L-1) \end{bmatrix} \tag{8.6}$$

- **Degree histogram**: Frequencies are calculated separately for each specific **degree** value, resulting in N-size histograms, where N is the number of different degrees in the network, the histogram is computed according to Equations 8.7 and 8.8.

$$K_D^L(l,k^G) = \sum_{i=1}^{N} \sum_{t=0}^{T} f(D(i,t),l) \cdot \begin{cases} 1, & if k^G = k_i^G \leq \frac{l+1}{L} \\ 0, & \text{otherwise.} \end{cases} \tag{8.7}$$

, where $f(D(i,t),l)$ refers to Equation 8.5.

$$\vec{\theta}^{K}_{\mathscr{D}}(L) = \mu\left(\begin{bmatrix} K^{L}_{\mathscr{D}}(l,k_1) \\ K^{L}_{\mathscr{D}}(l,k_2) \\ \vdots \\ K^{L}_{\mathscr{D}}(l,k_{\max}) \end{bmatrix}\right) \tag{8.8}$$

- **Temporal histogram**: The histogram represents the **time** evolution pattern encompasses information about the nodes in the network at each iteration. Similar to the degree histogram, we can partition the histogram calculation for each time step of the D-TEP. This approach provides insights into the patterns observed during each step of the automaton's evolution. Thus, we can define a temporal histogram according to Equations 8.9 and 8.10.

$$T^{L}_{D}(l,t) = \sum_{i=1}^{N} f(D(i,t),l) \tag{8.9}$$

$$\vec{\theta}^{\tau}_{\mathscr{D}}(L) = \mu\left(\begin{bmatrix} \tau^{L}_{\mathscr{D}}(l,1) \\ \tau^{L}_{\mathscr{D}}(l,2) \\ \vdots \\ \tau^{L}_{\mathscr{D}}(l,T) \end{bmatrix}\right) \tag{8.10}$$

, where $0 \leq l \leq L-1$ and $0 \leq t \leq T-1$, with T representing the number of time steps the automaton has undergone. Furthermore, the histogram is further normalized to ensure it remains invariant to changes in network size. Similarly to the degree histogram discussed earlier, we propose a consolidated histogram for the temporal property by averaging all histograms obtained at each time step. These three histograms are calculated for both D-TEP and SD-TEP. To maximize classification accuracy, we concatenate all histograms for each TEP, resulting in a feature size of $6B$ for each graph, where B is the number of bins in the histograms. Equations 8.11 and 8.12 refers to the aggregated features of T-DEP and SD-TEP, respectively.

$$\vec{\Upsilon}^{\mathscr{D}}_{L} = \left[\vec{\theta}^{G}_{\mathscr{D}}(L), \vec{\theta}^{K}_{\mathscr{D}}(L), \vec{\theta}^{\tau}_{\mathscr{D}}(L)\right] \tag{8.11}$$

$$\vec{\Upsilon}^{\mathscr{S}}_{L} = \left[\vec{\theta}^{G}_{\mathscr{S}}(L), \vec{\theta}^{K}_{\mathscr{S}}(L), \vec{\theta}^{\tau}_{\mathscr{S}}(L)\right] \tag{8.12}$$

The proposed in [Zielinski *et al.* 2022] involves combining histograms with varying bin sizes to achieve an optimal discriminatory representation. For each bin size, there exist distinct feature vectors that represent the network. In order to identify the best representation

for the classification task, we conducted an experiment to explore this parameter by combining histograms with different numbers of bins, as outlined below:

$$\begin{aligned} \vec{\Theta}^{\mathscr{D}}_{[L_1,L_2,\ldots,L_n]} &= \left[\vec{\Upsilon}^{\mathscr{D}}_{L_1}, \vec{\Upsilon}^{\mathscr{D}}_{L_2}, \ldots, \vec{\Upsilon}^{\mathscr{D}}_{L_n}\right] \rightarrow \quad \text{D-TEP} \\ \vec{\Theta}^{\mathscr{S}}_{[L_1,L_2,\ldots,L_n]} &= \left[\vec{\Upsilon}^{\mathscr{S}}_{L_1}, \vec{\Upsilon}^{\mathscr{S}}_{L_2}, \ldots, \vec{\Upsilon}^{\mathscr{S}}_{L_n}\right] \rightarrow \quad \text{SD-TEP} \end{aligned} \tag{8.13}$$

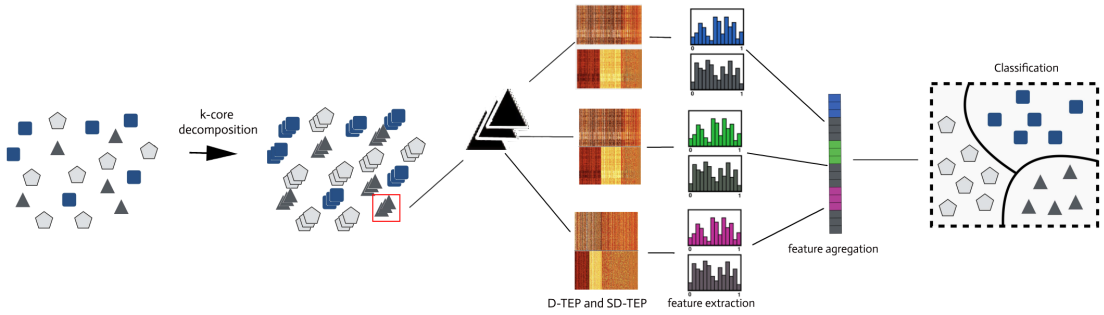
The, the graph signature involves concatenating the feature vectors of both D-TEP and SD-TEP, derived from the three histograms mentioned earlier. This concatenated signature combines the information from all three histogram representations to create a comprehensive feature vector:

$$\vec{\Omega}_{[L_1,L_2,\ldots,L_n]} = \left[\vec{\Theta}^{\mathscr{D}}_{L_1,L_2,\ldots,L_n}, \vec{\Theta}^{\mathscr{S}}_{L_1,L_2,\ldots,L_n}\right] \tag{8.14}$$

It is important to notice that Equation 8.15 represents the concatenated features of SD-TEP and D-TEP for one graph. Due to the multiscale approach, the final feature vector is the concatenation of the features of each subgraph from graph G:

$$\vec{\Omega}^{G} = \left[\vec{\Omega}^{G}_{k_{core}=1}, \vec{\Omega}^{G}_{k_{core}=\ldots}, \vec{\Omega}^{G}_{k_{core}=k_{max}}\right] \tag{8.15}$$

Figure 39 – Proposal Diagram. The final feature vector of a single graph G is composed by a combination of D-TEP and SD-TEP features from each subgraph of G.



Source: Developed by the author

Figure 39 illustrates the proposed method. First, the whole dataset is decomposed by the application of k-core decomposition. Then, each subgraph in the network dataset is assigned an initial state, represented by either zeros or ones (black and white vertices). By applying a specific rule, the automata evolve over a set number of time steps, and the densities of neighboring cells are observed at each iteration. D-TEPs and SD-TEPs are derived from this process. Subsequently, histograms are generated using various bin numbers (referred to as L). The ultimate feature vector of a single graph G is formed by merging the D-TEP and SD-TEP features obtained from each subgraph within G.

8.3 Experiments

For the experiments, we utilized six real-world network databases. These databases as described in 4 and were previously used in studies conducted by [Miranda, Machicao and Bruno 2016, Zielinski *et al.* 2022, Ribas, Machicao and Bruno 2020], allowing us to compare our results with theirs.

Furthermore, in order to ensure a fair comparison and improve the classification accuracy achieved in previous works, we utilized the same configuration parameters as the previous works. The LLNA was evolved for T = 350 timesteps, with the initial configuration s_0 randomly set using a normal distribution with 50% of cells alive. The same rules that were selected for evaluating the databases in the previous works were also used in this study.

Based on the LLNA proposal research by RIBAS; MACHICAO; BRUNO (2020), the following rules were employed for the Kingdom, Animal, Fungi, Plant, Protist, Firmicutes-Bacillis, and Actibacteria databases respectively: B02345678-S123468, B023468-S01468, B04-S1468, B0468-S0467, B0236-S123567, B0468-S0458, and B1237-S267.

Moreover, we used the k_{max}^{core} for each evaluated dataset presented in Chapter 7. Regarding the classification process, we employed a 10-fold cross-validation strategy. The algorithm tested all possible combinations of train-test data splits and calculated the average accuracy across all iterations. Since the split method is non-deterministic, we repeated this cross-validation procedure 100 times to ensure robustness.

For the classification task, we applied SVM (Support Vector Machines) algorithm. SVM employs decision boundaries and hyperplanes to find an optimal hyperplane that maximally separates two classes [Hearst *et al.* 1998].

8.4 Results

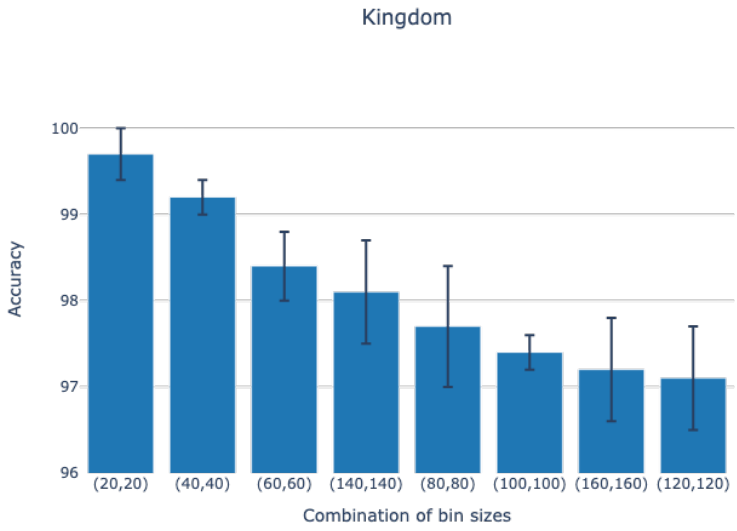
This section presents the results regarding our proposed approach.

8.4.1 Parameter evaluation

First, we evaluate the impact of the number of bins used to compute the histograms presented in Section 8.2. After analysing the best bin parameters, we can finally compare the approach with other literature. The set of bin values explored are L = {20, 40, 60, 80, 100, 120, 140, 160}. To exemplify the evaluation, we present Figure 40 that shows the accuracy of Kingdom dataset over the evaluated bin values. It is possible to notice that, for this case, obtain 99.7% $\pm$ 0.3 with $\Omega_{(20)}$. However, regarding the results in Table 13, the value of L = 80 also presents good results for the evaluated datasets.

Moreover, for Animal database exhibited the highest accuracy of 99.8% $\pm$ 0.5 with a

Figure 40 – Accuracies for Kingdom database using different number of bins.



Source: Developed by the author

Table 13 – Accuracies and standard deviations (in percentage) of different combinations of histogram sizes in the real world databases. The method taken here is aggregation, where we concatenate features from different K-cores. Blue colored results indicate best numbers for each database.

L	**Kingdom**	**Animal**	**Fungi**	**Plant**	**Firmicutes-Baccilis**	**Actinobacteria**
20	**99.7 ±0.3**	99.7 ±0.7	80.8 ±2.1	54.8 ±5.2	96.7 ±0.3	**98.3 ±0.7**
40	99.2 ±0.2	**99.8 ±0.5**	83.3 ±2.7	56.7 ±4.3	97.4 ±0.4	97.7 ±0.4
60	98.4 ±0.4	99.7 ±1.0	90.5 ±1.5	58.3 ±3.7	**98.0 ±0.3**	98.0 ±0.4
80	97.7 ±0.7	**99.8 ±0.5**	81.8 ±3.0	68.2 ±4.7	**98.0 ±0.3**	97.7 ±0.2
100	97.4 ±0.2	100.0 ±0.0	89.0 ±2.6	68.0 ±4.5	97.0 ±0.3	**98.2 ±0.6**
120	97.1 ±0.6	100.0 ±0.0	**90.8 ±2.5**	66.8 ±5.6	97.9 ±0.2	97.7 ±0.5
140	98.1 ±0.6	100.0 ±0.0	87.2 ±2.6	**72.8 ±5.4**	97.8 ±0.2	97.6 ±0.4
160	97.2 ±0.6	100.0 ±0.0	86.2 ±1.5	69.5 ±4.8	97.8 ±0.2	97.6 ±0.3

Source: Developed by the author

histogram size of 80. The Plant database achieved the highest accuracy of 72.8% ± 5.4 with a histogram size of 140, indicating improved performance. However, the result for L = 80 is 68.2% ± 4.7, which indicates a similar result to L = 140 if standard deviation is considered. Furthermore, the Firmicutes-Baccilis database achieved the highest accuracy of 98.0% ± 0.3 with histogram sizes of 60 and 80. Finally, the Actinobacteria database achieved the highest accuracy of 98.3% ± 0.7 with a histogram size of 20, which is one of our selected parameters.

Finally, Table 14 presents a comparative analysis of the accuracy and standard deviation of our proposed approach with three existing methods: LLNA, BP-LLNA, and D-TEP [Miranda, Machicao and Bruno 2016, Zielinski *et al.* 2022, Ribas, Machicao and Bruno 2020]. The results highlight the effectiveness of our approach in achieving higher accuracies and lower standard

Table 14 – Comparative accuracy and standard deviation of the proposed approach with three existing methods: LLNA, BP-LLNA and D-TEP

	PROPOSED		**LITERATURE**		
	$\vec{\Omega}_{(20)}$	$\vec{\Omega}_{(80)}$	LLNA	D-TEP	BP-LLNA
Kingdom	**99.7% ± 0.3**	97.7% ± 0.7	93.1% ± 5.38	96.24% ± 0.35	97.44% ± 3.98
Animal	99.7% ± 0.7	99.8% ± 0.5	77.25% ± 16.2	**100.00% ± 0.00**	84.87% ± 15.25
Fungi	80.8% ± 2.1	**81.8% ± 3.0**	54.58% ± 19.38	81.00% ± 4.38	76.17% ± 17.45
Plant	54.8% ± 5.2	68.2% ± 4.7	69.70% ± 4.67	**79.58% ± 2.12**	74.81% ± 5.64
Firmicutes-Baccilis	96.7% ± 0.3	**98.0% ± 0.3**	84.63% ± 2.00	95.73% ± 0.34	**98.30% ± 1.17**
Actinobacteria	**98.3% ± 0.7**	97.7% ± 0.2	91.48% ± 1.6	97.65% ± 0.29	95.13% ± 1.22

Source: Developed by the author

deviations in several databases.

For the Kingdom database, our proposed approach achieved an impressive accuracy of 99.7% ± 0.3, outperforming LLNA (93.10% ± 5.38), BP-LLNA (97.44% ± 3.98) and D-TEP (96.24% ± 0.35). In the Animal database, our approach achieved a remarkable accuracy of 99.7% ± 0.7, surpassing LLNA (77.25% ± 16.2) and BP-LLNA (84.87% ± 15.25) and competitive result with D-TEP. However, the inspiration for our approach acchieved 100% of accuracy for this dataset.

Moreover, for Fungi database, our approach obtained an accuracy of 81.8% ± 3.0, the highest result regarding all methodologies. However, our approach demonstrated comparable performance to LLNA in the Plant database, achieving an accuracy of 68.2% ± 4.7 compared to LLNA's 69.70% ± 4.67 but fails to superpass the outcome from D-TEP.

Furthermore, our proposed approach showed excellent accuracy in the Firmicutes-Baccilis and Actinobacteria databases, with accuracies of 98.0% ± 0.3 and 98.3% ± 0.7, respectively. These results outperformed all previous LLNA-based techniques.

In summary, the comparative analysis reveals the superior performance of our proposed in three out of six datasets and a competitive accuracy regarding Firmicutes-baccilis. The results highlights the overall importance of existing LLNA-based techniques as well as the combination of D-TEP (and SD-TEP) approach with multiscale analysis.

8.5 Conclusion

This research paper presents an improved feature extraction method that leverages the density time-evolution pattern (D-TEP) and state density time-evolution pattern (SD-TEP) from Life-like network automata. Additionally, this approach incorporates a multiscale decomposition technique to enhance the feature extraction process. By utilizing D-TEP and SD-TEP, our method captures the evolving density of alive neighbors and the state density of cells, respectively, providing a comprehensive representation of the network automata's dynamic behavior at different scales. By using the k-core decomposition, we model a different CA at each scale,

which represents a different perspective of the initial graph.

After the evaluation of the bin parameter, the results showed bythe proposed combination of global, degree, and temporal histograms features from the K-TEP, k-core time-evolution patterns, we demonstrate the robustness of our method for real datasets. The results are competitive with previous LLNA-based techniques showing the importance of the proposed approach. Furthermore, it is importance to highlight the limitation of the methods: as the size and complexity of the graph increase, the computation time also increases. However, since the average degree of metabolic networks are low (from Chapter 7), we were able to effectively compute the features. Hence, we introduce an advanced methodology that exhibits an enhanced discriminatory nature towards networks.

CHAPTER

9

CONCLUSION

As previously mentioned, the study of complex networks has become a valuable tool for understanding interconnected systems across various domains. Its versatility and broad applicability in scientific and sociological disciplines have generated significant interest among researchers. Complex networks can represent any system where there are connections between data points, making them highly applicable in practical scenarios.

To recover our goals, the primary objective of this project was to investigate and develop innovative methodologies, motivated by our previous work in computer vision and the SCG' area of interest, to surpass existing network classification techniques in the literature. The central aim was to create efficient techniques for extracting network features and evaluate their effectiveness based on classification accuracy. It is worth noting that our proposed approaches were designed to identify patterns across various applications, regardless of the original data format. This was based on the understanding that any system characterized by interconnected elements could be represented as a complex network. Two main proposals were introduced in this project: the utilization of the adjacency matrix as a key component of graph analysis and the application of multiscale techniques for graph analysis.

Our first proposed methodologies included two versions: binary and colored evaluation of graph matrices. First, Chapter 5 investigated the use of the adjacency matrix as the primary input for network analysis. The adjacency matrix provided a one-to-one representation of the network, but its unordered nature required an ordination based on node centrality to generate a suitable representation for image analysis. Visual analysis demonstrated that the proposed data transformation accentuated the patterns within the models. Furthermore, quantitative analysis was conducted on twelve different datasets, including seven real systems, and the results indicated that, in general, all proposed approaches and compared methods yielded satisfactory classification accuracies. The proposed approach of sorting and analyzing the adjacency matrix proved to be well-suited for further investigation in practical domains, despite the limitations imposed by increasing graph size, which consequently enlarged the image and increased computational time.

The findings suggested that a straightforward representation like the adjacency matrix served as a valuable source of information for network classification with results over 90% for several real datasets.

For the case of colored CN-based approach, Chapter 6 explored the application of an image generation method that combined vertex similarity, computed using the Jaccard index, with the sorting ordination of matrices as previously proposed method. The results from the proposed technique demonstrated the efficacy of these images, revealing interesting patterns and structures in the visual inspection and qualitative analysis over synthetic and real datasets, overcoming several classification results from the literature. The utilization of the sorting technique, applied in Chapters 5 and 6, instilled confidence in the consistency and stability of these patterns, regardless of the input format. This consistency allowed for the creation of a unified and coherent representation suitable for image-based analysis and later classification of the whole network with low parametrization.

Furthermore, Chapters 7 and 8 explores another computer vision inspired technique, the scale-decomposition. First, the approach from Chapter 7 involved decomposing the dataset into a set of subgraphs using k-core decomposition, which provided a multilayer perspective of the original dataset. This allowed for representation at different levels of detail by gradually removing vertices at each core.

The decomposed dataset was then subjected to feature extraction techniques, extracting either images (if sorted adjacency matrix was applied) or graphs (when classical metrics were used). These extracted features were utilized for network class labeling, and the performance of the proposed methodologies in synthetic and real datasets was compared to established techniques from the literature superpassing the outcome in many situations. The research findings demonstrated the competitiveness of the proposed methodologies compared to existing methods in the literature, specially LLNA-based techniques.

Finally, Chapter 8 presented an improved feature extraction method that leveraged the density time-evolution pattern (D-TEP) and state density time-evolution pattern (SD-TEP) from Life-like network automata. The approach also incorporated a multiscale decomposition technique to enhance the feature extraction process. By utilizing D-TEP and SD-TEP, the method captured the evolving density of alive neighbors and the state density of cells, respectively, providing a comprehensive representation of the dynamic behavior of the network automata at different scales. The k-core decomposition was used to model a different cellular automaton (CA) at each scale, representing a different perspective of the initial graph. The results demonstrated the effectiveness of the proposed combination of global, degree, and temporal histograms features from the k-core time-evolution patterns (K-TEP) for real datasets. The method showed competitiveness with previous LLNA-based techniques, highlighting the importance of the proposed approach. However, it is important to note the limitation of increased computation time with larger and more complex graphs.

The multiscale approach also provided important results regarding shape classification. Annex C models shape textures as complex networks and also uses the distance transform algorithm to enhance feature extraction the images. The results, after parameter selection, provides competitive results compared to the literature and can be fully accessed in [Ribas, Neiva and Bruno 2019].

Moreover, as a secondary goal, we model healthcare data as complex networks in order to evaluate the complexity of rare disease codification and virus transmission. First, Annex A shows the results of modeling the cross-reference between two different disease nomenclature to show the lack of representative on applied the ICD-10 to RDs in constrast to the ORPHA terminology. The web-app system allows users to interact with the network and its components in order to visually the model and retrieve CN statistics such as average degree and number os components.

Modeling is also used in Annex B and [Scabini *et al.* 2021]. In the work, CN are used to model the social dynamics of several domains in the society in order to evaluate the evolution os COVID-19 trasnmittion over different constraints. The results from the SIR-based model fitted the reality of the pandemic.

On general, our goals regarding graph classification and modeling were achieved. In summary:

- We conducted a investigation and analysis of pattern recognition methods based on complex networks, introducing new methodologies for characterizing complex networks using two-dimensional and multiscale representations
- The results are competitive with previous network classification in the literature, and carefully comparing the effectiveness of these methods with existing approaches found in the literature.
- We evaluated the methodologies in real networks, proving a base for further application of the techniques.
- The proposed methodologies surpass existing methods in the literature for characterizing complex networks, consistently yielding superior results.
- The proposed approaches, in general, require minimal parameterization compared to the compared methodology, reducing complexity.
- The potential for combining the proposed approaches with various image exploration techniques, including convolutional neural networks, enhances overall results.
- The application of CN knowledge were demonstrated in two distinct contexts within health data: rare diseases and the COVID-19 pandemic. Both modeling approaches from CN shows the effective of the structure to expose the complexity of real phenomena.

9.1 Future Work

Future work can focus on the incorporation of additional network metrics to enhance the CN-image and reveal intricate patterns in underlying network structures as well as different CNN methods applied to retrieve features from images. Furthermore, it is important to seek for new methodologies to minimize the standard deviation in classification for real datasets remains, a critical challenge to improve accuracy and robustness.

Moreover, it is important to note the limitation of increased computation time with larger and more complex graphs. However, the effectiveness of the methodology was demonstrated, particularly for networks with low average degree. The advanced approach provided an enhanced discriminatory nature towards networks, contributing to the field of complex network analysis. Future research directions can also address challenges related to dimensionality reduction and feature selection. By further refining and advancing the proposed approaches, the field of complex network analysis can benefit from improved methodologies and increased accuracy.

Regarding the modeling process, althought COVID-19 pandemic is now controlled due to the global immunization, the knowledge produced during the last years will be able to help future epidemic problems, specially those with human-human transmission.

In the context of RDs, many works can be developed with the use of CN. For instance, cross-reference modeling between disease terminologies can be extended to the new ICD-11 [Organization 2022].

In conclusion, this research significantly contributed to the classification and modeling of complex networks. The developed methodologies surpassed existing methods, showcasing their potential for various applications. The findings and future directions presented in this thesis open avenues for further exploration and refinement, advancing the field of complex network analysis and its applications.

9.2 Scientific publications

As the thesis has two branches, network classification and modeling, we developed scientific results in these two areas of graph theory.

Graph Classification

Chapters 5, 6, 7 and 8 are results regarding CN classification and while the first chapter is available freely at [Neiva and Bruno 2023] it is also being reviewed in Physica A. The following chapters are in the line for submission, while Annex C is published at Information Sciences, Annex A is accepted in a computr science conference and Annex B is published at Physica A.

