University of São Paulo
Interunit Bioinformatics Graduate Program

Heitor Baldo

Towards a Quantitative Theory of Digraph-Based Complexes and its Applications in Brain Network Analysis

PhD Thesis

This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 88887.464712/2019-00

São Paulo
2024

Heitor Baldo

Towards a Quantitative Theory of Digraph-Based Complexes and its Applications in Brain Network Analysis

Rumo a uma Teoria Quantitativa de Complexos Baseados em Dígrafos e suas Aplicações na Análise de Redes Cerebrais

Final Version

Ph.D. thesis presented to the Interunit Bioinformatics Graduate Program at the University of São Paulo to obtain the degree of Doctor of Science.

Concentration area: Bioinformatics

Supervisor: Prof. Dr. Koichi Sameshima
Faculdade de Medicina-USP
Co-supervisor: Prof. Dr. André Fujita
Instituto de Matemática e Estatística-USP

São Paulo
2024

Ficha catalográfica elaborada com dados inseridos pelo(a) autor(a)
Biblioteca Carlos Benjamin de Lyra
Instituto de Matemática e Estatística
Universidade de São Paulo

Baldo, Heitor
Towards a Quantitative Theory of Digraph-Based Complexes and its Applications in Brain Network Analysis / Heitor Baldo; orientador, Koichi Sameshima; coorientador, André Fujita. - São Paulo, 2024.
213 p.: il.

Tese (Doutorado) - Programa Interunidades de Pós-Graduação em Bioinformática / Instituto de Matemática e Estatística / Universidade de São Paulo.
Bibliografia
Versão corrigida

1. Graph Theory. 2. Digraph-Based Complexes. 3. Partial Directed Coherence. 4. Brain Connectivity. 5. Epilepsy. I. Sameshima, Koichi. II. Título.

Bibliotecárias do Serviço de Informação e Biblioteca
Carlos Benjamin