- Exploring ordered patterns in the adjacency matrix for improving machine learning on

complex networks - Neiva, M. B., & Bruno, O. M. (2023). arXiv preprint arXiv:2301.08364. [Neiva and Bruno 2023]. **At minor revisions at Physica A**.

- Enhancing Information on Ordered Patterns for Complex Network Classification through Exploration of Vertex Similarity - Neiva, M. B., & Bruno, O. M. - to be submitted
- Classifying Complex Networks using Multiscale analysis - Neiva, M. B., & Bruno, O. M. - to be submitted
- K-DTEP, a tool for classifying K-core generated networks - Neiva, M. B., Zielinski, K. M. C. & Bruno, O. M. - to be submitted
- Distance transform network for shape analysis - Ribas, L. C., Neiva, M.B. & Bruno, O. M. **Information Sciences** 470 (2019): 28-42. [Ribas, Neiva and Bruno 2019]

Graph Modeling

- ICD-10 - ORPHA: An Interactive Complex Network Model for Brazilian Rare Diseases - Neiva, M. B., Oliveira, B. M., Schmidt, A. M. Scheibe, V. M., Milke, J. C., Santos, M. L., Yamada, D. B., Filho, M. E. C., Soares, G. T., Ribeiro, Y. A., Bruno, O. M. Félix, T. M. & Alves, D. - **Accepted to HCIST** (International Conference on Health and Social Care Information Systems and Technologies in 2023).
- Social interaction layers in complex networks for the dynamical epidemic modeling of COVID-19 in Brazil. Scabini, L. F., Ribas, L. C., Neiva, M. B., Junior, A. G., Farfan, A. J., & Bruno, O. M. (2021). **Physica A:** Statistical Mechanics and its Applications, 564, 125498. [Scabini *et al.* 2021]

BIBLIOGRAPHY

ABADI, M.; AGARWAL, A.; BARHAM, P.; BREVDO, E.; CHEN, Z.; CITRO, C.; CORRADO, G. S.; DAVIS, A.; DEAN, J.; DEVIN, M. *et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. **arXiv preprint arXiv:1603.04467**, 2016. Citation on page 51.

ALVAREZ-HAMELIN, J. I.; DALL'ASTA, L.; BARRAT, A.; VESPIGNANI, A. k-core decomposition: A tool for the visualization of large scale networks. **arXiv preprint cs/0504107**, 2005. Citations on pages 45, 92, and 93.

ALVES, D.; YAMADA, D. B.; BERNARDI, F. A.; CARVALHO, I.; FILHO, M. E. C.; NEIVA, M. B.; LIMA, V. C.; FÉLIX, T. M. *et al.* Mapping, infrastructure, and data analysis for the brazilian network of rare diseases: protocol for the rarasnet observational cohort study. **JMIR Research Protocols**, JMIR Publications Inc., Toronto, Canada, v. 10, n. 1, p. e24826, 2021. Citation on page 134.

AMBROSIANO, J.; SIMS, B.; BARTLOW, A. W.; ROSENBERGER, W.; RESSLER, M.; FAIR, J. M. Ontology-based graphs of research communities: a tool for understanding threat reduction networks. **Frontiers in Research Metrics and Analytics**, Frontiers Media SA, v. 5, p. 3, 2020. Citation on page 136.

AMIB. **Brazilian intensive care medicine association: Updated data on ICU beds in Brazil**. 2020, visited on 2020–05–08. <https://www.amib.org.br/fileadmin/user_upload/amib/2020/abril/28/dados_uti_amib.pdf>. Citation on page 158.

ANDERSON, R. M.; ANDERSON, B.; MAY, R. M. **Infectious diseases of humans: dynamics and control**. [S.l.]: Oxford university press, 1992. Citation on page 145.

ATAER-CANSIZOGLU, E.; BAS, E.; KALPATHY-CRAMER, J.; SHARP, G. C.; ERDOGMUS, D. Contour-based shape representation using principal curves. **Pattern Recognition**, v. 46, n. 4, p. 1140–1150, 2013. Citation on page 163.

AURELIANO, W.; GIBBON, S. Judicialisation and the politics of rare disease in brazil: Rethinking activism and inequalities. **Critical Medical Anthropology: Perspectives in and from Latin America**, UCL Press, p. 248–269, 2020. Citation on page 136.

BACKER, J. A.; KLINKENBERG, D.; WALLINGA, J. Incubation period of 2019 novel coronavirus (2019-ncov) infections among travellers from wuhan, china, 20–28 january 2020. **Eurosurveillance**, European Centre for Disease Prevention and Control, v. 25, n. 5, p. 2000062, 2020. Citation on page 152.

BACKES, A. R.; BRUNO, O. M. Shape classification using complex network and multi-scale fractal dimension. **Pattern Recognition Letters**, v. 31, n. 1, p. 44–51, 2010. Citations on pages 163 and 168.

____. Shape skeleton classification using graph and multi-scale fractal dimension. In: SPRINGER. **International Conference on Image and Signal Processing**. [S.l.], 2010. p. 448–455. Citation on page 101.

____. Texture analysis using volume-radius fractal dimension. **Applied Mathematics and Computation**, Elsevier, v. 219, n. 11, p. 5870–5875, 2013. Citations on pages 28, 29, 55, 56, 57, and 165.

BACKES, A. R.; CASANOVA, D.; BRUNO, O. M. A complex network-based approach for boundary shape analysis. **Pattern Recognition**, Elsevier, v. 42, n. 1, p. 54–67, 2009. Citations on pages 28 and 55.

____. A complex network-based approach for boundary shape analysis. **Pattern Recognition**, v. 42, n. 1, p. 54–67, 2009. Citations on pages 163, 165, 168, 169, 171, 172, 173, 176, and 180.

____. Plant leaf identification based on volumetric fractal dimension. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 23, n. 6, p. 1145–1160, 2009. Citation on page 164.

____. Contour polygonal approximation using the shortest path in networks. **International Journal of Modern Physics C**, World Scientific, v. 25, n. 02, p. 1350090, 2014. Citation on page 56.

BACKES, A. R.; FLORINDO, J. B.; BRUNO, O. M. Shape analysis using fractal dimension: A curvature based approach. **Chaos**, v. 22, n. 4, p. 043103(1–8), 2012. Citation on page 163.

BACKES, A. R.; MARTINEZ, A. S.; BRUNO, O. M. Texture analysis based on maximum contrast walker. **Pattern Recogn. Lett.**, Elsevier Science Inc., New York, NY, USA, v. 31, n. 12, p. 1701–1707, Sep. 2010. ISSN 0167-8655. Citation on page 164.

____. Texture analysis using graphs generated by deterministic partially self-avoiding walks. **Pattern Recognition**, Elsevier, v. 44, n. 8, p. 1684–1689, 2011. Citation on page 55.

BAI, X.; LATECKI, L. J. Path similarity skeleton graph matching. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 30, n. 7, p. 1282–1292, 2008. Citation on page 164.

BAI, X.; YANG, X.; YU, D.; LATECKI, L. J. Skeleton-based Shape Classification Using Path Similarity. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 22, n. 04, p. 733–746, 2008. Citation on page 163.

BAI, Y.; YAO, L.; WEI, T.; TIAN, F.; JIN, D.-Y.; CHEN, L.; WANG, M. Presumed asymptomatic carrier transmission of covid-19. **Jama**, 2020. Citation on page 151.

BAILEY, N. T. *et al.* **The mathematical theory of infectious diseases and its applications**. [S.l.]: Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975. Citation on page 145.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **science**, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999. Citations on pages 28, 37, 43, 59, 67, and 92.

BASTOS, S. B.; CAJUEIRO, D. O. Modeling and forecasting the covid-19 pandemic in brazil. **arXiv preprint arXiv:2003.14288**, 2020. Citation on page 145.

BATAGELJ, V.; ZAVERŠNIK, M. Generalized cores. **arXiv preprint cs/0202039**, 2002. Citation on page 45.

BEHRISCH, M.; SCHRECK, T.; PFISTER, H. Guiro: user-guided matrix reordering. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 26, n. 1, p. 184–194, 2019. Citation on page 64.

BENGIO, Y. Practical recommendations for gradient-based training of deep architectures. **Neural Networks: Tricks of the Trade: Second Edition**, Springer, p. 437–478, 2012. Citation on page 50.

BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 35, n. 8, p. 1798–1828, 2013. Citation on page 50.

BERGSTRA, J.; BASTIEN, F.; BREULEUX, O.; LAMBLIN, P.; PASCANU, R.; DELALLEAU, O.; DESJARDINS, G.; WARDE-FARLEY, D.; GOODFELLOW, I.; BERGERON, A. *et al.* Theano: Deep learning on gpus with python. In: CITESEER. **NIPS 2011, BigLearning Workshop, Granada, Spain**. [S.l.], 2011. v. 3, n. 0. Citation on page 51.

BHADANGE, M. S.; KALSHETTY, Y. R. Querying images by content using color, texture and shape. Citation on page 164.

BHATRAJU, P. K.; GHASSEMIEH, B. J.; NICHOLS, M.; KIM, R.; JEROME, K. R.; NALLA, A. K.; GRENINGER, A. L.; PIPAVATH, S.; WURFEL, M. M.; EVANS, L. *et al.* Covid-19 in critically ill patients in the seattle region—case series. **New England Journal of Medicine**, Mass Medical Soc, 2020. Citation on page 152.

BIASOTTI, S.; MARINI, S.; SPAGNUOLO, M.; FALCIDIENO, B. Sub-part correspondence by structural descriptors of 3d shapes. **Computer-Aided Design**, Elsevier, v. 38, n. 9, p. 1002–1019, 2006. Citation on page 164.

CARMI, S.; HAVLIN, S.; KIRKPATRICK, S.; SHAVITT, Y.; SHIR, E. A model of internet topology using k-shell decomposition. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 104, n. 27, p. 11150–11154, 2007. Citation on page 93.

CARMI, S.; HAVLIN, S.; SONG, C.; WANG, K.; MAKSE, H. A. Energy-landscape network approach to the glass transition. **Journal of Physics A: Mathematical and Theoretical**, IOP Publishing, v. 42, n. 10, p. 105101, 2009. Citation on page 101.

CASANOVA, D. **Redes complexas em visão computacional com aplicações em bioinformática**. Phd Thesis (PhD Thesis) — Universidade de São Paulo, 2018. Citation on page 28.

CASTRO, M. C.; CARVALHO, L. R. de; CHIN, T.; KAHN, R.; FRANCA, G. V.; MACARIO, E. M.; OLIVEIRA, W. K. de. Demand for hospitalization services for covid-19 patients in brazil. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020. Citation on page 144.

CHAKRABARTI, K.; ORTEGA-BINDERBERGER, M.; PORKAEW, K.; MEHROTRA, S. Similar shape retrieval in mars. In: IEEE. **Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on**. [S.l.], 2000. v. 2, p. 709–712. Citation on page 164.

CHALUMEAU, T.; COSTA, L. d.; LALIGANT, O.; MERIAUDEAU, F. Texture discrimination using hierarchical complex networks. In: DAMIANI, E.; YATONGNON, K.; SCHELKENS, P.; DIPANDA, A.; LEGRAND, L.; CHBEIR, R. (Ed.). **Signal Processing for Image Enhancement and Multimedia Processing**. [S.l.]: Springer US, 2008, (Multimedia Systems and Applications Series, v. 31). p. 95–102. Citation on page 165.

CHALUMEAU, T.; COSTA, L. d. F.; LALIGANT, O.; MERIAUDEAU, F. Texture discrimination using hierarchical complex networks. In: **International Conference on Signal Image Technology and Internet based Systems (SITIS)**. [S.l.]: Springer US, 2006. p. 543–550. Citation on page 165.

____. Texture discrimination using hierarchical complex networks. **Signal Processing for Image Enhancement and Multimedia Processing**, Springer, p. 95–102, 2008. Citation on page 28.

CHOLLET, F. **Keras: Deep learning library for theano and tensorflow.** 2015. Available: <https://keras>. Accessed: 29/03/2023. Citation on page 51.

COLLOBERT, R.; BENGIO, S.; MARIÉTHOZ, J. **Torch: a modular machine learning software library**. [S.l.], 2002. Citation on page 51.

CONDORI, R. H.; BRUNO, O. M. Analysis of activation maps through global pooling measurements for texture classification. **Information Sciences**, Elsevier, v. 555, p. 260–279, 2021. Citation on page 61.

CORNEA, N. D.; SILVER, D.; MIN, P. Curve-skeleton properties, applications, and algorithms. **IEEE Transactions on visualization and computer graphics**, IEEE Computer Society, v. 13, n. 3, p. 0530–548, 2007. Citation on page 164.

COSTA, L. d. F. Complex networks, simple vision. **arXiv preprint cond-mat/0403346**, 2004. Citations on pages 55, 64, 71, and 137.

COSTA, L. d. F.; JR, O. N. O.; TRAVIESO, G.; RODRIGUES, F. A.; BOAS, P. R. V.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, Taylor & Francis, v. 60, n. 3, p. 329–412, 2011. Citations on pages 28, 55, 57, 77, 134, and 137.

COSTA, L. d. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. **Advances in physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007. Citations on pages 37, 55, 78, 91, 92, 165, and 168.

COSTA, L. d. F. D.; JR, R. M. C. **Shape analysis and classification: theory and practice**. [S.l.]: CRC Press, Inc., 2000. Citation on page 163.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE transactions on information theory**, IEEE, v. 13, n. 1, p. 21–27, 1967. Citation on page 52.

CROKIDAKIS, N. Data analysis and modeling of the evolution of covid-19 in brazil. **arXiv preprint arXiv:2003.12150**, 2020. Citation on page 145.

DECRETO Nº 64.881. visited on 2020–04–30. <https://www.saopaulo.sp.gov.br/wp-content/uploads/2020/03/decreto-quarentena.pdf>. Citation on page 154.

DENG, L.; YU, D. Foundations and trends in signal processing: Deep learning–methods and applications. 2014. Citation on page 49.

DIJKSTRA, E. W. A note on two problems in connexion with graphs. In: **Edsger Wybe Dijkstra: His Life, Work, and Legacy**. [S.l.: s.n.], 2022. p. 287–290. Citations on pages 36 and 38.

DOROGOVTSEV, S. N.; GOLTSEV, A. V.; MENDES, J. F. F. K-core organization of complex networks. **Physical review letters**, APS, v. 96, n. 4, p. 040601, 2006. Citation on page 93.

DOROGOVTSEV, S. N.; MENDES, J. F. **Evolution of networks: From biological nets to the Internet and WWW**. [S.l.]: OUP Oxford, 2013. Citation on page 59.

ERDÖS, P.; RÉNYI, A. On random graphs. **Publicationes Mathematicae**, v. 6, p. 290–297, 1959. Citation on page 165.

ERDOS, P.; RÉNYI, A. On the evolution of random graphs. **Publ. Math. Inst. Hung. Acad. Sci**, v. 5, n. 1, p. 17–60, 1960. Citations on pages 59, 68, and 82.

ERDÖS, P.; RéNYI, A. On the evolution of random graphs. **Publ. Math. Inst. Hung. Acad. Sci**, v. 5, p. 17–61, 1960. Citation on page 165.

ERDÖS, P.; RÉNYI, A. On the strenght of connectedness of a random graph. **Acta Mathematica Scientia Hungary**, v. 12, p. 261–267, 1961. Citation on page 165.

ERDŐS, P.; RÉNYI, A. *et al.* On the evolution of random graphs. **Publ. Math. Inst. Hung. Acad. Sci**, v. 5, n. 1, p. 17–60, 1960. Citations on pages 28 and 41.

EULER, L. Solutio problematis ad geometriam situs pertinentis. **Commentarii academiae scientiarum Petropolitanae**, p. 128–140, 1741. Citation on page 36.

EUROPEAN Conference on Rare Diseases. 2005. Accessed on June 25, 2023. Available: <https://ec.europa.eu/health/ph_threats/non_com/ev_20050622_co01_en.pdf>. Citation on page 134.

EVERITT, B.; DUNN, G. **Applied Multivariate Data Analysis**. Wiley, 2001. (A Hodder Arnold Publication). ISBN 9780340741221. Available: <https://books.google.com.br/books?id=6n2nskP-aZ4C>. Citations on pages 171 and 175.

FABBRI, R.; COSTA, L. D. F.; TORELLI, J. C.; BRUNO, O. M. 2d euclidean distance transform algorithms: A comparative survey. **ACM Computing Surveys (CSUR)**, ACM, v. 40, n. 1, p. 2, 2008. Citations on pages 164, 165, and 166.

FALOUTSOS, M.; FALOUTSOS, P.; FALOUTSOS, C. On power-law relationships of the internet topology. **ACM SIGCOMM computer communication review**, ACM New York, NY, USA, v. 29, n. 4, p. 251–262, 1999. Citation on page 92.

FARABET, C.; COUPRIE, C.; NAJMAN, L.; LECUN, Y. Learning hierarchical features for scene labeling. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 35, n. 8, p. 1915–1929, 2012. Citation on page 49.

FERGUSON, N.; LAYDON, D.; GILANI, G. N.; IMAI, N.; AINSLIE, K.; BAGUELIN, M.; BHATIA, S.; BOONYASIRI, A.; PEREZ, Z. C.; CUOMO-DANNENBURG, G. *et al.* Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. **Imperial College London**, 2020. Citations on pages 149 and 153.

FERRARI, V.; TUYTELAARS, T.; GOOL, L. V. Object detection by contour segment networks. In: **Computer Vision–ECCV 2006**. [S.l.]: Springer, 2006. p. 14–28. Citation on page 165.

FERREIRA, C. R. The burden of rare diseases. **American journal of medical genetics Part A**, Wiley Online Library, v. 179, n. 6, p. 885–892, 2019. Citation on page 134.

FLORY, P. J. Molecular size distribution in three-dimensional polymers. **Journal of the American Chemical Society**, v. 63, p. 3083–3090, 1941. Citation on page 165.

FRANCA, E. B.; PASSOS, V. M. d. A.; MALTA, D. C.; DUNCAN, B. B.; RIBEIRO, A. L. P.; GUIMARAES, M. D.; ABREU, D. M.; VASCONCELOS, A. M. N.; CARNEIRO, M.; TEIXEIRA, R. *et al.* Cause-specific mortality for 249 causes in brazil and states during 1990–2015: a systematic analysis for the global burden of disease study 2015. **Population health metrics**, BioMed Central, v. 15, n. 1, p. 1–17, 2017. Citation on page 135.

GARDNER, M. The fantastic combinations of jhon conway's new solitaire game'life. **Sc. Am.**, v. 223, p. 20–123, 1970. Citation on page 58.

GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. **Proceedings of the national academy of sciences**, National Acad Sciences, v. 99, n. 12, p. 7821–7826, 2002. Citations on pages 37 and 92.

GIUGLIANI, R.; VAIRO, F. P.; RIEGEL, M.; SOUZA, C. F. D.; SCHWARTZ, I. V.; PENA, S. D. Rare disease landscape in brazil: report of a successful experience in inborn errors of metabolism. **Orphanet journal of rare diseases**, BioMed Central, v. 11, n. 1, p. 1–8, 2016. Citation on page 134.

GOH, W.-B. Strategies for shape matching using skeletons. **Computer vision and image understanding**, Elsevier, v. 110, n. 3, p. 326–345, 2008. Citation on page 164.

GONÇALVES, W. N.; BACKES, A. R.; MARTINEZ, A. S.; BRUNO, O. M. Texture descriptor based on partially self-avoiding deterministic walker on networks. **Expert Systems with Applications**, Elsevier, v. 39, n. 15, p. 11818–11829, 2012. Citations on pages 28, 29, and 57.

GONÇALVES, W. N.; BRUNO, O. M. Dynamic texture segmentation based on deterministic partially self-avoiding walks. **Computer Vision and Image Understanding**, Elsevier, v. 117, n. 9, p. 1163–1174, 2013. Citation on page 165.

GONÇALVES, W. N.; MACHADO, B. B.; BRUNO, O. M. A complex network approach for dynamic texture recognition. **Neurocomputing**, Elsevier, v. 153, p. 211–220, 2015. Citations on pages 29, 56, 57, 64, and 165.

GONÇALVES, W. N.; MARTINEZ, A. S.; BRUNO, O. M. Complex network classification using partially self-avoiding deterministic walks. **Chaos: An Interdisciplinary Journal of Nonlinear Science**, American Institute of Physics, v. 22, n. 3, p. 033139, 2012. Citations on pages 56 and 57.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT press, 2016. Citation on page 50.

GRAZIANO, M. The organization of behavioral repertoire in motor cortex. **Annu. Rev. Neurosci.**, Annual Reviews, v. 29, p. 105–134, 2006. Citation on page 48.

GUO, Z.; ZHANG, L.; ZHANG, D. A completed modeling of local binary pattern operator for texture classification. **IEEE transactions on image processing**, IEEE, v. 19, n. 6, p. 1657–1663, 2010. Citations on pages 46 and 61.

GUYE, M.; BETTUS, G.; BARTOLOMEI, F.; COZZONE, P. J. Graph theoretical analysis of structural and functional connectivity mri in normal and pathological brain networks. **Magnetic Resonance Materials in Physics, Biology and Medicine**, Springer, v. 23, p. 409–421, 2010. Citation on page 101.

HAGBERG, A.; CONWAY, D. Networkx: Network analysis with python. **URL: https://networkx. github. io**, 2020. Citation on page 137.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778. Citations on pages 49 and 50.

HEARST, M. A.; DUMAIS, S. T.; OSUNA, E.; PLATT, J.; SCHOLKOPF, B. Support vector machines. **IEEE Intelligent Systems and their Applications**, IEEE, v. 13, n. 4, p. 18–28, 1998. Citations on pages 107, 171, and 175.

HILAGA, M.; SHINAGAWA, Y.; KOHMURA, T.; KUNII, T. L. Topology matching for fully automatic similarity estimation of 3d shapes. In: ACM. **Proceedings of the 28th annual conference on Computer graphics and interactive techniques**. [S.l.], 2001. p. 203–212. Citation on page 164.

HIRSCH, J.; NICOLA, G.; MCGINTY, G.; LIU, R.; BARR, R.; CHITTLE, M.; MANCHIKANTI, L. Icd-10: history and context. **American Journal of Neuroradiology**, Am Soc Neuroradiology, v. 37, n. 4, p. 596–599, 2016. Citation on page 135.

HU, M. K. Visual pattern recognition by moment invariants. **IRE transactions on information theory**, IEEE, v. 8, n. 2, p. 179–187, 1962. Citations on pages 47, 61, and 164.

HUANG, C.; WANG, Y.; LI, X.; REN, L.; ZHAO, J.; HU, Y.; ZHANG, L.; FAN, G.; XU, J.; GU, X. *et al.* Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. **The Lancet**, Elsevier, v. 395, n. 10223, p. 497–506, 2020. Citation on page 151.

HUBERMAN, B. A.; ROMERO, D. M.; WU, F. Social networks that matter: Twitter under the microscope. **arXiv preprint arXiv:0812.1045**, 2008. Citation on page 92.

IBGE. **Censo Demográfico: Tabela 137 - População residente, por religião**. 2010, visited on 2020–04–18. <https://sidra.ibge.gov.br/tabela/137>. Citation on page 149.

____. **Censo Demográfico: Tabela 185 - Domicílios particulares permanentes por situação e número de moradores**. 2010, visited on 2020–04–20. <https://sidra.ibge.gov.br/tabela/185>. Citation on page 147.

____. **Pesquisa Nacional por Amostra de Domicílios Contínua trimestral: Tabela 5918 - População, por grupos de idade**. 2019, visited on 2020–05–11. <https://sidra.ibge.gov.br/tabela/5918>. Citation on page 147.

INEP. **Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira: DADOS DO CENSO ESCOLAR: Ensino Médio brasileiro tem média de 30 alunos por sala**. 2018, visited on 2020–04–5. <http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/dados-do-censo-escolar-ensino-medio-brasileiro-tem-media-de-30-alunos-por-sala/21206>. Citation on page 149.

INLOCO. **Mapa brasileiro da COVID-19**. 2020, visited on 2020–05–15. <https://mapabrasileirodacovid.inloco.com.br/pt/>. Citations on pages 155, 156, 160, and 161.

IPEA. **Instituto de Pesquisa Econômica Aplicada: Sistema de Indicadores de Percepção Social (SIPS)**. 2011, visited on 2020–04–15. <http://www.ipea.gov.br/portal/images/stories/PDFs/livros/livros/livro_sistemaindicadores_sips_01.pdf>. Citation on page 148.

JIA, Y.; SHELHAMER, E.; DONAHUE, J.; KARAYEV, S.; LONG, J.; GIRSHICK, R.; GUADARRAMA, S.; DARRELL, T. Caffe: Convolutional architecture for fast feature embedding. In: **Proceedings of the 22nd ACM international conference on Multimedia**. [S.l.: s.n.], 2014. p. 675–678. Citation on page 51.

Johns Hopkins University. **Coronavirus Resource Center**. 2020, visited on 2020–05–18. <https://coronavirus.jhu.edu/map.html>. Citations on pages 17, 18, 144, 154, and 155.

JUNIOR, J. J. d. M. S.; BACKES, A. R. Shape classification using line segment statistics. **Information Sciences**, Elsevier, v. 305, p. 349–356, 2015. Citation on page 175.

KANEHISA, M.; SATO, Y.; KAWASHIMA, M.; FURUMICHI, M.; TANABE, M. Kegg as a reference resource for gene and protein annotation. **Nucleic acids research**, Oxford University Press, v. 44, n. D1, p. D457–D462, 2016. Citation on page 60.

KHOTANZAD, A.; HONG, Y. H. Invariant Image Recognition by Zernike Moments. **Ann. Oper. Res. Pattern Anal. Machine Intell. IEEE Trans. Pattern Anal. Machine Intell. J . Robotics Res. J . Robotics Res. J . ACM Networks I . J . Stoker**, v. 12, n. 14, p. 13–118, 1990. Citation on page 164.

KIM, G. un; KIM, M.-J.; RA, S. H.; LEE, J.; BAE, S.; JUNG, J.; KIM, S.-H. Clinical characteristics of asymptomatic and symptomatic patients with mild covid-19. **Clinical Microbiology and Infection**, 2020. ISSN 1198-743X. Available: <http://www.sciencedirect.com/science/article/pii/S1198743X20302688>. Citation on page 144.

KITSAK, M.; GALLOS, L. K.; HAVLIN, S.; LILJEROS, F.; MUCHNIK, L.; STANLEY, H. E.; MAKSE, H. A. Identification of influential spreaders in complex networks. **Nature physics**, Nature Publishing Group UK London, v. 6, n. 11, p. 888–893, 2010. Citation on page 45.

KOHAVI, R. *et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. **Ijcai**. [S.l.], 1995. v. 14, n. 2, p. 1137–1145. Citation on page 54.

KONIG, D. Graphok es matrixok (hungarian)[graphs and matrices]. **Matematikai és Fizikai Lapok**, v. 38, p. 116–119, 1931. Citation on page 36.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Communications of the ACM**, AcM New York, NY, USA, v. 60, n. 6, p. 84–90, 2017. Citations on pages 49 and 50.

LAI, C.-C.; LIU, Y. H.; WANG, C.-Y.; WANG, Y.-H.; HSUEH, S.-C.; YEN, M.-Y.; KO, W.-C.; HSUEH, P.-R. Asymptomatic carrier state, acute respiratory disease, and pneumonia due to severe acute respiratory syndrome coronavirus 2 (sarscov-2): Facts and myths. **Journal of Microbiology, Immunology and Infection**, Elsevier, 2020. Citation on page 152.

LAN, L.; XU, D.; YE, G.; XIA, C.; WANG, S.; LI, Y.; XU, H. Positive rt-pcr test results in patients recovered from covid-19. **Jama**, 2020. Citation on page 151.

LAUER, S. A.; GRANTZ, K. H.; BI, Q.; JONES, F. K.; ZHENG, Q.; MEREDITH, H. R.; AZMAN, A. S.; REICH, N. G.; LESSLER, J. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. **Annals of internal medicine**, 2020. Citations on pages 151 and 152.

LE, D.-H.; DANG, V.-T. Ontology-based disease similarity network for disease gene prediction. **Vietnam Journal of Computer Science**, Springer, v. 3, n. 3, p. 197–205, 2016. Citations on pages 136 and 138.

LEIBE, B.; SCHIELE, B. Analyzing appearance and contour based methods for object categorization. In: IEEE. **Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on**. [S.l.], 2003. v. 2, p. II–409. Citations on pages 171 and 172.

LESKOVEC, J.; KREVL, A.; DATASETS, S. Snap datasets: Stanford large network dataset collection. **URL: http://snap. stanford. edu/data/index. html**, 2014. Citation on page 60.

LIAO, S. X.; PAWLAK, M. On image analysis by moments. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 18, n. 3, p. 254–266, 1996. Citation on page 164.

LINHARES, C. D.; PONCIANO, J. R.; PAIVA, J. G. S.; TRAVENÇOLO, B. A.; ROCHA, L. E. A comparative analysis for visualizing the temporal evolution of contact networks: a user study. **Journal of Visualization**, Springer, v. 24, p. 1011–1031, 2021. Citation on page 64.

LINTON, N. M.; KOBAYASHI, T.; YANG, Y.; HAYASHI, K.; AKHMETZHANOV, A. R.; JUNG, S.-m.; YUAN, B.; KINOSHITA, R.; NISHIURA, H. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data. **Journal of clinical medicine**, Multidisciplinary Digital Publishing Institute, v. 9, n. 2, p. 538, 2020. Citation on page 152.

LONCARIC, S. A survey of shape analysis techniques. **Pattern recognition**, Elsevier, v. 31, n. 8, p. 983–1001, 1998. Citation on page 163.

LU, G.; SAJJANHAR, A. Region-based shape representation and similarity measure suitable for content-based image retrieval. **Multimedia Systems**, Springer, v. 7, n. 2, p. 165–174, 1999. Citation on page 164.

MACHICAO, J.; JR, E. A. C.; MIRANDA, G. H.; AMANCIO, D. R.; BRUNO, O. M. Authorship attribution based on life-like network automata. **PloS one**, Public Library of Science San Francisco, CA USA, v. 13, n. 3, p. e0193703, 2018. Citations on pages 57, 64, and 102.

MARINHO, V. Q.; HIRST, G.; AMANCIO, D. R. Authorship attribution via network motifs identification. In: IEEE. **2016 5th Brazilian conference on intelligent systems (BRACIS)**. [S.l.], 2016. p. 355–360. Citation on page 101.

MAURER, C. R.; QI, R.; RAGHAVAN, V. A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 25, n. 2, p. 265–270, 2003. Citation on page 166.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, p. 115–133, 1943. Citations on pages 47 and 48.

MILGRAM, S. The small world problem. **Psychology today**, New York, v. 2, n. 1, p. 60–67, 1967. Citation on page 146.

MING-HUA, C.; PING-FAN, Y. Multiscaling approach based on morphological filtering. **Science in China Series A-Mathematics, Physics, Astronomy & Technological Science**, Science China Press, v. 32, n. 4, p. 492–503, 1989. Citation on page 165.

MIRANDA, G. H. B.; MACHICAO, J.; BRUNO, O. M. Exploring spatio-temporal dynamics of cellular automata for pattern recognition in networks. **Scientific Reports**, Nature Publishing Group UK London, v. 6, n. 1, p. 37329, 2016. Citations on pages 28, 29, 56, 57, 58, 59, 64, 65, 70, 72, 78, 84, 85, 92, 95, 99, 102, 103, 104, 107, 108, and 165.

MIZUMOTO, K.; KAGAYA, K.; ZAREBSKI, A.; CHOWELL, G. Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020. **Eurosurveillance**, European Centre for Disease Prevention and Control, v. 25, n. 10, p. 2000180, 2020. Citation on page 151.

MOKHTARIAN, F.; BOBER, M. **Curvature scale space representation: theory, applications, and MPEG-7 standardization**. [S.l.]: Springer Science & Business Media, 2013. Citation on page 164.

MOLINER, A. M.; WALIGÓRA, J. The european union policy in the field of rare diseases. **Public health genomics**, S. Karger AG Basel, Switzerland, v. 16, n. 6, p. 268–277, 2014. Citation on page 134.

MOORE, C.; NEWMAN, M. E. Epidemics and percolation in small-world networks. **Physical Review E**, APS, v. 61, n. 5, p. 5678, 2000. Citation on page 146.

NEIVA, M. B. **Métodos de pré-processamento de texturas para otimizar o reconhecimento de padrões**. Phd Thesis (PhD Thesis) — Universidade de São Paulo, 2016. Citation on page 29.

NEIVA, M. B.; BRUNO, O. M. Exploring ordered patterns in the adjacency matrix for improving machine learning on complex networks. **arXiv preprint arXiv:2301.08364**, 2023. Citations on pages 16, 63, 78, 79, 80, 81, 83, 84, 90, 92, 94, 96, 97, 99, 114, and 115.

NEIVA, M. B.; GUIDOTTI, P.; BRUNO, O. M. Improving lbp and its variants using anisotropic diffusion. **arXiv preprint arXiv:1703.04418**, 2017. Citation on page 29.

NEIVA, M. B.; VACAVANT, A.; BRUNO, O. M. Improving texture extraction and classification using smoothed morphological operators. **Digital Signal Processing**, Elsevier, v. 83, p. 24–34, 2018. Citation on page 29.

NEWMAN, M. E. Modularity and community structure in networks. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 103, n. 23, p. 8577–8582, 2006. Citation on page 92.

____. Complex systems: A survey. **arXiv preprint arXiv:1112.1440**, 2011. Citation on page 92.

NEWMAN, M. E.; WATTS, D. J. Scaling and percolation in the small-world network model. **Physical review E**, APS, v. 60, n. 6, p. 7332, 1999. Citation on page 146.

OJALA, T.; PIETIKÄINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. **Pattern recognition**, Elsevier, v. 29, n. 1, p. 51–59, 1996. Citation on page 46.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, v. 24, n. 7, p. 971–987, 2002. Citation on page 61.

ORGANIZATION, W. H. **International Classification of Diseases (ICD) Information Sheet**. World Health Organization, 2022. Accessed on February 25, 2023. Available: <http://www.who.int/classifications/icd/factsheet/en/>. Citations on pages 30, 114, and 135.

ORGANIZATION, W. H. *et al.* **International classification of diseases:[9th] ninth revision, basic tabulation list with alphabetic index**. [S.l.]: World Health Organization, 1978. Citation on page 134.

ORPHANET: Orphadata. 2022. Available: <http://www.orphadata.org/cgi-bin/index.php>. Citations on pages 134 and 136.

OSOWSKI, S. *et al.* Fourier and wavelet descriptors for shape recognition using neural networks—a comparative study. **Pattern Recognition**, Elsevier, v. 35, n. 9, p. 1949–1957, 2002. Citation on page 174.

PALLA, G.; DERÉNYI, I.; FARKAS, I.; VICSEK, T. Uncovering the overlapping community structure of complex networks in nature and society. **nature**, Nature Publishing Group UK London, v. 435, n. 7043, p. 814–818, 2005. Citation on page 101.

PASTOR-SATORRAS, R.; CASTELLANO, C.; MIEGHEM, P. V.; VESPIGNANI, A. Epidemic processes in complex networks. **Reviews of modern physics**, APS, v. 87, n. 3, p. 925, 2015. Citation on page 137.

PERRONE, G.; UNPINGCO, J.; LU, H.-m. Network visualizations with pyvis and visjs. **arXiv preprint arXiv:2006.04951**, 2020. Citation on page 138.

PETERSEN, J. Die theorie der regulären graphs. 1891. Citation on page 36.

PLOTZE, R. d. O.; FALVO, M.; PÁDUA, J. G.; BERNACCI, L. C.; VIEIRA, M. L. C.; OLIVEIRA, G. C. X.; BRUNO, O. M. Leaf shape analysis using the multiscale minkowski fractal dimension, a new morphometric method: a study with passiflora (passifloraceae). **Canadian Journal of Botany**, NRC Research Press, v. 83, n. 3, p. 287–301, 2005. Citation on page 164.

PORTAL da Transparência - Painel COVID Registral. visited on 2020–05–13. <https://transparencia.registrocivil.org.br/registral-covid>. Citations on pages 18 and 156.

PORTAL, E. O. D. **COVID-19 Coronavirus data**. 2020, visited on 2020–05–18. <https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data/resource/55e8f966-d5c8-438e-85bc-c7a5a26f4863>. Citations on pages 18, 154, and 155.

QI, C.; KARLSSON, D.; SALLMEN, K.; WYSS, R. Model studies on the covid-19 pandemic in sweden. **arXiv preprint arXiv:2004.01575**, 2020. Citation on page 145.

RAIN, J.-C.; SELIG, L.; REUSE, H. D.; BATTAGLIA, V.; REVERDY, C.; SIMON, S.; LENZEN, G.; PETEL, F.; WOJCIK, J.; SCHÄCHTER, V. *et al.* The protein–protein interaction map of helicobacter pylori. **Nature**, Nature Publishing Group UK London, v. 409, n. 6817, p. 211–215, 2001. Citation on page 101.

REDNER, S. How popular is your paper? an empirical study of the citation distribution. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 4, n. 2, p. 131–134, 1998. Citation on page 92.

RIBAS, L. C.; JUNIOR, J. J.; SCABINI, L. F.; BRUNO, O. M. Fusion of complex networks and randomized neural networks for texture analysis. **arXiv preprint arXiv:1806.09170**, 2018. Citation on page 102.

RIBAS, L. C.; MACHICAO, J.; BRUNO, O. M. Life-like network automata descriptor based on binary patterns for network classification. **Information Sciences**, Elsevier, v. 515, p. 156–168, 2020. Citations on pages 58, 59, 64, 65, 70, 72, 78, 84, 85, 92, 95, 99, 102, 103, 107, and 108.

RIBAS, L. C.; NEIVA, M. B.; BRUNO, O. M. Distance transform network for shape analysis. **Information Sciences**, v. 470, p. 28–42, 2019. ISSN 0020-0255. Available: <https://www.sciencedirect.com/science/article/pii/S0020025518306418>. Citations on pages 30, 64, 113, 163, and 171.

____. Distance transform network for shape analysis. **Information Sciences**, Elsevier, v. 470, p. 28–42, 2019. Citation on page 115.

RIBAS, L. C.; SCABINI, L.; BRUNO, O. M. A complex network approach for fish species recognition based on otolith shape. In: IEEE. **2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)**. [S.l.], 2022. p. 1–5. Citations on pages 55, 56, 64, and 101.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958. Citation on page 48.

____. **Principles of neurodynamics. perceptrons and the theory of brain mechanisms**. [S.l.], 1961. Citation on page 48.

ROSENFELD, A.; PFALTZ, J. Sequential operations in digital picture processing. **Journal of the ACM**, v. 13, n. 4, p. 471–494, 1966. Citation on page 164.

RUSS, J. C. **The image processing handbook**. [S.l.]: CRC press, 2016. Citation on page 165.

SACHS, H.; STIEBITZ, M.; WILSON, R. J. An historical note: Euler's königsberg letters. **Journal of Graph Theory**, Wiley Online Library, v. 12, n. 1, p. 133–139, 1988. Citation on page 36.

SAFAR, M.; SHAHABI, C.; SUN, X. Image retrieval by shape: a comparative study. In: IEEE. **Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on**. [S.l.], 2000. v. 1, p. 141–144. Citation on page 164.

SAIN, S. R. **The nature of statistical learning theory**. [S.l.]: Taylor & Francis, 1996. Citation on page 52.

SAúDE, M. da. **Portaria n^{o} 199, de 30 de janeiro de 2014**. 2014. Accessed on July 4, 2023. Available: <https://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0199_30_01_2014.html>. Citations on pages 134, 135, and 137.

SCABINI, L.; RIBAS, L.; RIBEIRO, E.; BRUNO, O. Deep topological embedding with convolutional neural networks for complex network classification. In: **Network Science: 7th International Winter Conference, NetSci-X 2022, Porto, Portugal, February 8–11, 2022, Proceedings**. Berlin, Heidelberg: Springer-Verlag, 2022. p. 54–66. ISBN 978-3-030-97239-4. Available: <https://doi.org/10.1007/978-3-030-97240-0_5>. Citations on pages 64 and 92.

SCABINI, L. F.; FISTAROL, D. O.; CANTERO, S. V.; GONÇALVES, W. N.; MACHADO, B. B.; JR, J. F. R. Angular descriptors of complex networks: A novel approach for boundary shape analysis. **Expert Systems with Applications**, Elsevier, v. 89, p. 362–373, 2017. Citations on pages 55, 56, 57, 64, and 92.

SCABINI, L. F.; RIBAS, L. C.; NEIVA, M. B.; JUNIOR, A. G.; FARFAN, A. J.; BRUNO, O. M. Social interaction layers in complex networks for the dynamical epidemic modeling of covid-19 in brazil. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 564, p. 125498, 2021. Citations on pages 30, 101, 113, 115, 137, and 147.

SCHRIML, L. M.; ARZE, C.; NADENDLA, S.; CHANG, Y.-W. W.; MAZAITIS, M.; FELIX, V.; FENG, G.; KIBBE, W. A. Disease ontology: a backbone for disease semantic integration. **Nucleic acids research**, Oxford University Press, v. 40, n. D1, p. D940–D946, 2012. Citation on page 136.

SCHRIML, L. M.; MUNRO, J. B.; SCHOR, M.; OLLEY, D.; MCCRACKEN, C.; FELIX, V.; BARON, J. A.; JACKSON, R.; BELLO, S. M.; BEARER, C. *et al.* The human disease ontology 2022 update. **Nucleic acids research**, Oxford University Press, v. 50, n. D1, p. D1255–D1261, 2022. Citation on page 136.

SEBASTIAN, T. B.; KLEIN, P. N.; KIMIA, B. B. Recognition of shapes by editing their shock graphs. **IEEE Transactions on pattern analysis and machine intelligence**, IEEE, v. 26, n. 5, p. 550–571, 2004. Citation on page 171.

SEIDMAN, S. B. Network structure and minimum degree. **Social networks**, Elsevier, v. 5, n. 3, p. 269–287, 1983. Citation on page 93.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on image data augmentation for deep learning. **Journal of big data**, SpringerOpen, v. 6, n. 1, p. 1–48, 2019. Citation on page 99.

SIDDIQI, K.; PIZER, S. **Medial representations: mathematics, algorithms and applications**. [S.l.]: Springer Science & Business Media, 2008. Citation on page 164.

SILVA, E. N. d.; SOUSA, T. R. V. Economic evaluation in the context of rare diseases: is it possible? **Cadernos de Saúde Pública**, SciELO Brasil, v. 31, p. 496–506, 2015. Citation on page 134.

SILVA, P. C. V. da; VELÁSQUEZ-ROJAS, F.; CONNAUGHTON, C.; VAZQUEZ, F.; MORENO, Y.; RODRIGUES, F. A. Epidemic spreading with awareness and different timescales in multiplex networks. **Physical Review E**, APS, v. 100, n. 3, p. 032313, 2019. Citation on page 146.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014. Citations on pages 49, 50, 51, and 61.

SMITH, D. M.; ONNELA, J.-P.; LEE, C. F.; FRICKER, M. D.; JOHNSON, N. F. Network automata: Coupling structure and function in dynamic networks. **Advances in Complex Systems**, World Scientific, v. 14, n. 03, p. 317–339, 2011. Citation on page 102.

SOO, I. H.-Y.; LAM, M. K.; RUST, J.; MADDEN, R. Do we have enough information? how icd-10-am activity codes measure up. **Health Information Management Journal**, SAGE Publications Sage UK: London, England, v. 38, n. 1, p. 22–34, 2009. Citation on page 135.

SORENSEN, T. A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. **Biol. Skar.**, v. 5, p. 1–34, 1948. Citation on page 79.

SURVEILLANCES, V. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19)—china, 2020. **China CDC Weekly**, v. 2, n. 8, p. 113–122, 2020. Citation on page 151.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2015. p. 1–9. Citations on pages 49 and 50.

TANG, J.; ZHANG, C.; LUO, B. Shape representation and distance measure based on relational graph. In: IEEE. **Hybrid Intelligent Systems, 2006. HIS'06. Sixth International Conference on**. [S.l.], 2006. p. 20–20. Citation on page 165.

TOMPSON, J. J.; JAIN, A.; LECUN, Y.; BREGLER, C. Joint training of a convolutional network and a graphical model for human pose estimation. **Advances in neural information processing systems**, v. 27, 2014. Citation on page 49.

TORELLI, J. C.; FABBRI, R.; TRAVIESO, G.; BRUNO, O. M. A high performance 3d exact euclidean distance transform algorithm for distributed computing. **International Journal of Pattern Recognition and Artificial Intelligence**, World Scientific, v. 24, n. 06, p. 897–915, 2010. Citation on page 165.

TORRES, R. d. S.; FALCAO, A. X.; COSTA, L. d. F. A graph-based approach for multiscale shape analysis. **Pattern Recognition**, Elsevier, v. 37, n. 6, p. 1163–1174, 2004. Citations on pages 164 and 175.

VASANT, D.; CHANAS, L.; MALONE, J.; HANAUER, M.; OLRY, A.; JUPP, S.; ROBINSON, P. N.; PARKINSON, H.; RATH, A. Ordo: an ontology connecting rare disease, epidemiology and genetic data. In: RESEARCHGATE. NET. **Proceedings of ISMB**. [S.l.], 2014. v. 30. Citation on page 136.

VEGA-REDONDO, F. **Complex social networks**. [S.l.]: Cambridge University Press, 2007. Citation on page 145.

WAKAP, S. N.; LAMBERT, D. M.; OLRY, A.; RODWELL, C.; GUEYDAN, C.; LANNEAU, V.; MURPHY, D.; CAM, Y. L.; RATH, A. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. **European Journal of Human Genetics**, Nature Publishing Group, v. 28, n. 2, p. 165–173, 2020. Citations on pages 134 and 136.

WALLACE, T. P.; WINTZ, P. A. An efficient three-dimensional aircraft recognition algorithm using normalized fourier descriptors. **Computer Graphics and Image Processing**, Elsevier, v. 13, n. 2, p. 99–126, 1980. Citation on page 164.

WANG, X.; FENG, B.; BAI, X.; LIU, W.; LATECKI, L. J. Bag of contour fragments for robust shape classification. **Pattern Recognition**, Elsevier, v. 47, n. 6, p. 2116–2125, 2014. Citations on pages 163 and 175.

WANG, Z.; CHI, Z.; FENG, D. Shape based leaf image retrieval. **IEE Proceedings-Vision, Image and Signal Processing**, IET, v. 150, n. 1, p. 34–43, 2003. Citation on page 164.

WATTS, D. Six degrees of separation. **The Science of a Connected Age, London: Vintage**, 2003. Citation on page 37.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. **Nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citations on pages 28, 37, 41, 42, 59, 68, 69, 83, and 168.

WAXMAN, B. M. Routing of multipoint connections. **IEEE journal on selected areas in communications**, IEEE, v. 6, n. 9, p. 1617–1622, 1988. Citation on page 59.

WEINREICH, S. S.; MANGON, R.; SIKKENS, J.; TEEUW, M. E.; CORNEL, M. Orphanet: a european database for rare diseases. **Nederlands tijdschrift voor geneeskunde**, v. 152, n. 9, p. 518–519, 2008. Citations on pages 31, 134, 135, 136, 139, and 140.

WHO. **World Health Organization - Coronavirus disease 2019 (COVID-19): situation report, 73**. [S.l.], 2020. Citation on page 152.

____. **World Health Organization - Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations**. 2020, visited on 2020–05–07. <www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations>. Citation on page 144.

____. **World Health Organization Coronavirus Disease (COVID-19) Dashboard**. 2020, visited on 2020–05–18. <https://covid19.who.int/>. Citations on pages 18, 154, 155, and 156.

WOLFRAM, S. Statistical mechanics of cellular automata. **Reviews of modern physics**, APS, v. 55, n. 3, p. 601, 1983. Citation on page 102.

WORLDOMETER. **Brazil Coronavirus Cases**. 2020, visited on 2020–05–18. <https://www.worldometers.info/coronavirus/country/brazil/>. Citations on pages 18, 154, and 155.

WU, W.-Y.; WANG, M.-J. J. Detecting the dominant points by the curvature-based polygonal approximation. **CVGIP: Graphical Models and Image Processing**, Elsevier, v. 55, n. 2, p. 79–88, 1993. Citation on page 174.

YANG, J.; YAO, C.; MA, W.; CHEN, G. A study of the spreading scheme for viral marketing based on a complex network model. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 389, n. 4, p. 859–870, 2010. Citation on page 27.

ZHAO, H.; KONG, M.; LUO, B. Shape representation based on polar-graph spectra. In: **Intelligent Computing in Signal Processing and Pattern Recognition**. [S.l.]: Springer, 2006. p. 900–905. Citation on page 165.

ZHAO, J.; YU, H.; LUO, J.; CAO, Z.; LI, Y. Complex networks theory for analyzing metabolic networks. **Chinese Science Bulletin**, Springer, v. 51, p. 1529–1537, 2006. Citation on page 60.

ZHENJIANG, M. Zernike moment-based image shape analysis and its application. **Pattern Recognition Letters**, v. 21, n. 2, p. 169–177, 2000. Citation on page 164.

____. Zernike moment-based image shape analysis and its application. **Pattern Recognition Letters**, v. 21, n. 2, p. 169 – 177, 2000. ISSN 0167-8655. Available: <http://www.sciencedirect.com/science/article/pii/S0167865599001440>. Citation on page 175.

ZHOU, F.; YU, T.; DU, R.; FAN, G.; LIU, Y.; LIU, Z.; XIANG, J.; WANG, Y.; SONG, B.; GU, X. *et al.* Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. **The Lancet**, Elsevier, 2020. Citation on page 152.

ZIELINSKI, K. M. C.; RIBAS, L. C.; MACHICAO, J.; BRUNO, O. M. **A Network Classification Method based on Density Time Evolution Patterns Extracted from Network Automata**. 2022. Citations on pages 30, 58, 59, 64, 65, 70, 72, 84, 85, 92, 95, 97, 101, 102, 103, 105, 107, and 108.

ANNEX

A

ICD-10 - ORPHA: AN INTERACTIVE COMPLEX NETWORK MODEL FOR BRAZILIAN RARE DISEASES

Initial Consideration

This work is the outcome of combining the knowledge from the RARAS Network group with the advantages of complex networks. It presents a draft version of a paper accepted at HCIST 2023, which will be presented in Porto, Portugal in November 2023.

Addressing the importance of rare diseases in public health, both the World Health Organization and the Brazilian Health Ministry have prioritized this area. In 2014, the Brazilian government introduced Ordinance nº199, a national policy focusing on the care of rare patients. In Brazil, diseases are classified using the International Classification of Diseases 10th Revision, a widely used terminology in this context. However, alternative terminologies, such as the ORPHAcode provided by Orphanet, also exist for coding rare diseases.

This paper proposes a **complex network model** that utilizes the relationships between these terminologies to demonstrate that the International Classification of Diseases 10th Revision may be too generic for diagnosing rare Brazilian patients. Additionally, it highlights the absence of a perfect nomenclature for defining rare diseases, emphasizing the need for context-specific applications. Mapping the relationships between different terminologies is therefore crucial in establishing consistent semantic connections in biomedical ontologies, facilitating the execution of tasks that involve multiple terminologies.

A.1 Introduction

A Rare Disease (RD) is a medical condition that is characterized by a low prevalence compared to diseases that are commonly observed in the general population. The definition of RD varies across different countries. For the European Union, a disease must affect no more than 1 in 2000 people to be considered a RD [Moliner and Waligóra 2014], while in Brazil, RDs are defined according to the World Health Organization (WHO) criteria: 1.3 cases per 2000 individuals [Giugliani *et al.* 2016].

A report from the 2005 European Conference on Rare diseases found that, given the longevity of 323 RD, 25.7% of these are potentially lethal before five years of age, 36.8% drive to a reduced life expectancy, while only 37.5% are associated with a typical lifespan [European Conference on Rare Diseases 2005, Ferreira 2019]. Also, on a macro level, although individually rare, collectively, they affect up to 10% of the total population. Thus, RD significantly impacts public health worldwide [Alves *et al.* 2021].

RDs are a global health priority for WHO [Saúde 2014]. In Brazil, a developing country, up to 15 million people are affected by a rare condition, and authorities know the importance of supporting and treating patients [Silva and Sousa 2015]. In this context, Ordinance n^{o}199, from January 30, 2014, created the National Policy for Comprehensive Care for People with Rare Diseases. The goal is to present guidelines to guarantee universality, integrality, and equity for RD patients [Saúde 2014]. Moreover, the Brazilian Ministry of Health reports all diseases using the International Classification of Diseases 10th Revision (ICD-10). However, RD specialists commonly use Orphanet terminology to refer to a rare condition [Weinreich *et al.* 2008].

During the development of this study, the Brazilian Ministry of Health utilized the ICD-10 to classify RDs, which is a recognized international standard in the medical community [Organization *et al.* 1978]. However, only half of the RDs listed in the ICD-10 have unique codes [Wakap *et al.* 2020]. To overcome this challenge, healthcare professionals frequently use the ORPHA code taxonomy, which was created specifically for classifying RD [Orphanet: Orphadata 2022]. ORPHAcodes provide a cross-referenced with ICD codes, then researchers can gain valuable insights into the RD nomenclature's coding.

In this scenario, Complex Networks (CNs) are a valuable tool for analyzing correspondence and extracting features from the cross-reference model of the Brazilian Ministry of Health's RD list. Graph theory has been widely used to represent various systems and their connections, from social analysis to ontology construction [Costa *et al.* 2011]. Therefore, this study utilizes a graph theory model to compare two important terminologies for RDs, the ICD-10 and ORPHAcode, in order to understand the local set of RDs defined by the Brazilian Ministry of Health. Additionally, an interactive dashboard was developed to provide clinical staff and other interested users with the ability to visualize the network and examine important graph characteristics.

A.2 Background

The following sections describe the two terminologies used to build the complex network model: the ICD-10 and the ORPHAcode.

A.2.1 International Classification of Diseases 10th Revision (ICD-10)

The ICD-10 is the most widely used disease classification system worldwide, with the goal of standardizing disease codification and related health problems. The ICD is part of the WHO's efforts to universalize and organize data related to conditions, procedures, mortality, and morbidity [Organization 2022]. The ICD codes are language-independent, enabling statistical comparisons across different countries worldwide.

This standard has undergone several changes over the years to reflect advances in medicine and technology and the reality of the current health area. ICD-10 has more than 70,000 codes, representing a significant increase compared to ICD-9 [Hirsch *et al.* 2016]. However, a new version, ICD-11, has been available since January 2022, with classification changes that reflect modern society's diseases and scenarios [Organization 2022]. Nevertheless, it will take a long time to implement the novel nomenclature in a developing country, such as Brazil, where the previous terminology has been used since 1996 [Franca *et al.* 2017]. According to WHO, countries using an earlier version of ICD are expected to take up to five years to implement the new terminology [Organization 2022]. By the time this study was developed, the Brazilian government had been using ICD-10 to identify and report any patient's condition.

Assigning a code to a given condition is complex, with some conditions having multiple matching ICD-10 codes. The complexity poses difficulties for the medical team and complicates the search for treatments and diagnosis [Soo *et al.* 2009]. As researchers continuously search for new knowledge regarding RD, it is understandable that such codes appear since some conditions are only partially defined. Using ICD codes can make establishing epidemiology and prognosis challenging for RD due to its generality. For instance, a single ICD-10 code, Q87.0, is related to over 130 RDs with different conditions in the ORPHAcode [Saúde 2014]. Considering the difficulty of centralizing RD in a single ICD code, an alternative classification system, the ORPHAcode terminology, has shown interest. ORPHAcodes were specifically developed for classifying RDs and are cross-referenced with ICD codes. Below, we present the ORPHAcode terminology and explain why it is a better alternative for classifying RDs.

A.2.2 ORPHAcode

A group of academic researchers and professionals from 40 countries led by the National Institute of Health and Medical Research (Inserm) developed the Orphanet, a multilingual online observatory to accomplish RDs [Weinreich *et al.* 2008]. This website contains information on over 6100 RD codes, the ORPHAcodes, and is used globally to identify RDs. Orphanet is a

comprehensive resource for RD characteristics, prevalence, orphan drugs, health centers, and other rare disease ecosystem details.

The observatory also provides, for each disease, the inheritance, age of onset, diagnosis procedures, and clinical description. Still under development, As of 2018, the system accounted for 81.2% of annotated RD, and data collection is carried out through a systematic review of rare diseases using population-based studies, meta-analysis, and population surveys [Wakap *et al.* 2020]. The platform is supported by a relational database designed around the disease concept and maps the codes with other six terminologies, including ICD-10 [Orphanet: Orphadata 2022, Vasant *et al.* 2014]. Thus, we can navigate from different nomenclatures and establish a parallel among codes.

Regarding the focus of the system and the availability of cross-reference between ICD codes and the ORPHAcodes list, it became clear that the WHO terminology is not the best fit for RD definition. VASANT *et al.* (2014) shows that only around 500 RDs are listed in ICD-10, and half of them have generic codes, making it challenging for specialists to initiate trial treatments without additional information on symptoms and diagnosis [Vasant *et al.* 2014]. However, it is known that Brazil has 40 times more RDs [Aureliano and Gibbon 2020]. Despite this problem, the need to report ICD-10 to the Brazilian Ministry of Health remains a requirement, but the use of ORPHAcode provides a common language for a more uniform understanding of RD. Considering the context, the ORPHAcode nomenclature is a better option when dealing with RD, mainly due to its specificity [Weinreich *et al.* 2008]. Thus, the following section shows a visual and statistical model to demonstrate and quantify the nonlinear relationship between the ORPHAcode and ICD-10.

A.3 Proposed Method

In this section, the methodology of the proposal is explained, including the construction of the ICD10-ORPHA Brazilian Model according to the Ordinance n^{o}199 and the development of the web-app system implemented in this study.

A.3.1 ICD-10 - ORPHA Brazilian Model

The usage of ontology systems to link disease terminologies is not a novel concept. Usually, a Disease Ontology (DO) describes a set of diseases based on hierarchical characteristics of the condition, and it is associated with metadata, such as definition, symptoms, synonyms, and cross-references. It has been used in many studies to structure data [Schriml *et al.* 2012, Le and Dang 2016, Ambrosiano *et al.* 2020]. Furthermore, the ontology of a single disease can be associated with another one according to specific criteria defined by an expert in the domain [Schriml *et al.* 2022].

Moreover, researchers have noticed that several natural phenomena, such as social and biological arrangements, are composed of structures where elements are related to each other, given a specific link [Costa *et al.* 2011]. The methodology can cover a wide range of applications such as social dynamics and epidemic processes [Pastor-Satorras *et al.* 2015, Scabini *et al.* 2021].

Regardless of the structure of the data, when the information is not modeled as a complex network, the first step is to transform the raw data into a graph G, (G = {E,V}) , where V is the vertices (nodes), and E is the edges (links) connecting the elements in V. As mentioned, the Orphanet observatory has a cross-reference between ORPHA and ICD codes. Thus, the Brazilian model is created as follows:

$$CN_{ij} = \begin{cases} 1, & \text{if code i is related to code j} \\ 0, & \text{otherwise.} \end{cases} \tag{A.1}$$

It is essential to notice that an ORPHAcode will never directly link to another ORPHAcode due to the nature of the system. In addition, the network is undirected, which means that if node i is connected to node j, the contrary is also correct. Also, differing from ontology systems, in graph theory, one of the goals is also to compute statistical metrics to permit us to comprehend the graph quantitatively. The metrics computed in this study are defined as follows: degree, average degree, number of connected components, clustering coefficient and betweenness coefficient [Costa 2004]. There are a considerable number of metrics in graph theory but not all suit our study.

Next, we show how the proposed complex network model and the web-app system for its visualization were built. The algorithms were developed with Networkx, a package in Python for creating and manipulating complex networks. It also gives us access to different functions to compute each node's clustering coefficients, betweenness, and degree [Hagberg and Conway 2020]. Additionally, as seen in the next section, each connected component, i.e., subgraph where all pairs of nodes have finite path length, is decomposed as a single graph.

A.3.2 Interactive Web-app System

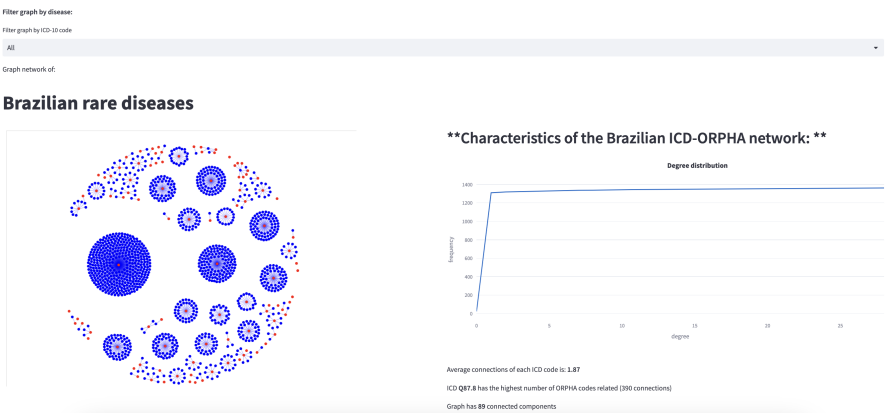
To enhance user interaction and facilitate the visualization of node distribution, a web-based application system was developed. First, the system features a select box that enables users to choose a RD based on an ICD-10 code list from Ordinance n^{o}199 of the Brazilian Ministry of Health [Saúde 2014]. Then, the graph component shows the selected disease, the number of related codes (node degree), the clustering coefficient, and the betweenness. To distinguish between ICD codes and ORPHAcodes identifiers, vertices from the international classification are colored in red, while the second terminology is painted in blue. The aim is to demonstrate the structure's complexity and build a visualization of the connections between the RDs in Brazil and Orphanet context, a commonly used terminology by geneticists and other clinical specialists.

The visualization is rendered by a graph library pyvis, an interactive network framework [Perrone, Unpingco and Lu 2020]. Users can zoom in, check the labels of each element in the network, and drag components around the canvas. The most basic metric in the image is the number of components.

A.4 Results and Discussion

Figure 41 provides an overview of the web-based system, which shows the correspondence between Rare Brazilian Diseases' ICD-10 codes from the Ministry of Health and Orphanet codes. As mentioned before, the Orphanet codes, or ORPHAcodes, are presented in blue, while the ICD-10 codes are in red. The web page also displays the cumulative degree distribution of the model, its the average degree (1.87), and the number of connected components (89). The plateau in degree distribution suggests that most diseases in the network have a low number of connections. However, since the average degree is not one, it is evident that many diseases do not have a one-to-one correspondence between ICD-10 and ORPHAcode terminologies. The link to the application is found online at https://tinyurl.com/bnv4zdma.

Figure 41 – The initial page of the interactive web-based system. On the left, Brazilian ICD-ORPHA Brazilian full network. On the right the cumulative distribution of the degree is shown.

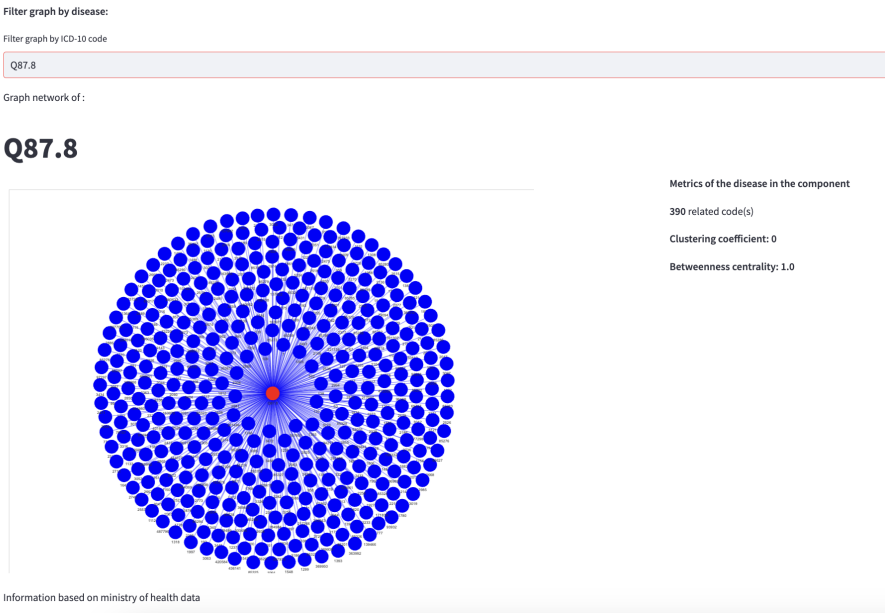


Source: Developed by the authors

In Figure 42, we can observe a section of the interactive system where users can select a specific disease from the dropdown menu at the top. Once a condition is selected, the corresponding subgraph is displayed, along with its degree, clustering coefficient, and betweenness centrality. The largest component, with central node in ICD-10 Q87.8 - Other specified congenital malformation syndromes, not elsewhere classified is shown in Figure 42. As we can see, Q87.8 has an extensive number of connections, with a total of 390 related ORPHAcodes. Two examples are ORPHAcode - 2669 - Nephrosis-deafness-urinary tract-digital malformations syndrome and ORPHAcode - 1270 - Bowen-Conradi syndrome [Le and Dang 2016]. Both diseases have different natures. While ORPHAcode 2669 ranges urinary tract anomalies, nephrosis, conductive deafness, and digital malformations, ORPHAcode 1270 is characterized by microcephaly, a dis-

tinctive facial appearance [Weinreich *et al.* 2008]. Showing that ICD-10 codes weakly represent RD.

Figure 42 – The largest component of the ICD-ORPHA network' page



Source: Developed by the authors

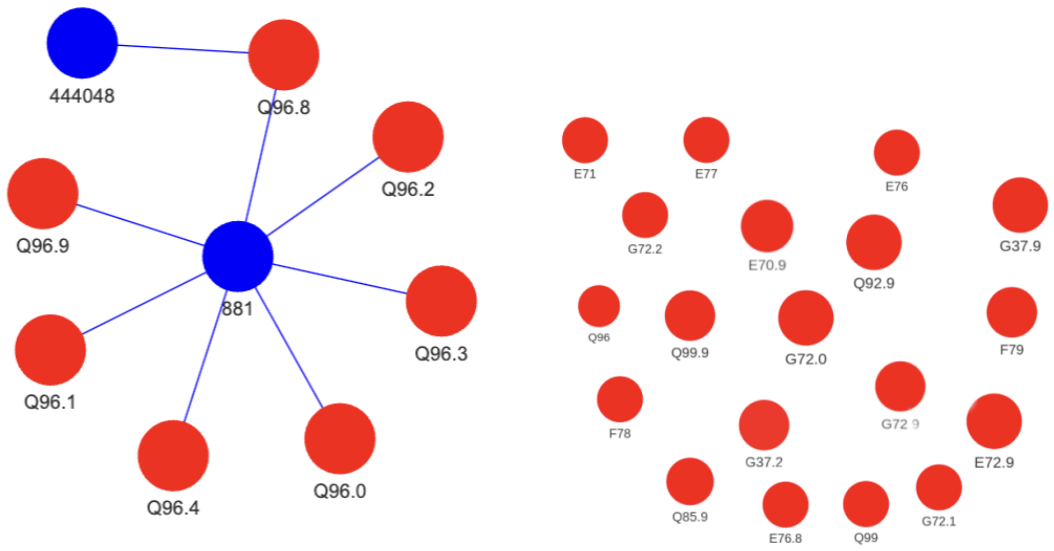
Furthermore, the graph metrics show that with a high betweenness, Q78.8 is in the path of any other two nodes in the subgraph. In contrast, a clustering coefficient of zero indicates a lack of connectivity between Q78.8 connections.

We have presented the largest component as an example, but many components possess the same star structure, where a central ICD node has numerous ORPHAcode connections in the Brazilian CN. For instance, Q87.0 (k = 120), Q87.1 (k = 67), G71.0 (k = 68), E77.8 (k = 50), G12.2 (k = 24) also have the same one-to-many relationship in the ICD-ORPHA model.

Secondly, another fascinating subgraph is represented by Figure 43a. In this subgraph, ORPHAcode 881, Turner syndrome, is associated with seven different ICD-10 codes (Q96.0, Q96.1, Q96.2, Q96.3, Q96.4, Q96.9, and Q96.8). This complex structure illustrates the possibility of a many-to-many relationship between ICD-10 and ORPHAcode identifiers, a disadvantage when using ICD-10 to represent RD. Furthermore, besides ORPHAcode 881, Q96.8 is associated with another ORPHAcode, the code 444048, which represents 46,XX ovarian dysgenesis-short stature syndrome. This aspect is the inverse of what we usually check in the entire graph, where the ICD-10 code is the component's main element (highest degree node) and this complexity shows the possibility of a many-to-many relationship between ICD-10 and ORPHAcode identifiers. The hybrid structure reflects the specialists' difficulty finding proper information, data and treatments to support patients. If there are several branches, the literature for a single ICD code can be imprecise for a specific condition.

Although most of the ICD-10 connects for one or more ORPHAcodes, 19 RDs in this

Figure 43 – (a) depicts the complexity of a single subgraph, highlighting a situation where one ICD-10 code is associated with two ORPHAcodes. On the other hand, (b) showcases instances where several ICD-10 codes from the Ministry of Health list have no corresponding ORPHAcode.



(a) (b)

Source: Developed by the authors

terminology does not match any classification in the Orphanet observatory. The third analysis case is shown in Figure 43b. In the Brazilian RD list, we found the following ICD-10 codes that match this characteristic: E72.9; G72.9; Q92.9; Q85.9; E70.9; G37.9; Q99.9; G72.0; G72.1; G72.2; F79; E76; E77; Q99; F78; Q96; E71; E76.8; and G37.2. According to specialists, G72.0 - Drug-induced myopathy, G72.1 - Alcoholic myopathy, and G72.2 - Myopathy due to other toxic agents are not considered rare. Also, ICD codes with the final .9 are non-specific diseases, meaning no definitive or well-defined diagnosis to the patient. At the same time, Orphanet provides greater specificity by having multiple branches for a single ICD-10 code [Weinreich *et al.* 2008]. To illustrate this, some examples of codes are E70.9 - Disorder of aromatic amino-acid metabolism, unspecified; G72.9 - Myopathy, unspecified; Q92.9 - Trisomy and partial trisomy of autosomes, unspecified; Q85.9 - Phakomatosis, unspecified; G37.9 - Demyelinating disease of central nervous system, unspecified; Q99.9 - Chromosomal abnormality, unspecified.

Furthermore, the categories: F79 (Unspecified intellectual disabilities); E76 (Disorders of glycosaminoglycan metabolism); E77 (Disorders of glycoprotein metabolism); Q99 (Other chromosome abnormalities, not elsewhere classified); F78 (Other intellectual disabilities); Q96 (Turner's syndrome); and E71 (Disorders of branched-chain amino-acid metabolism and fatty-acid metabolism) are groups of diseases which also do not define a conclusive diagnosis to the patient but appear in the RD list from the Ministry of Health. Finally, E76.8 (Other disorders of glucosaminoglycan metabolism) is another undetermined disease, while G37.2 (Central pontine myelinolysis) is not found in the Orphanet repository but remains in the Brazilian Rare Disease

list from the Ministry of Health.

A.5 Conclusion

The Orphanet, featuring a repository of over 6,100 RD codes, serves as a valuable resource for identifying and comprehending RDs. As Orphacode is tailored to the European context, it is essential to examine the approach employed by Brazil's Ministry of Health, which relies on ICD-10 for reporting. Our proposed web application offers visualization and the results of the complex network model created from ICD-10 and ORPHAcode terminologies in relation to Brazilian RDs. As illustrated in Figures 41, 42, and 43, devising a flawless terminology for RD classification remains challenging, particularly for developing countries. More specifically, the entire network consists of 89 components of various shapes. Generally, a single ICD-10 code corresponds to multiple ORPHAcodes, forming a star-shaped component. However, we demonstrate the opposite, where one ORPHAcode is linked to multiple ICD-10 codes. Additionally, this relationship is not star-shaped due to another connection present in element Q96.8.

These graph evaluations and complexities assist specialists in visualizing and quantifying the challenges in determining a definitive diagnosis for RDs, given the numerous generic codes found in the Brazilian ICD-10 rare disease classification. Furthermore, this highlights the need for ongoing research efforts to examine patient cases, clarify and encode unidentified conditions, improve the documentation of RD, determine their epidemiology, and strengthen national registries.

Although the ICD-11 has been released, the Brazilian Ministry of Health is expected to update its RD list in the coming years. This updated list will undoubtedly simplify cross-referencing and streamline the search for treatments and diagnosis, thanks to the improved terminology. Nevertheless, the majority of literature and medical facilities still utilize and report conditions based on ICD-10. This complexity must be documented and made accessible to Brazilian specialists to enable them to search for relationships within their national domain. Therefore, this technical contribution can be used to minimize data complexity in RDs international classifications.

In future work, we plan to assess the new ICD-11 codes for the Brazilian RD list and enrich the node with essential metadata, ultimately constructing a more robust ontology system that could serve as a foundation for disease diagnosis.

Acknowledgements

This study was funded by the National Council for Scientific and Technological Development – CNPq and the Ministry of Health of Brazil – MoH. The Raras Network group is

composed of 40 institutions in the five regions of Brazil. More information regarding them and institutions can be found at raras.org.br

ANNEX

B

SOCIAL INTERACTION LAYERS IN COMPLEX NETWORKS FOR THE DYNAMICAL EPIDEMIC MODELING OF COVID-19 IN BRAZIL

Initial Consideration

This Annex is a summarized version of the paper published at journal Physica A. The work presents a specific **network model** designed to study the spread of COVID-19 and social interactions. In this model, individuals are represented as nodes, and social contacts are represented as links. The network consists of different layers, each representing a specific type of social interaction. The simulation takes into account the lockdown or absence of lockdown measures in these layers. The paper applies this network model to study the COVID-19 epidemic in Brazil as a case study. By simulating the spread of the virus within the network, the model provides insights into the impact of different social interactions on the epidemic's dynamics. The findings of this study contribute to the field of epidemiology and provide important insights for policymakers and public health officials in managing and controlling the spread of the virus.

B.1 Introduction

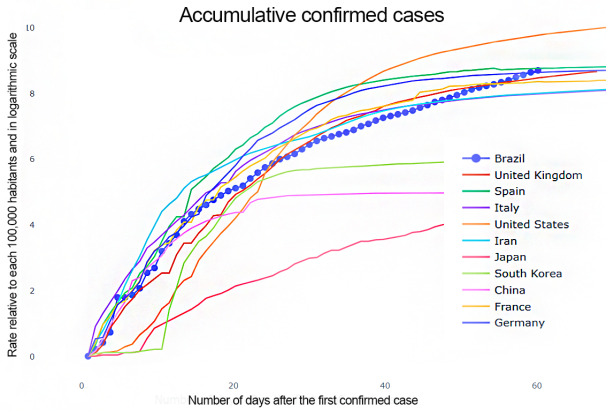
Although we have experienced several pandemics throughout history, COVID-19 is the first major pandemic in the Modern Era. The 21st century is marked by globalization and an intricate and intense social network, which connects in one way or another to everyone on the planet.

The form of propagation and contagion of the Sars-CoV-2 virus occurs by direct contact

between individuals, through secretions, saliva, and especially by droplets expelled during breathing, speeching, coughing, or sneezing. The virus also spreads by indirect contact, when such secretions reach surfaces, food, and objects [WHO 2020, visited on 2020-05-07]. Besides, infected people take a few days to manifest symptoms, which can be severe or as mild as a simple cold. There is even a large proportion of infected people who remain asymptomatic [Kim *et al.* 2020]. This makes it practically impossible to quickly identify the infected and apply effective measures to limit the spread of the disease. Also, Sars-CoV-2 was discovered in December 2019, which makes it very recently in the face of the current epidemic. Little is known about the COVID-19 disease, which appears to be highly lethal, with no drugs to prevent or treat. The concern is greater since direct (individual - individual) and indirect (individual - objects - individual) social relations are the means of spreading the disease. Thus, the social interaction structure is the key to create strategies and guide health organizations and governments to take appropriate actions to combat the disease.

One of the main concerns is overloading the health system. According to [Castro *et al.* 2020], there are only 9 hospital beds per 100,000 people in the North region while Southeast accounts for 21 hospital beds. The treatment of severe cases requires the use of respirators/ventilation in intensive care units (ICU), and if simultaneous infections occur there will be no beds to meet the demand and a possibly large number of victims. Thus, it is urgent to develop models and analyses to try to predict the evolution of the virus. Also, as noted in Figure 44, Brazil is running towards being the next epicenter of the pandemic. It has already exceeded the number of cases in important countries such as Germany, China, Japan, Italy, Iran, South Korea, and France (the rates consider the population size of each country and are on a logarithmic scale).

Figure 44 – Total number of cases reported in Brazil compared to other countries (May 5, 2020 [Johns Hopkins University 2020, visited on 2020-05-18]). It is possible to notice that Brazil is surpassing countries such as Italy, South Korea, Japan, and China, and it is reaching the relative number of cases in the United Kingdom and France. As of the date of this study, the United States is the epicenter of the pandemic.



Source: Developed by the author

Since COVID-19 presents a unique and unprecedented situation, this work proposes a specific model for the current pandemic. Based on the classic epidemic model SIR, also extended

to SID [Qi *et al.* 2020], SIASD [Bastos and Cajueiro 2020] and SIQR [Crokidakis 2020], we propose a more realistic model to better represent the effects of the COVID-19 disease by adding more infection states. The proposed approach also considers social structures and demographic data for complex network modeling. Each individual is represented as a node and edges represent social interaction between them. The multi-layer structure is implemented by different edges representing specific social activities: home, work, transports, schools, religious activities, and random contacts. The probability of contagion is composed of a dynamic term, which depends on the circumstances of the social activity considered, and a global scaling factor β for controlling characteristics such as isolation, preventive measures, and social distancing.

The proposed model can be used to analyze any society given sufficient demographic data, such as medium/big cities, countries, or regions. Here we analyze in depth the Brazilian data. The SIR model is applied through the network using an agent model, and each iteration of the system is simulated using the 24-hour pattern, allowing us to understand the dynamics of the disease throughout the days. The results show the importance of social distancing recommendations to flatten the curve of infected people over time. This is currently maybe the only way to avoid a collapse of the health system in the country.

B.2 Epidemic Propagation on Complex Networks

Created from a mixture of graph theory, physics, and statistics, CNs are capable to analyze not only the elements themselves but also their environment to find patterns and obtain information about the dynamics of a system.

Usually, applications with complex networks consist of two main steps: i) transform the real structure into a complex network, and ii) analyze the model and extract its features or understand its dynamics. One natural phenomenon that has a straight forward connection to a complex network in society. People are connected due to several aspects such as members of a family, religious groups, co-workers, members of the same school, or faculty, among other social relationships. Therefore CNs have been widely employed for social network analysis [Vega-Redondo 2007].

Extended from social interactions, the epidemic spread has also been studied by researchers in the last decades. In this context, one of the best known and widely used epidemic models in infectious diseases is the susceptible-infected-recovered (SIR) model, which is composed of three categories of individuals [Bailey *et al.* 1975, Anderson, Anderson and May 1992].

- **Susceptible:** the ones who are not infected but could change its status to a state to infected if in contact with a sick person combined with a probability β of contagion
- **Infected:** the ones that have the disease

- **Recovered:** usually after some time, a person recovers from the illness and it is not able to be infected again due to the immunity process (in this case, this is an assumption of the process). The recovery rate of infected people is aligned with a probability of γ

Also, the model can be described as

$$\frac{ds}{dt} = -\beta is, \ \frac{di}{dt} = \beta is - \gamma i, \ \frac{dr}{dt} = -\gamma i \tag{B.1}$$

where s, i and r represents the ratio of susceptible, infected and recovered people in the population, respectively. Usually, the problem is solved with differential equations, however, agent-based techniques in networks can represent the nature of the spread of viral diseases in a more complex scenario.

If a network is fully connected, meaning that $e(v_i, v_j) = \{1, \forall_{i,j} 0 < i, j <= N\}$, Equation B.2 fits the structure perfectly. However, in the real world, not everyone is connected and people only contract the disease if in contact with an infected individual or object. This is why a complex network approximates the dynamics of real viruses and can help us to understand the disease behavior. There are various approaches to represent people and society as networks, named social network analysis. Small world networks [Moore and Newman 2000] can be used as a good approximation of the social connections. In 2000, Moore [Moore and Newman 2000] emphasized that the use of small-world networks, where the distance among two elements is usually small in comparison to the size of the population, showed a faster spread of the viral disease than classical diffusion methods. The approximation of real social phenomena was first explained by Milgram [Milgram 1967] in [Milgram 1967], the sociologist is the author of the well-known idea that there are up to six people separating any two individuals in the world, which reinforces the importance of analyzing the epidemic spread from a graph view. In [Newman and Watts 1999], the authors used small-world networks to simulate a SIR model, however, they considered that every contact with an infected person resulted in contamination, which is not realistic. Therefore, other researchers improved the model over the years, adding new constraints to approximate the simulation to real scenarios [Silva *et al.* 2019].

The SIR model on networks works as follows: each node represents a person and, the elements are connected according to some criteria and the epidemic propagation happens through an agent-based approach. It starts from a random node, and for each time step nodes with the susceptible state can contract the disease from a linked infected node with a predefined probability. The same idea occurs with the recovered category. After a certain period, a node can recover or can be removed from the system (case of death) according to a certain probability. At the end of the evolution of a SIR model applied to a network, the number of nodes in each SIR category (susceptible, infected and recovered) can be calculated for each unit of time evaluated and then compare these data with real information, for example, the hospital capabilities of

the health system. Also, the probability of infection and recovery can be adjusted over time considering social distancing, hygiene, and health conditions.

B.3 Proposed Model: COmplexVID-19

The proposed model extends the SIR model to a more realistic scenario to achieve a better correlation to the COVID-19 disease, since the model was created specifically for the disease, we named the model as COmplexVID-19. Our strategy is based on a multi-layer network to represent the Brazilian demography and its different characteristics of social relationships. Each layer is composed of a set of groups representing how people interact in a given social context. In the network, a node represents a person and the edges are the social relationships between persons, and they are also the means through which the disease can be transmitted. The virus spreads from an infected node to neighboring nodes at each iteration step (1 step = 1 day), according to a given infection probability. First, we describe how the layers are built based on social data from Brazil.

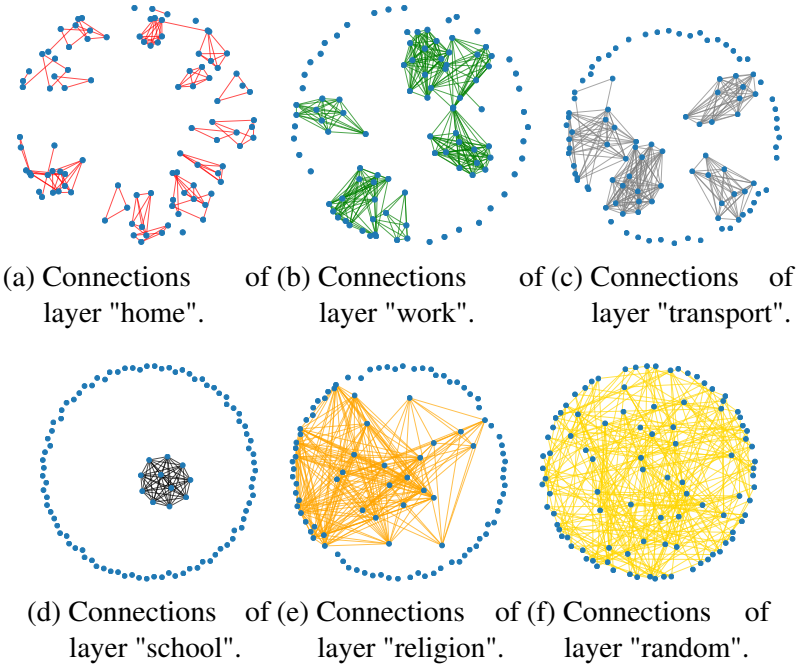
B.3.1 Network Layer Structure Over Brazilian Demography

To define the different social relations, the first information needed is the age distribution so that groups such as schools and work can be separated. We consider the Brazilian age distribution in relation to the total population in 2019 [IBGE 2019, visited on 2020-05-11], details are given on [Scabini *et al.* 2021]. This distribution is used to define an age group for each node, which is then used to determine its social activities through the creation of edges on different layers. In this approach, each network-layer represents a kind of social relationship or activity that influences the transmission of the COVID-19. In this way, it is possible to evaluate and understand what is the impact of each social activity in the epidemic propagation. Basically, in this work, a network layer is represented by a set of edges connecting some nodes. The following social activities are considered, composing 6 different layers: home, work, transport, school, religious activities and random activities.

The first layer represents home interactions and is composed of a set of groups with varying size which are fully connected internally. These groups have no external connections, i.e. the network starts with disconnected components representing each family. To create each group, we consider the Brazilian family size distribution for 2010 [IBGE 2010, visited on 2020-04-20], the year with more detailed information on family sizes from 1 up to 14 members. We consider the probability of a family having sizes from 1 to 10, therefore the probability of a family having 10 persons is the sum of the higher sizes. The first layer is then created following the family size distribution and ensuring that each family has at least 1 adult. Figure 45 (a) shows the structure of such a layer built for a population of $n = 100$.

A large fraction of the population in any country needs to work or practice some kind of

Figure 45 – Each social layer of the proposed multi-layer network. The nodes are people and do not change across layers, and the weighted connections represent social contact which may lead to infection according to the edge weight (probability value between $[0, 1)$.



(a) Connections of layer "home". (b) Connections of layer "work". (c) Connections of layer "transport".

(d) Connections of layer "school". (e) Connections of layer "religion". (f) Connections of layer "random".

Source: Developed by the author

economic activity, which also means interacting with other people. Thus, work represents one of the most important factors of social relations, which is also very important in an epidemic scenario. To represent the work activity we propose a generic layer to connect people with ages from 18 to 59 years, i.e. 60% of the total population in the case of Brazil. There is a wide variety of jobs and companies, therefore it is not trivial to create a connection rule that precisely reflects the real world. Here, we consider an average scenario with random groups of sizes around $[5, 30]$, uniformly distributed, and internally connected (such as the "home" layer). An example of this layer is shown on Figure 45 (b), using $n = 100$. Although the nodes of a group are fully connected, the transmission of the virus depends directly on the edge weights.

Collective transports are essential in most cities, however, it is one of the most crowded environments and plays an important role in an epidemic scenario also due to the possibility of geographical spread, as vehicles are constantly moving around. The third layer we propose represents this kind of transports, such as public transports, and includes people that do not possess or use a personal vehicle. In Brazil the number of people using public transport depends on the size of the city, with 64.98% in the capitals and 35.89% in other cities [IPEA 2011, visited on 2020-04-15], with an average use of around 1.2 hours a day [1]. Here we consider the average of the population between the two cases (50%), randomly sampled, to participate in the "transports" layer. Random groups are created with sizes between $[10, 40]$, uniformly sampled, and the nodes

[1] <http://g1.globo.com/bom-dia-brasil/noticia/2015/02/brasileiros-gastam-em-media-1h20-por-dia-em-transportes-publicos
html>

within each group are fully connected. This variation of sizes is considered to represent cases such as low and high commuting times, and also the differences between vehicle sizes. Other factors such as agglomeration and contact intensity are discussed in Section B.3.1.1. This layer is illustrated on Figure 45 (c).

Schools are another environment of great risk for epidemic propagation. The proposed layer considers the characteristics of schools from primary to high school and how children interact. We consider that all persons from 0 to 17 years (24% of the Brazilian population) participate in this layer, and the size of the groups, which represents different school classes, varies uniformly between $[16, 30]$ [INEP 2018, visited on 2020-04-5]. This layer is illustrated on Figure 45 (d).

Brazil is a very religious country, in which by 2010 only around 16.2% of the population claimed not to belong to any religion [IBGE 2010, visited on 2020-04-18]. 64.6% claimed to be catholic and 22.2% to be protestant, summing up to 86.8% of the total population. Here we consider that nearly half of these people (40% of the total population) actively participate in religious activities (weekly). The distribution of religious temple sizes is defined as a Pareto distribution in the interval $[10, 100]$. Taking into account that wage distribution follows the Pareto distribution approximately, we model real estate predominance according to their capacity. The assumption here is that building costs (for churches, offices, homes, etc.) have a linear relationship to their internal capacity, and thus any given capacity has a power-law relationship with the number of such buildings within a region.

We consider a random layer to represent all kinds of contacts not related to the specific previous social layers. This includes small direct contacts (person-to-person) and indirect contacts (individual - objects - individual) that may happen throughout the week, such as random friend/neighbor meetings, shopping, and other activities that involve surface contacts. For that $5n$ new random edges are created, that can connect any node. On the one hand, this yields an average of 5 random connections to each node, which can randomly connect any other node. On the other hand, the impact of this layer on the epidemic is smaller than the others, as it represents rapid contacts in comparison to the other activities described, thus its infection probability is smaller. In the following section we discuss the details concerning this aspect, deriving from the edge weights of each layer. In Figure 45 (f) an example of this layer is shown. The overall structure of social interactions in our model can be compared to the statistical analysis in [Ferguson *et al.* 2020], however here we introduce a more detailed model of social contacts with specific layers and connection patterns to better fit the particularities of a given country or city.

B.3.1.1 Infection Probabilities

Unlike the traditional SIR model, which consists of a single β term to describe the probability of infection, here we propose a dynamic strategy to better represent the real world and the new COVID-19 disease. The idea is to incorporate important characteristics in the context

of epidemic propagation according to each layer. Firstly, to a given layer a fixed probability term is calculated to represent its characteristic of social interaction. For this, we considered 3 local terms: the contact time per week, the average number of people close to each other (agglomeration level), and the total number of people involved in the respective activity. Considering two nodes v_x and v_y, connected at group j of layer i, its edge weight is then defined by

$$e(v_x, v_y) = \frac{t_i}{168}\frac{k_i}{n_{ij}}\beta \tag{B.2}$$

where t_i represents the average weekly contact time on layer i, k_i is the agglomeration level (average number of nearby people) and n_{ij} represents the size of the group j in which the nodes participates on layer i. The first fraction represents the contact time normalized by the total time of the week ($24 * 7 = 168$), and the second fraction represents the proportion among the local people closest to the total number of people on that activity group.

The first part of the infection probability equation is multiplied by a β term, which scales the original probability. The β term is then the only parameter to tune the infection rates for the entire network, and the other properties are specific for the studied society, based on its population characteristics and the nature of the activities (layers). Table 15 shows these specific properties that we considered for the Brazilian population, and how the infection probabilities are calculated for each layer. In the table, we have the following information: who or how many people are part of the activity represented by a layer (column "who", discussed in the previous section); contact time according to activity (column "Time of contact"); the average number of people close to each other in each activity (column "Nearest", represents the agglomeration level); the number of connections between people (column "Group size"); the probability of infection (column "Probability").

Table 15 – Specific brazilian properties considered to compose each social layer and calculate their probability of infection, i.e. the edge weights of each layer.

Activity	**Who**	**Time (t_i)**	**Nearest (k_i)**	**Group size ($n_{i,j}$)**	**Infection Prob. (e)**
Home	everyone	3 hours a day	$k_i = n_{i,j}$	$[1,10]$	$(\frac{21}{168}\frac{n_{i,j}}{n_{i,j}})\beta$
Work	18 to 59 years	8 hours a day, 5 days	3	$[5,30]$, uniform	$(\frac{40}{168}\frac{3}{n_{i,j}})\beta$
Transports	50%, random	1.2 hours a day	8	$[10,40]$, uniform	$(\frac{8.4}{168}\frac{8}{n_{i,j}})\beta$
Schools	0 to 17 years	4 hours a day, 5 days	5	$[16,30]$, uniform	$(\frac{20}{168}\frac{5}{n_{i,j}})\beta$
Religion	40%, random	2 hours a week	6	$[10,100]$, Pareto	$(\frac{2}{168}\frac{6}{n_{i,j}})\beta$
Random	5 per person	1 hour a week	1	1-to-1 contacts	$(\frac{1}{168})\beta$

Source: Developed by the author

B.3.2 Dynamics Modeling

The proposed model is a variant of the SIR approach where we include new possible states, structural and dynamic mechanisms after the new findings on COVID-19. The traditional SIR model consists of 3 states: Susceptible, Infected, and Recovered. To better represent the

intrinsic dynamics of the new epidemic, we considered 7 states according to reported distributions of the clinical spectrum [Lauer *et al.* 2020, Surveillances 2020]:

- **Susceptible**: Traditional case, it means that a person can be infected at any time. This is the initial state of every node.
- **Infected - asymptomatic**: People who do not show any symptoms (30% of the total cases of infection) and remain contagious for up to 18 days (they may recover after 8 days). This is the most dangerous case for the epidemic spreading because the person is not aware of its infection.
- **Infected - Mild**: 55% of the cases, present mild and moderated symptoms with no need for hospitalization, remain contagious for up to 20 days, and may recover after 10 days of infection.
- **Infected - Severe**: 10% of the cases, present strong symptoms, and need hospitalization, remain contagious for up to 25 days. Has a death rate of 15% and may recover after 20 days.
- **Infected - Critical**: Present worst symptoms and remain contagious for up to 25 days, need ICU and Ventilation, have a death rate of 50% and may recover after 21 days.
- **Recovered**: People who went through one of the infection cases and overcame the disease, ceasing to contaminate and supposedly becoming immune. These nodes no longer interact with other nodes anymore and are therefore removed from the network.
- **Dead**: People who went through severe or critical cases and eventually died. These nodes are also removed from the network.

Estimates for the proportion of asymptomatic cases vary from 18% (95% confidence, $[15.5, 20.2\%]$) [Mizumoto *et al.* 2020] to 34% (95% confidence, $[8.3, 58.3\%]$) [Huang *et al.* 2020]. Considering the confidence intervals, here we roughly approximate it to an average of 30% of the total number of infected cases. However, it is very difficult to study asymptomatic cases due to several reasons, such as the lack of available tests and the difficulty in identifying potential cases, which would include every person who had contact with known symptomatic cases. Some studies indicate that asymptomatic cases may remain contagious for up to 25 days, with an incubation period of 19 days [Bai *et al.* 2020], but the viral load may be smaller at the end of the infection. Here we take an optimistic approach considering that they may recover (become immune and cease to contaminate) uniformly after 8 days of infection, up to around 18 days. As for the recovered nodes, we are considering that people become immune or at least acquire a long-term resistance to the virus, up to a maximum of 300 days (limit of our simulations). However, this should be taken cautiously as these properties are not yet fully understood [Lan *et al.* 2020].

B.3.2.1 Dynamic Evolution

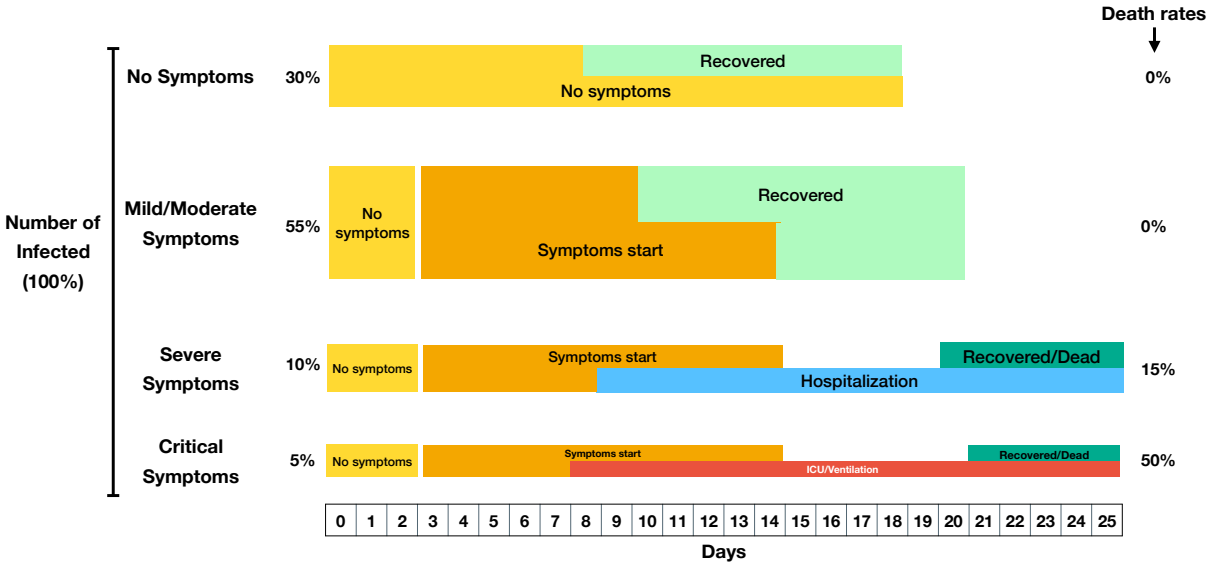
The infection grows through the contact (edges) between infected and susceptible nodes, and the probability of being infected is the edge weight. If infection occurs, then one of the 4 infection cases are chosen based on the probability described above (30%, 55%, 10% and 5%). This distribution plays an important role in the structure and dynamics of the network. The node structure of asymptomatic cases does not change during the simulation, except for the time it takes to cease contamination and recover. It means that as these persons are not aware of their contamination, they will remain acting normally on the network (according to the active layers and edge weights). Their contagious time varies from 1 to 18 days after infection.

Concerning the other cases (mild, severe, and critical), we consider the incubation time of the virus, the recovery time, the contagion time, the death rates of each case, and the usual action taken by the infected person or health professionals at hospitals. Various works [Linton *et al.* 2020, Backer, Klinkenberg and Wallinga 2020, Lauer *et al.* 2020] point out that the average incubation period of COVID-19 is around 5 days, but some cases may take much less or more time. The official WHO report [WHO 2020] states that the average incubation time is around 5 to 6 days, with cases up to 14 days. The results in [Lauer *et al.* 2020] show that the average shape of the incubation time follows a log-normal distribution (Weibull distribution) with an average of 6.4 days and a standard deviation of 2.3 days. In this context, we consider the day when an infected person begins to show symptoms by randomly sampling from this distribution (1000 repetitions), with cases varying from 2 to 14 days.

For mild cases, the nodes are isolated at home, maintaining the connections of the first layer, and then only 20% of the cases are diagnosed. Considering the ratio of diagnosed cases, patients who are asymptomatic or with mild symptoms of COVID-19 may not seek health care, which leads to the underestimation of the burden of COVID-19 [Lai *et al.* 2020]. Moreover, our diagnosis rule is also based on the fact that ongoing tests in Brazil are increasing more slowly than in most European countries and the USA (tests are being performed mostly on people that need hospitalization). If a given case is severe or critical, the patient goes to a hospital and is fully isolated, i.e. we remove all of its connections. This is a rather optimistic assumption, considering that these patients still may infect the hospital staff. Concerning the time that patients usually stay at hospitalization/ICU, the works [Bhatraju *et al.* 2020, Zhou *et al.* 2020] points to an average of 14 days for all cases. For standard hospitalization, we considered a minimum of 6 days and a maximum of 16 days of stay, and for the ICU/Ventilation, a minimum of 7 and a maximum of 17 days of stay. The time of each case will depend on the day the symptoms start and the day of recovering/death. Figure 46 illustrates all the infected states and mechanisms described here. This configuration results in an overall lethality of 4%. It is important to stress that here we consider a maximum of 25 days of infection time, which is the time frame based on most studies we have seen so far in the literature. We are still at the beginning of the pandemic and a better characterization of the long-term impact is very difficult. Nonetheless, the available information

allows to represent the most obvious features of the Sars-CoV-2 virus and to evaluate its main impacts on society.

Figure 46 – Configuration considered for the dynamic evolution of each type of infected node in the proposed SIR model. Each overlapping region is treated as a combined probability distribution that defines when one phase ends for the other to begin.



Source: Developed by the author

To simulate the reduction or increase of social distancing/quarantine, we remove/include some layers of the network, or change their edge weights. Similarly to the approach on [Ferguson *et al.* 2020] to improve home contact when in quarantine, we increase the home layer edge weights by 20% for each removed layer. To balance that we considered a smaller number of hours of contact in the base calculation for the home layer (3 hours a day), also taking into consideration that this layer has full contact between people of the same family. When the home contacts are increased according to our approach of layer removal, the time/intensity of contacts may increase up to its double.

B.4 Results

For each experiment with the proposed model, we consider the average and standard deviation (error) of 100 random repetitions to extract statistics of infection, death, and hospitalization time. Due to the random nature of these networks, it is possible that extreme cases occur within the repetitions, i.e. when the infection starts at a node that is not capable of further propagation, leading the epidemic to end at few iterations. Considering the real data we know that this is not the case, at least not for Brazil, therefore we manually remove these networks and they are not considered for the average/error calculations. It is important to notice, however, that this rarely happens, in all our experiments we noticed a maximum of 4 networks of this kind. Due to time and hardware constraints, our simulation considers 100,000 nodes, and the results need to

be scaled up by a factor of 57 to match the Brazilian population statistics. This factor was empirically found by approximating the model results in the number of reported cases in Brazil. It is important to stress that for better statistics it should be considered the largest possible number of nodes to represent a population, i.e. the ideal case would be $n = total\ country/city\ population$. However, the computational cost of the simulation grows directly proportional to the number of nodes and edges of the network, and considering the critical situation of the moment at hand, 100,000 nodes are our limit to promptly present results of the epidemic dynamics.

In the experiments when varying the social distancing, the same network is considered in each iteration, i.e. comparisons of including/excluding layers are made in the same random network. We considered the epidemic began on February 26, which is the day the first confirmed case was officially reported. It is important to emphasize that we made various optimistic assumptions throughout the model construction and simulation, such as to consider that people are behaving with more caution by reducing direct contact, wearing masks, and doing proper home/hospital isolation when infected. It is also important to notice that we are not considering the number of available ICU/regular hospitalization beds for the death count, i.e. all the critical and severe cases are effectively treated. It is not trivial to estimate the direct impact of these numbers on the epidemic, however, this is an essential factor that directly impacts the number of deaths. Here we focus on the impacts of different actions on the overall epidemic picture, such as the increase and reduction of cases, deaths, and occupied beds in hospitals.

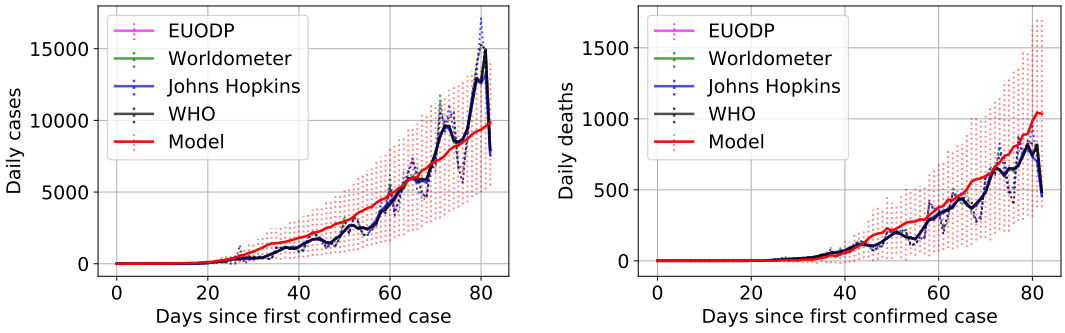
The social network starts normally, with all its layers and the original infection probabilities. The infection starts at a node with the closest degree to the average network degree and propagates at iterations of 1 day (up to 300 days). We consider an optimistic scenario, in which people are aware of the virus since the beginning, thus the initial infection probability is $\beta = 0.3$. This represents a natural social distancing, a reduction of direct contacts that could cause infection (hugs, kisses, and handshakes), and also precautions when sneezing, coughing, etc. We empirically found that this initial value of β yields results with a higher correlation to the Brazilian pandemic. A moderated quarantine is applied after 27 days, representing the isolation measures applied on March 24 by most Brazilian states, such as São Paulo [DECRETO Nº 64.881 visited on 2020-04-30]. To simulate this quarantine we remove the layers of religious activities and schools and reduce the contacts on transports and work down to 30% of its initial value, i.e. $\beta = 0.09$. The remaining activities on these layers represent services that could not be stopped, such as essential services, activities that are kept taking higher precautionary measures, and also those who disrespect the quarantine.

B.4.1 Comparison to Real Data

We compare the output of the model in the first 83 days with real data available from the Brazilian epidemic (up to May 18) [Portal 2020, visited on 2020-05-18, Worldometer 2020, visited on 2020-05-18, Johns Hopkins University 2020, visited on 2020-05-18, WHO 2020,

visited on 2020-05-18]. The model achieves a significant overall similarity within its standard deviation. The greatest difference in the number of diagnosed cases at the last 10 days may be related to the increase in the number of tests being performed in Brazil, or yet, the constant decrease of isolation levels in the country (below 50% for most days of the past month) [Inloco 2020, visited on 2020-05-15]. We considered here a fixed isolation level around what was observed in the first days after the government decrees in Brazil, but data in ref. shows that these levels are constantly changing. Therefore, the number of diagnosed cases and deaths for the remaining simulation may be greater than the reported on this paper (see the "keep isolation" scenario in the next section).

Figure 47 – Comparison between the proposed model output for the first 83 days (up to May 18) to 4 different data sources of the Brazilian COVID-19 numbers: EU Open Data Portal (EUODP) [Portal 2020, visited on 2020-05-18], Worldometer [Worldometer 2020, visited on 2020-05-18], Johns Hopkins University [Johns Hopkins University 2020, visited on 2020-05-18], and World Health Organization (WHO) [WHO 2020, visited on 2020-05-18]. The dotted lines represent the standard deviation, in the case of the real data the curve is the average over a 5-day window, and the solid lines the real raw data. The greatest average number of deaths produced by the proposed model may be related to underdetection (See Figure 48).



(a) Daily new cases.

(b) Daily new deaths.

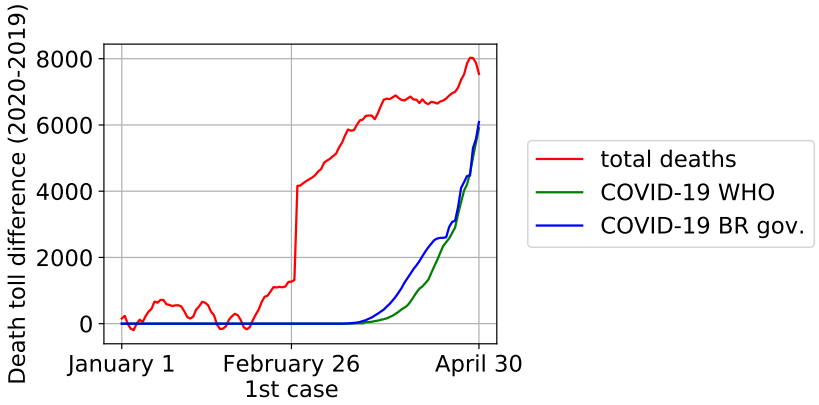
Source: Developed by the author

Concerning the daily death toll, the average number of the proposed model is greater than the official numbers. This is somehow expected, considering that the underdetection rates may be greater in contrast to the fewer number of tests being performed. To better understand this, we analyzed the number of death in Brazil from January 1 to April 30, comparing cases between 2019 and 2020, the results are shown in Figure 48. It is possible to observe a clear increasing pattern after February 26, which is the day of the first officially confirmed case of COVID-19 in Brazil. This indicates that the real death toll for the disease may be significantly greater than the official numbers.

B.4.2 Future Actions and its Impacts

After the initial epidemic phase, we consider 4 possible actions that can be taken after 90 days (May 26): a) Do nothing more, maintaining the current isolation levels; b) Stop isolation,

Figure 48 – To understand the impact of the COVID-19 underdetection in Brazil, we considered the official death records of 2019 and 2020 at the same period (January 1 to April 30) [Portal da Transparência - Painel COVID Registral visited on 2020-05-13]. Then the total death difference is compared to the COVID-19 records of the WHO [WHO 2020, visited on 2020-05-18] and the Brazilian government [Portal da Transparência - Painel COVID Registral visited on 2020-05-13] data. The largest difference that appears right after the first confirmed case may indicate a significant underdetection of COVID-19 cases.

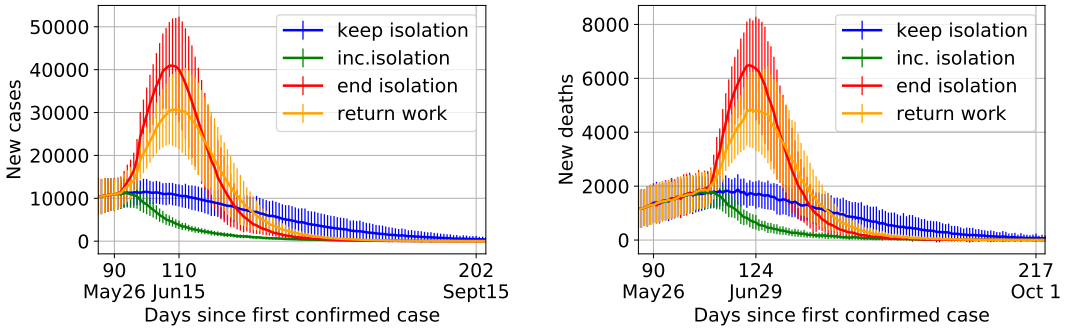


Source: Developed by the author

returning activities to normal (initial network layers and weights); c) Return only work activities, restoring the initial probability of the layer; or d) Increase isolation, stopping the remaining activities in the work and transports layers (home and random remains). Firstly, we analyze the impacts on the number of daily new cases and deaths, results are shown in Figure 49. As previously mentioned, at the start of the COVID-19 pandemic, Brazil was performing a fewer number of tests by an order of magnitude, in comparison to other countries with similar epidemic numbers, therefore we considered as diagnosed only the severe and critical cases, which are pronounced subjects for testing, and 20% of the mild cases. The total infection ratio is discussed later. Considering keeping the current isolation levels, the peak of daily new cases occurs around 100 days after the first case (June 5), with around 11,000 confirmed cases. After 202 days (September 15), the average daily cases is around 500, and it goes below 100 daily cases after around 237 days (October 19). The peak of daily new deaths occurs around 118 days (June 23), with an average of 1900 deaths, and goes below 100 new occurrences after around 210 days (September 24). It is important to stress that this is a hypothetical scenario where the isolation level remains the same from day 27 to 300, which is hardly true in the real world where it is constantly changing [Inloco 2020, visited on 2020-05-15]. The total numbers after the last day (300) account for 946,830 ($\pm 10,507$) diagnosed cases and 149,438 ($\pm 3,124$) deaths.

When we consider the return of all activities after 90 days, the number of cases and deaths grows significantly in an exponential fashion. The peak occurs at 108 days (June 13) with an average of 40,937 ($\pm$ 11,010) new cases, and at 122 days (June 27) with an average of 6,484 ($\pm$ 1,739) new deaths. Although the peak of cases/deaths and the decrease of the numbers occur early, in this case, the final result is critically worse, with a total of 1,340,367 ($\pm$ 18,513)

Figure 49 – Daily statistics in 4 possible scenarios after 90 days (May 26): Keep isolation levels; Increase isolation (stop work and public transports); End isolation (returns work and transport to normal and return school and religion); and return work (only the work layer is returned to normal).



(a) Daily new cases (diagnosed)

(b) Daily new deaths

Source: Developed by the author

diagnosed cases and 212,105 (± 4,359) deaths. Here it is important to notice that we considered that all the activities return after 90 days and remain fully operational until the last day (300). Moreover, we do not account for the overloading of hospitals, which directly impacts the final death count. Therefore, the number of deaths may be considerably higher. Another possible scenario is the return of only the work layer, keeping reduced transports and no schools and religious activities, however, the pattern is similar to returning all activities, considering the growth time, peak, and decay time. The final numbers in this case are 1,253,119 (± 26,009) diagnosed cases and 197,756 (± 5,693) deaths.

If the isolation is strictly increased after 90 days (lockdown), the infection and death counts drop significantly in comparison to the other approaches. Moreover, the recovering time is much faster, as daily new cases stop earlier than the other scenarios. The peak of daily new cases happens around day 93 (June 1), and of daily new deaths around day 106 (June 11). The total numbers of diagnosed cases and deaths after day 300 are, respectively, 552,855 (± 195,802) and 87,059 (± 30,871).

Considering the hospitalization time described in the scheme of Figure 46 it is possible to estimate the number of occupied beds for regular hospitalization (severe cases) and ICU/Ventilation (critical cases). We also show the difference between the cumulative growth of diagnosed and undiagnosed cases and recovered cases. The same approach as the previous experiment is considered (except for "return work") with 3 possible actions after 90 days (May 26), results are shown in Figure 50. The overall pattern of results is similar to the previously observed for the number of diagnosed cases and deaths. It is possible to notice that the number of undiagnosed cases is much higher than the diagnosed cases. This reflects the number of asymptomatic cases and the lack of tests for mild cases. In the worst scenario, which means ending the isolation, the total infected number may go above 5 million cases. The recovered rate is directly proportional

to the infected rate, as one needs to be infected to either die or become resistant to the disease. If the infected rate is high, so is the recovered rate, e.g. the scenarios of keeping or ending isolation, and a high recovered rate also helps in mitigating the epidemic propagation (natural immunization). However, increasing isolation decreases the propagation much faster than natural immunization, with a considerably smaller death toll. It is also possible to observe the differences at the start of effective recovering, i.e. when the recovered rate surpasses the infected rates, this is due to the early increase in isolation levels.

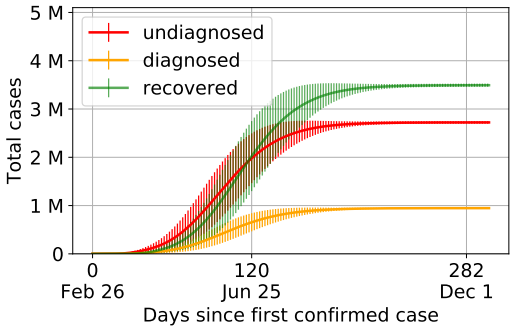
The peak of hospitalization occupancy occurs around a week before the death peaks, in any scenario. In this case, ICUs are very important because critical patients are treated there, which represents the cases of higher death rates. Within the "end isolation" setting, patients may occupy up to an average of 215,285 ($\pm$ 48,682) regular beds and 109,520 ($\pm$ 24,647) ICU beds. These numbers are by far greater than entire Brazil's capacity, as publicly-available and private ICU beds sum up to 45,848 [AMIB 2020, visited on 2020-05-08]. Even considering the better scenario, i.e. the lower bound of the standard deviation, the number of occupied ICU beds may reach around 86,000, which is also critical for Brazil's capacity (almost 2 times it's capacity). In this setting of "end isolation", the healthcare system would surely collapse.

When the isolation levels are kept, the numbers are significantly lower. However, the occupancy of 66,110 ($\pm$ 16,759) regular beds and 33,470 ($\pm$ 7,926) ICU beds is still critical for the Brazilian health system. Considering the creation of new provisional ICU units and good patient logistics, the situation may still remain under control during the peak of hospitalization occupancy. However, the results show that the hospital occupancy is prolonged considerably in this scenario, and they may stay functioning around their maximum capacity for up to a month (with an average of occupied ICU beds above 30,000). When increasing the isolation the peak of occupied beds is smaller, with an average of 63,226 ($\pm$ 20,682) regular beds and 31,816 ($\pm$ 10,592) ICU beds. Moreover, the shape of the curve throughout the days is different and the final numbers are considerably smaller. The peak also occurs around a week earlier and then decreases much faster. This scenario would be preferable as it has much more chances of not overloading the Brazilian healthcare system, relieving the hospital occupancy considerably faster and, therefore, contributing to the reduction of the number of deaths.

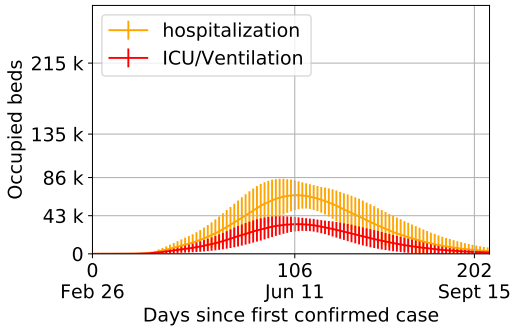
B.5 Conclusion

This work presents a new approach for the modeling of the COVID-19 epidemic dynamics based on multi-layer complex networks. Each node represents a person, and edges are social interactions divided into 6 layers: home, work, transports, schools, religions, and random relations. Each layer has its own characteristics based on how people usually interact in that activity. The propagation is performed using an agent-based technique, a modification of the SIR model, where weights represent the infection probability that varies depending on the layers and the

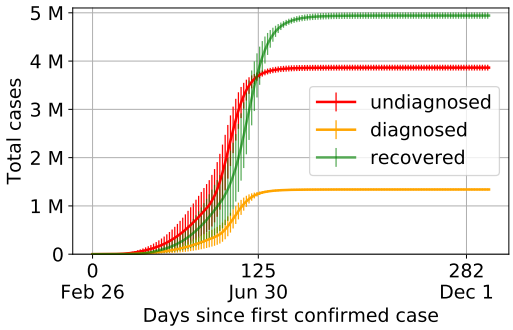
Figure 50 – Total number of infected and recovered cases and evolution of hospital beds utilization in 3 possible scenarios after 90 days (May 26): (a) Keep isolation level (no schools and religion, reduced work and transports), (b) End isolation (return schools, religion, work and transport to normal) or (c) Increase isolation (stop work and public transports)).



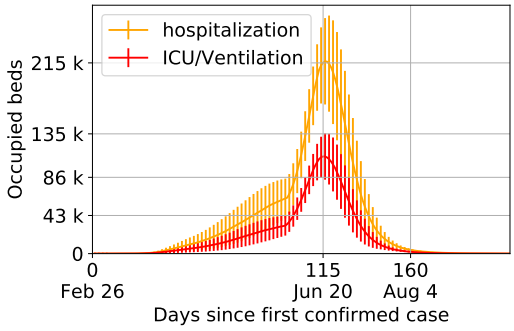
(a) Keep isolation level - cases



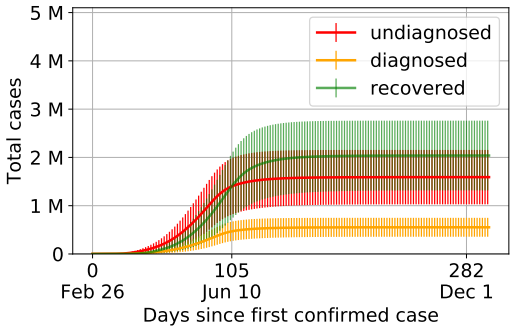
(b) Keep isolation level- hospitalization



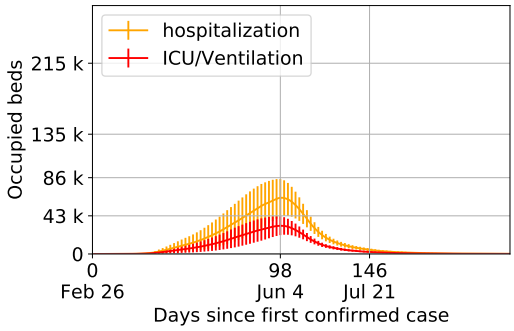
(c) End isolation, return all activities - cases



(d) End isolation, return all activities - hospitalization



(e) Increase isolation - cases



(f) Increase isolation - hospitalization

Source: Developed by the author

groups the node interacts, scaled by a β term that controls the chances of infection. The network structure is built based on demographic statistics of a given country, region, or city, and the propagation simulation is performed at time iterations, that represent days. Here, we studied in depth the case of the Brazilian epidemic considering its population properties and also specific events, such as when the first isolation measures were taken, and the impacts of future actions.

Brazil is a large and populated country with a wide variety of geographical location types,

climates, and it also has a lengthy border with other countries to the west. It is a challenging setting for any epidemiological study. Here we consider an average over all the country population, as we adjust the model output to match some statistics of the epidemic official reports. Brazil is performing fewer tests in comparison to other countries at the same epidemic scale, however, it is known that testing for infection is always limited, either due to the low number of tests or to the velocity of infections which the testing procedure cannot keep up to. We then considered that only hospitalization cases and 20% of the mild cases are diagnosed. Asymptomatic cases are not diagnosed and keep acting normally in the network, considering the active layers. Regarding the isolation of infected nodes, we take some optimistic assumptions: Mild cases (even those not diagnosed) are aware of its symptoms and isolate themselves at home. Severe and critical cases are eventually hospitalized, and then fully isolated from the network (removal of all its edges).

Under the described scenario, the network starts with all its layers and $\beta = 0.3$, representing that people are aware of the virus since the beginning (even before isolation measures). After 27 days of the first confirmed case, the first isolation measures are taken where schools and religious activities are stopped and work and transports keep functioning at 30% of the initial scale (achieved further reducing the β term). Different actions are then considered after 90 days of the first case: keep the current isolation levels, increase isolation, end isolation returning all activities to 100%, or returning only the work activities. The results show that keeping approximately the current isolation levels results in a prolonged propagation, as we are near the estimated peak (around June 5) with an average of 11,000 daily new cases and 1900 daily new deaths, and an average of 946,830 diagnosed cases (up to 3,6 million infected) and 149,438 deaths until the end of the year. In this scenario, hospitals may exceed its maximum capacity around June 11, but the efficient implementation of new ICU beds and good logistic management of patients may still keep the situation under control. However, this is a very optimistic assumption, considering that our definition of "keep isolation" considers social isolation above 50% as registered at the beginning of the Brazilian quarantine [Inloco 2020, visited on 2020-05-15]. The social isolation levels in Brazil are constantly decreasing even when we are still in a state of moderated quarantine, and it is possible to observe average isolation below 50% in most days of the past month (middle of April to middle of May 2020). Moreover, the results show that this prolonged scenario may cause hospitals to keep functioning at maximum capacity for up to a month. When analyzing other possible scenarios the situation may be considerably different. Relaxing isolation measures from now on causes an abrupt increase in the daily growth of cases and deaths, up to 5 times higher in comparison to the current isolation levels. Even if only work activities return while schools, religion, and transport activities remain inactive/reduced, the impact is very similar to returning all the activities, with a possible number of above 1,34 million diagnosed cases (up to 5,2 million infected), and around 212,105 deaths until the end of the year. This is, again, a very optimistic assumption as we do not consider the hospital overflow to calculate the death toll. Considering this aspect, ICU beds may be fully occupied in early June, and around the middle of the month their demand may reach up to 134,000 beds, which is around 3 times higher

than the entire country's capacity. The other alternative, which is the increase of isolation levels (lockdown), appears to be the only alternative to stop the healthcare system from entering a very critical situation. In this scenario, the growth in the number of daily cases and deaths would be mitigated, and faster. As we are near the peak of new cases at current isolation levels, estimated to be between the beginning and middle of June, increasing the isolation levels does not cause a significant impact on when the peak occurs or its magnitude. However, the disease spreading and the occurrences of new cases decrease much faster in this scenario in comparison to any other scenario studied here, with a difference of months. Moreover, the final numbers are considerably smaller, with an average of 552,855 diagnosed cases (up to 2.1 million infected) 87,059 deaths until the end of the year.

Although the proposed method includes various demographic information for the network construction, and an improved SIR approach to COVID-19, it still does not cover all factors that impact the epidemic propagation. As future works, one may consider more information such as the correlation between the age distribution within the social organization and the clinical spectrum of the 4 infection types (e.g. severe and critical cases are mostly composed of risk groups). Another possible improvement consists of increasing n (number of nodes of the networks), e.g. using a value near the real population of the studied society, which we avoided here due to hardware and time constraints (graph processing is costly). Another important point regarding the obtained results is related to the "keep isolation" scenario, which may be underestimated as we take various optimistic assumptions and also consider a fixed isolation level based on previously observed data, while most recent data shows that these levels are decreasing [Inloco 2020, visited on 2020-05-15]. Therefore, during the network evolution, a possible improvement is the use of dynamic isolation levels to better represent reality. It is also possible to consider various scenarios for future actions, such as 2 or more measures of increasing/reducing isolation. This may allow the discovering of new epidemic waves if social activities return too soon after the isolation period, such as what happened in 1918 with the Spanish flu.

ANNEX

C

DISTANCE TRANSFORM NETWORK FOR SHAPE ANALYSIS

Initial Consideration

This Annex is part of the primary and the secondary goal of this thesis. It models the distance-transform isolines of the shapes into complex networks and use the network features to classify them. It is a novel shape descriptor that combines network analysis and Euclidean distance transform. Experimental results demonstrate the robustness and high performance of the proposed shape descriptor. The method proves to be effective in capturing important shape characteristics and achieving accurate shape representation. This approach offers a promising tool for shape analysis and recognition tasks in various applications. The Annex presents a version of the paper that can be fully read at [Ribas, Neiva and Bruno 2019].

C.1 Introduction

Shape is a classical visual attribute and it is the most important feature for object characterization with its first studies dated from the 60's [Backes, Florindo and Bruno 2012, Ataer-Cansizoglu *et al.* 2013, Backes and Bruno 2010, Backes, Casanova and Bruno 2009]. In addition, the interest on this feature is also inspired by biological systems. This is because many biological processes such as biochemical reactions, involves the matching of shapes [Costa and Jr 2000].

Different approaches have been proposed along the years to represent shapes which can be classified into three main categories: skeleton-based, region-based and contour-based [Loncaric 1998]. This classification is made based on how features are extracted from the shape [Wang *et al.* 2014]. Skeleton-based methods, for instance, use the medial axes of the shape to extract its features. These methods have as advantage the robustness for shapes with occlusion and articulation. Examples of this category are: the Graph with Fractal [Bai *et al.* 2008] and Path

Similarity [Bai and Latecki 2008], among others [Goh 2008, Siddiqi and Pizer 2008, Cornea, Silver and Min 2007, Hilaga *et al.* 2001, Biasotti *et al.* 2006]. Region-based techniques, on the other hand, focus on the whole image to extract features (e.g., Zernike moments [Zhenjiang 2000, Khotanzad and Hong 1990] and those based on Hu moments [Hu 1962, Liao and Pawlak 1996], [Bhadange and Kalshetty, Lu and Sajjanhar 1999, Safar, Shahabi and Sun 2000, Chakrabarti *et al.* 2000]). Although, the ability to apply the method in generic shapes, the category cannot distinguish among objects that are very similar.

Finally, contour-based methodologies, which are considered in this paper, uses only the contour information of the shape to extract characteristics. Most of these methods, consider the contour as an ordered set of connected points, an intuitive way to deal with a sequence of dots. However, contour-based suffers when silhouette it not complete. The lack of points or occlusion of a shape region affect the results [Backes, Martinez and Bruno 2010]. Some examples of this category include Fourier descriptors [Wallace and Wintz 1980], Curvature Scale Space [Mokhtarian and Bober 2013], Multi-scale fractal dimension [Plotze *et al.* 2005, Torres, Falcao and Costa 2004] and based on Centroid–contour distance [Wang, Chi and Feng 2003].

To avoid this drawback, it was proposed the use of complex networks on shape contour [Backes, Casanova and Bruno 2009]. The network as shape descriptor method, uses the contour-based approach in the same way as the other methods in this group, but with the advantage that the contour elements does not need to be a sequence. Thus, it can recognize partial or corrupted silhouettes.

This paper is inspired in the network model proposed in [Backes, Casanova and Bruno 2009] and adds an extra component: The Euclidean distance transform. The latter method calculates the minimum distance from an image pixel to a region of interest [Rosenfeld and Pfaltz 1966]. It is a well known method largely used on computer vision, shape analysis and pattern recognition [Fabbri *et al.* 2008]. The EDT, has a strong link with morphological mathematics and can be understood as series of consecutive dilatations. Both the distance transforms and morphological mathematics has been successfully used for decades in shape analysis [Fabbri *et al.* 2008].

The idea of combining distance transform and networks is to create a method that can take advantage of both approaches. The networks proved to be robust and powerful to shape analysis [Backes, Casanova and Bruno 2009], on the other hand, the distance transform of a shape contour can contain richer information than the original contour since it carries all the information of the contour wave propagation and collision [Fabbri *et al.* 2008].

The proposed method obtains the distance transform map of contour and models each radius of dilatation as a network. Then, degree measurements from the network, modified by some transformation, is used to characterize the network. For shape characterization, informations about wave propagation and collision of the contour are added by the use of networks modeled from different radiuses of dilatation. This approach proved to be powerful and robust to describe

the shape given the results achieved on the experiments performed. This experiments were evaluated in two well-know benchmarks: generic shapes and ETH-80. Also, analysis were performed in plant species classification using the leaf shape contour, a challenger database due the variability of species present in nature. Additionally, the shapes were intentionally re-shaped to analyze characteristics such noise tolerance, scale invariance, rotate invariance and robustness. Tests with other literature methods were done in order to compare with proposed method.

C.2 Background

C.2.1 Networks

The networks field can be described as the intersection between graph theory and tools of statistical mechanics that gives a natural multidisciplinarity to it, combining Computer Science, Mathematics and Physics [Costa *et al.* 2007]. In the 50's, researches in graph theory conducted by [Erdös and Rényi 1961, Erdös and Rényi 1960, Erdös and Rényi 1959, Flory 1941] provided the basis of the research field. The approach was noticed to be very interesting due to its property of transcending from the traditional reductionist methodology to have the capacity to deal with different aspects of the problem at the same time (such as multiple iterations, actors and variables) [Miranda, Machicao and Bruno 2016].

The complex networks have been successfully adopted to develop methods in computer vision and pattern recognition. These methods, such as [Gonçalves, Machado and Bruno 2015, Gonçalves and Bruno 2013, Backes and Bruno 2013], include the study of modeling a problem into a complex networks, analysis of their topological structure and feature extraction. For instance, a texture image can be modeled as a network and its patterns represented by network connectivity. Then, the feature extraction is performed in terms of traditional measures of connectivity. In the literature, the problems of computer vision based on CN includes: texture analysis (e.g. [Chalumeau *et al.* 2006, Chalumeau *et al.* 2008]), refine edge (e.g. [Ferrari, Tuytelaars and Gool 2006]), boundary shape analysis (e.g. [Backes, Casanova and Bruno 2009, Zhao, Kong and Luo 2006, Tang, Zhang and Luo 2006]), etc.

C.2.2 Euclidean Distance Transform

The next method presented in this paper that will be further used in the proposal is the Euclidean Distance Transform. The technique have been used in tasks such as computer vision, graphics, shape analysis, pattern recognition, among others [Fabbri *et al.* 2008]. In the activity of shape analysis, for instance, the method have been used to match objects with the advantage of having better and smoother results when transformed images were compared [Torelli *et al.* 2010]. Also, EDT is able to compute morphological operations such as dilations and erosions [Russ 2016]. Thus, as shown in [MING-HUA and PING-FAN 1989], multiscale can be created by the application of these operations. According to [MING-HUA and PING-FAN 1989], the

appearance of an object is very dependable on the scale analyzed which makes the addition of different scales very important to obtain a complete comprehend of the object observed. For this reason, EDT is applied in this context to enrich the representation of the shape.

First of all, to understand the method a simple definition must be exposed. Basically, distance transform of a binary image I finds the minimum distance between background pixels B and an object C in foreground. Usually, the output of the EDT is named distance map S, which will be used in this paper. The distance map is composed of radiuses of dilation r, where each radius is a set of points $S_r \in S$ that have the same minimum distance to the object C.

The points in S_r, that forms an isoline, are a very important source of analysis in the distance map. Each radius represents how the object progresses along the evaluated dilatation. These radiuses contains important information about the propagation of waves of the EDT and the complexity of the contour. The "shock" between the waves are directly associated to specific characteristics of the object in analysis such as curves and size. Therefore, each radius of dilation can be analyzed separately, and roughly understood as a different scale. Consequently, characteristics of different 'scales' can be combined to obtain more robust features.

Although its proven importance, the computational complexity of the EDT can be high. In this way, several researchers have proposed methods to reduce cost by creating different implementations of exact and non-exact distance transform [Fabbri *et al.* 2008]. The paper uses a linear implementation of exact euclidean distance transform from [Maurer, Qi and Raghavan 2003] and available on Matlab.

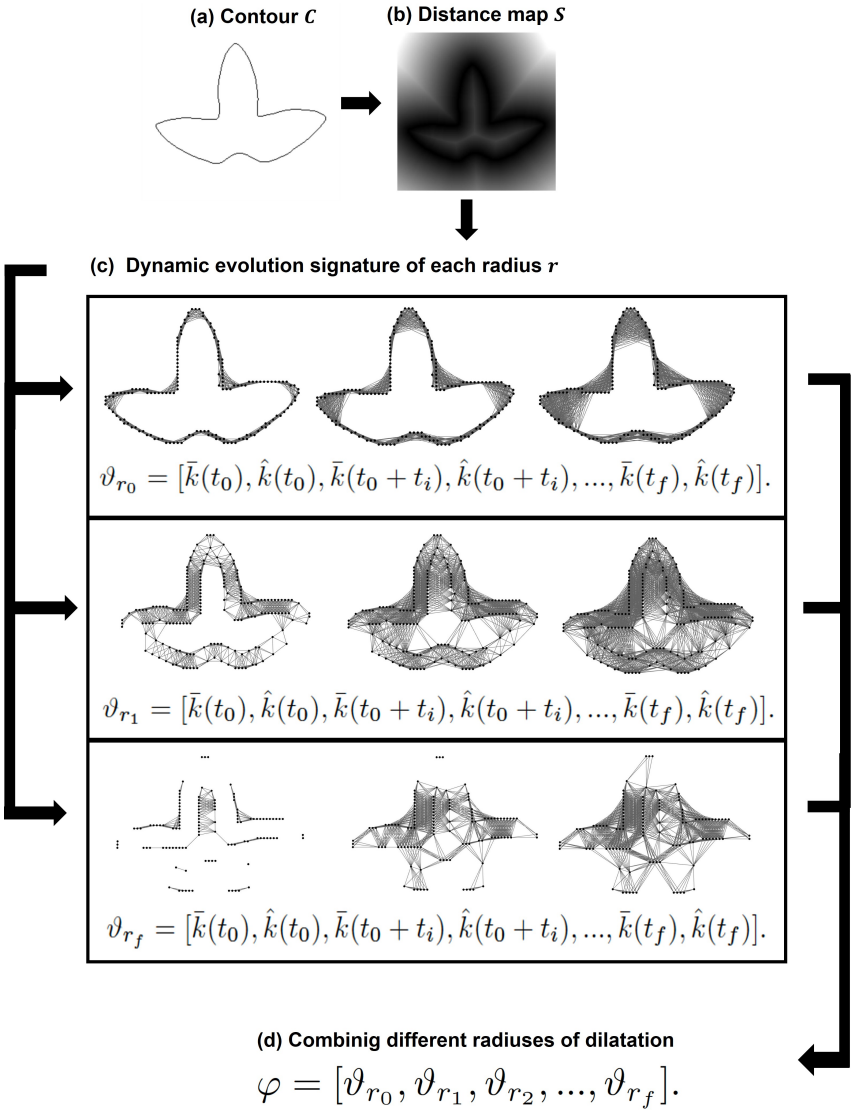
C.3 Distance Transform Network

In this section, we describe the distance transform network method for boundary shape analysis. The proposed method, also referenced here as DTN, combines distance transform and networks taking advantage of the robustness of the former and the extra shape information provided by the latter, in order to create a strong and robust shape descriptor. In the following subsections are described: (i) the network model, (ii) the analysis and metrics of the network, (iii) the composition of the feature vector and finally (iv) the parameters are evaluated.

C.3.1 Network model

In order to obtain the distance map and model it as a network, the Euclidean distance transform is applied on the image I, considering the contour C as the object of the interest (Figure 51(a)). From the output is obtained the distance map S (Figure 51(b)). An interesting approach of EDT on images is to evaluate how an object progresses along different radiuses of dilatation (distances), analyzing each radius separately and then combining its features for a robust classification. In this way, the proposal of this work is to analyze each radius of dilatation r of the distance map separately. Each r is composed of a set of points $S_r \in S$ and each point is

Figure 51 – Summarization of the proposed method.



Source: Developed by the authors

addressed as $s_i = [x_i, y_i]$, where x_i and y_i represent discrete values, the coordinates of the point i in the map.

In order to use the complex networks theory for the problem of shape analysis, given a subset $S_r \in S$ of a radius of dilatation r, this is modeled as a graph $G_r = (E_r, V_r)$. By using the radiuses of dilatation separately, new information of the shape related to propagation of waves of the EDT are added for analysis. Therefore, a graph is built, where each point $s_i \in S_r$ of a radius of dilatation r is represented as a vertex $v \in V_r$ of the graph (i.e., $S_r = V_r$). The set of non-directed edges $E_r : V_r \times V_r$ is defined by connections of all vertices in V_r to each to other. For each edge $e_{i,j} \in E_r$ that connects the vertices i and j, a non-negative weight w_{ij} defined by Euclidean distance between them is assigned according to:

$$w_{ij} = d(s_i, s_j) = \sqrt{(x_i - x_j) + (y_i - y_j)}. \tag{C.1}$$

Thus, the network is represented by a $N \times N$ weight matrix W_r,

$$W_r([w_i, w_j]) = d(s_i, s_j). \tag{C.2}$$

For the purpose of invariance, the weight matrix is normalized into the interval $[0,1]$, according to the highest weight:

$$W_r = \frac{W_r}{\max_{w_{ij} \in W_r}}. \tag{C.3}$$

Initially, the set of edges E_r connects all vertices from the network G_r, therefore, all vertices have the same number of connections, i.e., a regular behavior. However, a regular network has not any topological property and it is not considered a complex network. In this way, it is necessary to transform this regular network into a complex network, highlighting important properties that characterize the problem studied.

An approach to transform the network is to apply a threshold t on its edges, which produces a new set of edges E' [Costa *et al.* 2007]. This transformation consists in the selection of edges whose weights are smaller than a given threshold t. In Backes et al. [Backes, Casanova and Bruno 2009], it was showed that this approach allows to convert a initially regular network that models a contour in a complex network that presents small-world properties [Watts and Strogatz 1998]. The small-world complex network model presents two basic properties: (i) small-world and (ii) high clustering coefficient properties. The clustering coefficient is a measure related to the number N_Δ of connected triple and the number N_3 of triangles in the network. A connected triple is founded when two vertices i and j are connected, and i is also connected to k. A triangle is when j and k are also connected. The clustering coefficient is given by $CC = \frac{3N_\Delta}{N_3}$.

As a consequence of the high clustering coefficient is the existence of a small average geodesic path. A geodesic path is the shortest path in the network connecting two vertices. Once the complex network is obtained from a radius of dilatation, measures of its topological property can be extracted for pattern recognition. This task is described in the following sections.

C.3.2 Dynamic Evolution Signature

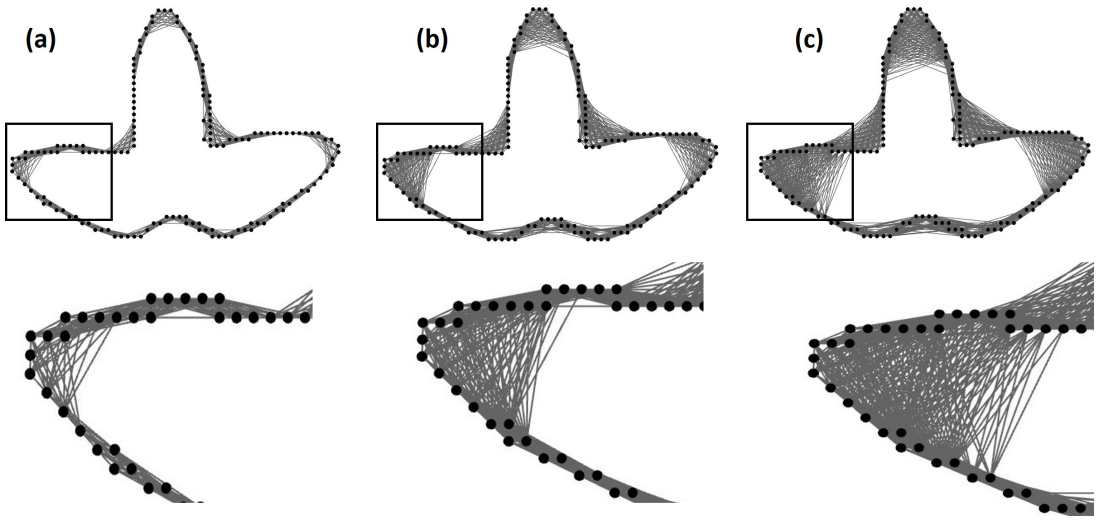
In image and object characterization, the complex network can provide tools for extraction of relevant information. Different complex networks may present a large range of characteristics, making the task of modeling the dynamics more difficult [Backes and Bruno 2010]. Although the modeling of the dynamic of a complex network is a difficult task, an interesting approach to obtain additional information about its structure is to apply sequential transformations over the network and compute some properties in each step [Costa *et al.* 2007]. In this way, the proposal of this work is: given a network that models a radius of dilatation of the EDT as described in previous section, apply a transformation $\delta_t(W_r)$ with different values of t

over the network and use degree measures in each stage of the network evolution as a feature vector (Figure 51(c)). This transformation consists in the application of a threshold t on each element of the weighted matrix W_r, obtaining as result an unweighted matrix A^r C.4 [Backes, Casanova and Bruno 2009].

$$A_t^r = \delta_t(W_r) = \forall w \in W_r \begin{cases} a_{ij} = 0 & \text{if } w_{ij} \leq t \\ a_{ij} = 1 & \text{if } w_{ij} > t \end{cases} \tag{C.4}$$

The shape characterization using the network is performed using several transformation $\delta_t(W_r)$ with a set of threshold $T, t \in T$. This set is defined by a initial threshold t_0 and incremented at a regular interval t_i until a final threshold t_f. Figure 52 shows the network dynamic evolution of a radius of dilatation for different threshold values. As can be seen, as the threshold increases, the number of edges in the curves is also increases.

Figure 52 – Radius of dilatation $r = 0$ modeled as a transformed network by different values of threshold. (a) $t = 0.15$, (b) $t = 0.2$ and (c) $t = 0.25$.



Source: Developed by the authors

Once obtained the network dynamic evolution, we use a feature vector composed of degree measurements for shape characterization.

C.3.2.1 Degree Descriptors

For characterization of a radius of dilatation r modeled as a network, degree descriptors are computed from the unweighted matrix A_t^r in each step t. The average degree $\bar{k}$ and the max degree $\hat{k}$ measurements, presented in Section C.3.1, are used to compose the feature vector. Before computing these measures, a normalization of the vertices degree by the number of vertices in the network is performed (Equation C.5). The objective of this normalization is to

reduce the influence of the network size in the measurements since the size of the network for each radius and threshold analysed if different.

$$\forall k_i = \frac{k_i}{N} \tag{C.5}$$

To describe the topology of the network G_r, that models a radius of dilatation r, we present a feature vector ϑ_r composed of the average degree $\bar{k}$ and the max degree $\hat{k}$. This feature vector consists of the concatenation of $\bar{k}$ and $\hat{k}$ from each step t of the network evolution:

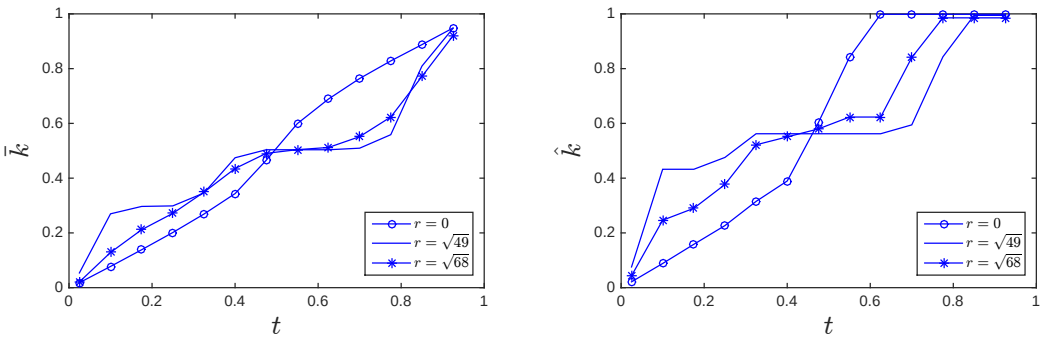
$$\vartheta_r = [\bar{k}(t_0), \hat{k}(t_0), \bar{k}(t_0+t_i), \hat{k}(t_0+t_i), ..., \bar{k}(t_f), \hat{k}(t_f)]. \tag{C.6}$$

C.3.3 Combining Different Radiuses of Dilatation

To obtain a robust feature vector with different information about the shape contour, we concatenate feature vectors ϑ_r for different values of radiuses of dilatation r (Figure 51(d)). Figure 53 presents the feature vector ϑ_r composed by the average degree and max degree of a shape for different radiuses of dilatation r. As can be seen, each radius modeled as network contains different information about the shape, which increases the performance in the final classification task. To create the feature vector, the radius values ranges between the interval $r_0 \leq r \leq r_f$, which is evaluated in Section C.3.4. Thus, a combined feature vector φ that considers information from different networks modeled using different values of radius r is given by:

$$\varphi = [\vartheta_{r_0}, \vartheta_{r_1}, \vartheta_{r_2}, ..., \vartheta_{r_f}]. \tag{C.7}$$

Figure 53 – Feature vector ϑ_r composed by the average degree and max degree of a shape for different radiuses of dilatation r



(a) Average degree feature (b) Max degree features

Source: Developed by the authors

The final vector φ is able of represent a shape considering desirable properties in the shape classification task, which are rotation and scale invariance, obtained through edge and

degree normalization, noise intolerance, reached by the application of the EDT and degradation invariance, from the fact that the network model does not require the extraction of ordered points from the contour.

C.3.4 Parameter Evaluation

A parameter evaluation is performed. DTN method assumes the following parameters: (i) set of thresholds T of the network dynamic evolution, and (ii) the initial and final radius of dilatation, r_0 and r_f. To accomplish this task, we use the three databases considered in this paper: ETH80, Leaves and Generic shapes. The classification of the feature vectors was performed using LDA classifier [Everitt and Dunn 2001]. The full experimental analysis can be refered in [Ribas, Neiva and Bruno 2019]. In summary, according to the plots, the method achieves the best performance in the three database for $r_0 = 0$ and $r_f = 113$.

This analysis suggests that the parameters found for the proposed method will achieve good results in any classification basis. Therefore, we believe that it is not necessary a comprehensively search for the parameters of the proposed method for different datasets.

C.4 Experimental Setup

To evaluate the proposal of this paper, three different databases were tested: Kimia-99 [Sebastian, Klein and Kimia 2004], ETH-80 [Leibe and Schiele 2003] and Leaves [Backes, Casanova and Bruno 2009]. The last database is also evaluated under different conditions: rotation, continuous and random degradation, noise and scale. Also, to compare results of the novel method with other shape classification algorithms seven different methods were chosen from the literature (among classic and new ones). Finally, the machine learning algorithms Linear Discriminant Analysis (LDA) [Everitt and Dunn 2001] and Support Vector Machine (SVM) [Hearst *et al.* 1998] are used for classification.

C.4.1 Shape Databases

Generic shapes

The first database has 99 shape images classified in 9 classes with 11 shapes each [Sebastian, Klein and Kimia 2004]. It contains shapes such as rabbits, men, airplanes, tools and fish. In the set, images were obtained by projections from 3D shapes, which can cause appearance or disappearance of some parts of it. Also, some images in the database suffers from bad segmentation and rotation adding a challenge on object analysis due to deformation. Samples shapes of this database are presented in Figure 54.

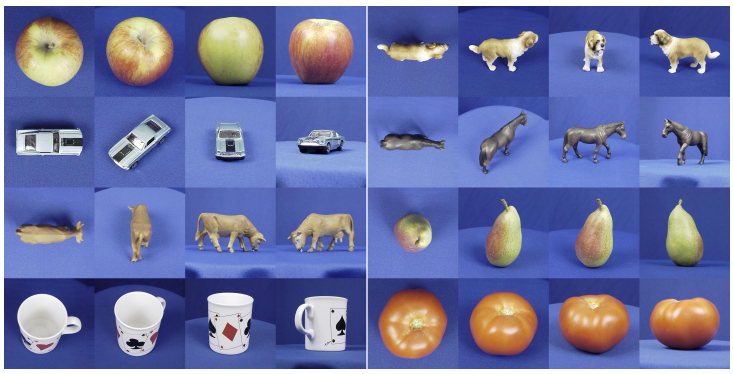
Figure 54 – Example of shapes image from the generic shapes database.

Source: Developed by the authors

ETH-80

ETH-80 databases contains 3280 separated in 8 labels [Leibe and Schiele 2003]. Originally in 3D, the authors also provide segmentation masks (used in the experiments). Categories are constructed by human-made and natural objects. For each type of item, different styles are used. For instance, in the category of cups, the database presents small and bigger cups while for cars, different brands of toy cars are available. In addition, for each object, 41 images were taken in different viewpoints including upper view. This setup adds variance in rotation and the final contour can be very different for each type of object. Samples images are shown in Figure 55.

Figure 55 – Example of some shapes image from the ETH-80.



Source: Developed by the authors

Leaves

This database contains 600 images of 30 leaf species. As different species of plants may contain similar shapes, the classification of this database is a challenging task. Also, in the digitalization of the images, overlaps can occurs in the adjacency of the objects. Figure 56 shows some sample images of this database proposed in [Backes, Casanova and Bruno 2009]. This database is more suitable for practical applications due to the presence of one type of major category object (leaves) divided in sub categories which turns the analysis very hard. Furthermore, in order to evaluate rotation, scale and noise invariance, this database has also supplementary sets containing features generated as follows:

- Rotation: The first modified database of leaves contains images rotated at 7°, 35°, 132°, 201° and 298° resulting in a new set with 3600 shapes. Figure 57(a) shows a sample object rotated by different angles.

- Scale: scaled by a factor of 125%, 150%, 175% and 200% original images generates a different database with original and scaled images. Figure 57(b) shows a sample object scaled by different factors.

- Noise: different levels of uniform noise were added on the images. The noise was generated according to an interval [-l...l] where l is the level of the noise. In the experiments, four different levels were applied generating a new set with 80 images for each class. Figure 57(c) shows examples of the same object with different levels of noise.

- Random degradation: Levels of degradation, from 15% to 65%, are used to produce this new set. In this case, degradation is applied randomly in the contour and, therefore, sequential points are harder to be obtained. Each level of degradation produces 20 modified images and samples of them are shown in Figure 57(e).

- Continuous degradation: this last modified database is produced by continuous degradation in Leaves database. It creates gaps in the contour making recognition harder. Levels of degradation are applied as the same as previously random one but in a continuous manner. Therefore, robustness to this effect can be evaluated. Figure 57(d) shows a sample of this approach.

Figure 56 – Some of leaves images from the Leaves database

Source: Developed by the authors

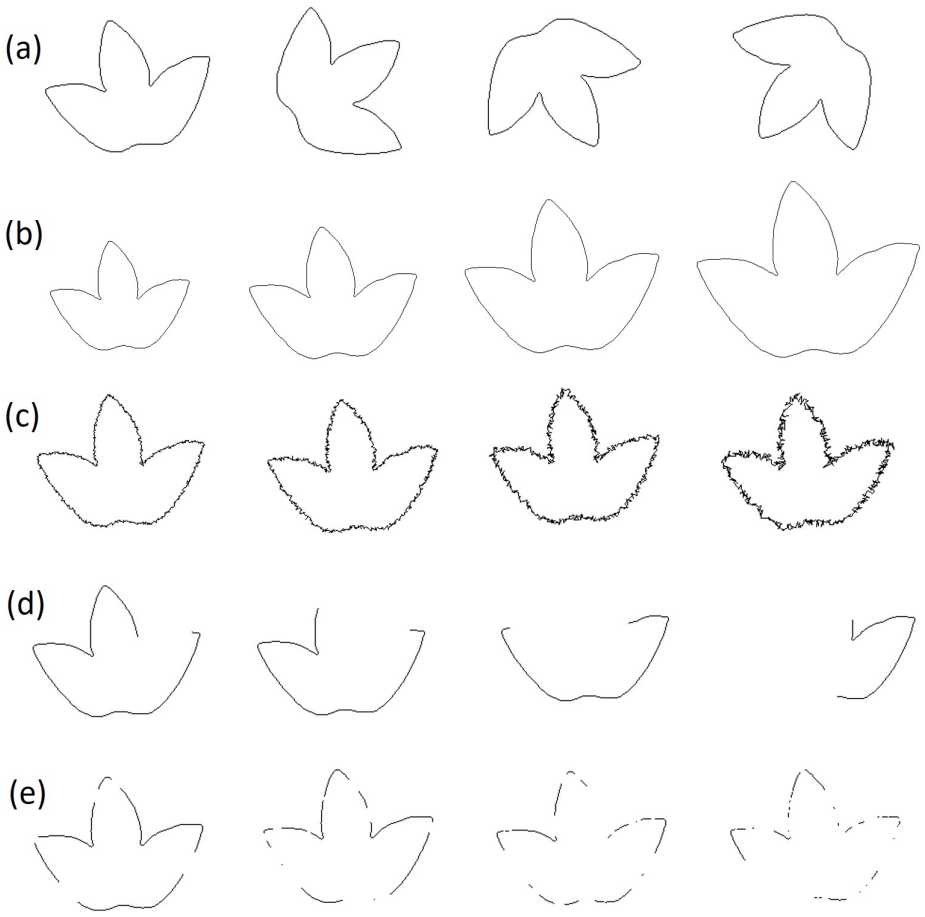
C.4.2 Shape Classification Methods

For any new proposed method it is essential to compare the new results with performances of well-known shape classification methods. Therefore, seven methods from the literature are described in this section and compared with the proposal latter in Section C.5.

C.N. degree

Like the proposed method, this feature extractor also models the image as a network. However, it uses the contour points (not the distance map) and distances between them to create the network and threshold distances to reach robustness. Features are obtained concatenating measures of the network such as degree and joint degree [Backes, Casanova and Bruno 2009].

Figure 57 – Example of applied artifacts on a contour. (a) rotated contours, (b) scaled contours, (c) noisy contours and contours with (d) continuous degradation and (e) random degradation.



Source: Developed by the authors

Fourier Descriptor

Fourier shape descriptor computes the spectral transform of the contour and uses this information to extract features [Osowski *et al.* 2002]. In the experiments, 20 most significant coefficients of the spectrum were used as features of the image.

Curvature Descriptors

The method analyses the contour as a curve extracting features such as maximum and minimum points which can represent important characteristics of the shape like changes in the directions [Wu and Wang 1993]. The total size of feature vector obtained for each image in this method is 25.

Zernike Moments

Very simple but useful, this technique computes a set of zernike moments (order 0 to 7) from the image and uses them as feature vector. Zernike moments have rotation invariance

properties and represent the magnitude of orthogonal complex moments of the shape [Zhenjiang 2000].

Multi-Scale Fractal Dimension

Complexity of the shape can be measured by fractal dimension. In this method, the contour is understood as a curve and changes in it are related to high complexity [Torres, Falcao and Costa 2004]. Feature vector contains the 50 most relevant points of the shape to represent the object in an one-dimensional vector.

Segment Analysis

This simple but powerful method describes the contour based on straight lines segments statistics. It considers portions of continuous points and computes the length of the straight line between extreme points of the portion [Junior and Backes 2015]. Average and standard deviation are computed from each line. Different portions are reached by different predefined percentages of the contour size. Final feature vector contains 34 features.

Bag of Contour Features

This mid-level modeling of shape decomposes the image in a set of contour fragments and each of them is coded according to a shape context descriptor [Wang *et al.* 2014]. The model uses shape pooling to compute an histogram for each shape and the method outputs a high number of features for each image which are labeled in the original paper by SVM.

C.4.3 Classification Setup

The classification of the feature vectors obtained by the methods is performed using two well-known methods LDA [Everitt and Dunn 2001] and SVM [Hearst *et al.* 1998] following a stratified 10-fold cross-validation scheme. In the experiments of rotation and scale it is used a leave-one-out cross-validation scheme. Only in these two setups, the processed samples (i.e. rotated and scaled) from a same shape are used for test and the rest for training. This scheme was adopted to avoid that information about the test samples were used for training.

C.5 Results and Discussion

This section compares the performance method with the methods described in Section C.4.2. In the following, it will be presented and discussed the results obtained for the three shape databases used (Section C.4.1).

The first results are presented for the ETH-80 database, which contains images of natural and human-made objects such as fruits, vegetables, animals and vehicles. Table 16 shows the

results for ETH-80 recognition according to all methods tested. As can be noticed, the proposed method, DTN, obtained the higher recognition rate among all shape descriptor methods. Results show that the addition of dilation information of the shape improved robustness to viewpoint variance which occurs in the database. The C.N. degree method which was the inspiration for this proposal was increased in almost 20% of CCR using the LDA classifier and 10% using the SVM classifier. BCF algorithm also shows high classification rate for the set but uses a large number of features and has a higher computational cost.

Table 16 – Comparison of proposed method with literature methods for ETH-80 and generic shapes database.

Methods	ETH-80		Generic shapes	
	LDA	SVM	LDA	SVM
Proposed Method	92.02 (± 1.68)	93.47 (± 1.10)	99.00 (± 3.16)	98.00 (± 4.03)
C.N. degree	72.34 (± 2.34)	82.92 (± 2.58)	96.00 (± 6.99)	95.95 (± 5.16)
Fourier	79.26 (± 1.56)	86.12 (± 1.89)	93.88 (± 7.07)	97.97 (± 4.21)
Curvature	62.43 (± 2.03)	72.01 (± 2.61)	78.77 (± 15.2)	84.84 (± 12.5)
Zernike	82.86 (± 2.01)	87.98 (± 1.34)	95.00 (± 5.27)	87.98 (± 1.34)
M. S. fractal dimension	73.96 (± 1.65)	76.73 (± 1.71)	95.00 (± 9.71)	76.73 (± 1.71)
Segment analysis	78.71 (± 2.53)	80.00 (± 2.46)	98.00 (± 4.21)	98.80 (± 3.16)
BCF	-	91.49	-	-

Source: Developed by the authors

The second database, Generic shapes database, showed a high performance for all methods evaluated in this work (see Table 16). Although the simplicity of segment statistics algorithm, it outputs one of the highest CCR among all the methods compared. By using the SVM classifier the proposed method obtained a CCR very close of the segment analysis method. However, again the DTN outperforms all results using LDA classifier. The curvature descriptor has the lowest accuracy rate for this database and highest standard deviation.

The third experiment was conducted into the Leaves databases, where four situations was considered: the original database, scale, rotation, degradation (random and continuous) and noise. Besides the variations of the dataset, it has an interesting characteristic for shape analysis benchmark. It is composed by real leaves from the Brazilian flora (30 species of plants) [Backes, Casanova and Bruno 2009]. Due to the similarities between the leaves from different species and the high variety within the leaves in the same species, it is a very difficult database capable to analyze very well the powerfulness of the shape analysis methods.

Table 17 presents the correct classification rate achieved by DTN and compared methods, when applied to the original Leaves databases, rotated and scaled database. In the original database the results show a higher performance of the proposed method when compared to all other methods, independent of the classifier. For instance, the proposed method improves the recognition performance from 84 % to 94.16 % over the C.N. degree method and from 83.33 % to 94.16 % over the third top method, the Segment analysis method using LDA classifier.

In order to confirm the properties of rotation and scale invariance discussed in Section C.3.3, we performed experiments in modified databases from original Leaves. The experimental

results using rotated and scaled images confirm the great invariance of the proposed method over these kinds of transformations. DTN obtained the best CCR in the two databases independent on the classifier used, LDA or SVM. Notice that the proposed method achieved similar CCR for the two classifiers, while other methods were sensitive to the machine learning algorithms. It suggested a consistent feature vector and consequently a good invariance to scale and rotation.

The results from noise tolerance experiment using LDA classifier are presented in Table 18. The proposed method demonstrates a great capacity for shape classification even in contours with a high quantity of noise. In the experiments using noise rate Level 1, Level 2 and in the combination of all levels, the proposed method has the best performance compared to other methods. However, using the noise rate Level 3 and Level 4 the C.N. degree method has a slight superiority in the performance compared to DTN. We also emphasize that the Curvature method applies a low-pass filter when it is being computed, reducing the noise of the contour. However, the proposed method do not need this kind of preprocessing and it is still able to achieve a good performance in noisy shapes.

Finally, one important property that needed to be evaluated is the robustness of the method to degradation. The degradation in the shape is the lack of information in its contour. In this way, we evaluate the capacity of a method to recognize a shape when it is not complete. This problem can be easily found due to bad image acquisition. Figure 58(a) presents the correct classification rate in function of continuous degradation level for various methods using the LDA classifier. As the degradation increases the correct classification rate decreases. As noticed in Figure 58(a), the correct classification rate is achieved by DTN and maintains a superior CCR when compared to other methods for almost all degradation levels. These results show a great robustness of the proposed method compared to all other methods.

Another experiment was also performed to evaluate the robustness against random contour degradation. In this experiment, points of different places of the shape contour are removed. Figure 58(b) shows the results of robustness for several method for this type of artifact. These results also show a high robustness of the proposed method compared to other methods. The results of Fourier, Curvature and Segment statistics methods were not presented in this experiment, due the requirement of sequential extraction of points in the contour which was not possible in this case. This is a problem often found in real applications. Thus, it is important to emphasize that the proposed method does not depended of a contour extraction in a sequential way.

C.6 Conclusion

In this paper, we have proposed a new method for shape classification based on networks and Euclidean distance transform. Given a contour, or even a subset (dots in sequence or not) of a contour, the DTN method applies the Euclidean distance transform and models each radius of

Table 17 – Comparison of proposed method with literature methods for Leaves database in different type of experiments.

Type of experiment	Methods	LDA	SVM
Original	Proposed Method	94.16 (± 4.24)	93.16 (± 2.41)
	C.N. degree	84.00 (± 5.67)	85.16 (± 3.63)
	Fourier	74.66 (± 6.97)	83.16 (± 4.61)
	Curvature	77.00 (± 6.42)	81.66 (± 3.33)
	Zernike	69.66 (± 5.76)	76.33 (± 5.07)
	M. S. fractal dimension	71.16 (± 4.51)	73.66 (± 5.81)
	Segment analysis	83.33 (± 5.44)	83.50 (± 4.26)
	Bag of contour fragments	71.00 (± 0.45)	65.00 (± 4.84)
Rotated	Proposed Method	89.06	88.67
	C.N. degree	80.61	84.28
	Fourier	67.53	77.17
	Curvature	76.31	83.33
	Zernike	65.53	76.64
	M. S. fractal dimension	58.17	68.58
	Segment analysis	81.47	76.22
Scaled	Proposed Method	92.38	93.08
	C.N. degree	81.46	85.04
	Fourier	63.25	85.29
	Curvature	76.92	84.46
	Zernike	53.38	71.33
	M. S. fractal dimension	65.25	68.50
	Segment analysis	82.04	79.75

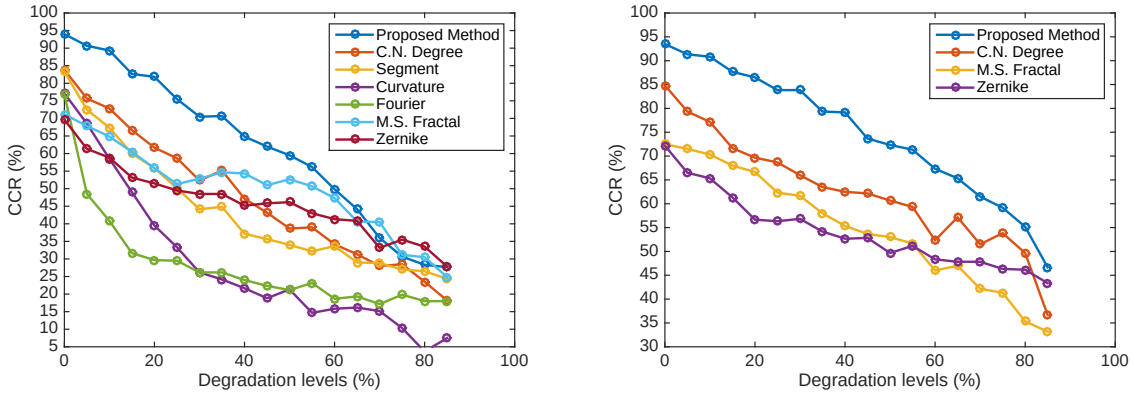
Source: Developed by the authors

Table 18 – Comparison of the proposed method with literature methods for Leaves database corrupted by different levels of noise using the LDA classifier.

Method	Level 1	Level 2	Level 3	Level 4	All levels
Proposed Method	88.33 (± 3.23)	80.83 (± 5.10)	77.16 (± 3.51)	74.83 (± 5.35)	86.29 (± 2.18)
C.N. degree	79.50 (± 5.55)	78.16 (± 4.87)	78.16 (± 5.29)	76.66 (± 5.38)	77.70 (± 2.76)
Fourier	65.50 (± 4.84)	57.83 (± 6.03)	51.00 (± 6.99)	45.16 (± 5.23)	49.58 (± 1.94)
Curvature	70.33 (± 6.02)	70.16 (± 6.35)	66.33 (± 6.27)	63.16 (± 6.05)	69.25 (± 2.42)
Zernike	67.50 (± 5.89)	66.83 (± 5.35)	66.83 (± 6.77)	63.83 (± 5.61)	62.33 (± 3.86)
M.S fractal dimension	65.00 (± 4.64)	63.50 (± 5.63)	63.33 (± 5.27)	59.66 (± 5.48)	64.20 (± 2.37)
Segment analysis	75.50 (± 6.80)	72.16 (± 5.72)	67.00 (± 5.37)	70.00 (± 5.09)	68.95 (± 1.35)

Source: Developed by the authors

Figure 58 – Comparison of robustness in contour degradation for various methods using LDA classifier.



(a) Robustness in continuous degradation

(b) Robustness in random degradation

Source: Developed by the authors

dilation as a network. The shape is effectively represented by a feature vector composed of the degree measures from the network dynamic evolution of different radiuses of dilation.

We have demonstrated the performance and the robustness of the proposed method in well-know databases of the literature and different properties. First, a study was carried out to evaluate the effect of each parameter of the proposed method on performance. The evaluation parameters demonstrated that the proposed method is not sensitive to the parameters setup, once it was used the same parameters for all databases. This is a very important characteristic, since it suggests that the proposed method could be used with the preconfigured parameters and the user does not need to setup anymore.

The experimental results were compared with other literature methods and it was proved that the proposed method outperforms the existing ones. In addition, experimental results of rotation, scale, noise and degradation demonstrated that DTN is very robust in shape recognition, and outperforms the compared methods in all aspects.

Considering the last experiments, on Leaves database, where rotation, scale, noise and degradation were present and the good performance obtained into the discrimination of the three shape databases considered, we can conclude that the method is robust by its ability to the invariant to different artifacts:

- *Rotation invariance:* the method has a rotation invariance characteristic which was achieved by the normalization of the weight matrix W. In the experiments, it was normalized within the interval $[0, 1]$ maintaining the proportion of the edges weights. Thus, considering images in different rotations, the weight of the largest edge was preserved, ensuring the same properties for the different sets of edges E.

- *Scale invariance:* the property of scale invariance was also noticed in the proposed method, the DTN. It was all, again, due to the normalization of the matrix W. The normalization

was responsible to rescale the edges, creating scale tolerance in the classification. For instance, two images in different scales produce different numbers of points in its contour. Once $S = V$, two radiuses of dilatation S_r^a and S_r^b of similar contours in different scales, produce networks with different number of vertices. Thus, the degree k_i is directly affected by the number of network vertices N. This problem was solved with the normalization of the degree k_i by the number of vertices in the modeled network, as presented previously in Section C.3.2.1 and also used in [Backes, Casanova and Bruno 2009].

- *Noise tolerance:* it is a fact that in the image acquisition process, errors in the contour can be added. This occurs because this task is not perfect, causing a variety of interference and noise. However, the way the contour (i.e., the radiuses of dilatation from contour) was modeled as network allowed us to analyze the shape without decreasing the performance in classification.

- *Degradation robustness:* in the DTN method, the network does not have information about space and sequence of the points of each radius of dilatation. From the distance map, the modeled patterns are equal in the feature space, in theory. This property allowed the method to not require for a sequential extraction of the points of contour. This is important, once the radiuses of dilatation are not continuous. Thus, it is only needed the coordinates of points of radiuses of dilatation. Specially for random degradation, the method have showed an as advantageous technique compared to some other shape descriptors such as Fourier, Curvature and Segments descriptor that were not able to classify the shapes due the sequential requirement. This characteristic is very interesting for real world problems, that needs to deal with degraded contours or even with contours produced in a sparse and random way.

The Distance Transform Network (DTN) proved its good performance. The capacity to deal with degraded contours and random dots contours, and finally the convenience of no parameter setup needed, make it a very good option for shape analysis.


ICMC
USP
SÃO CARLOS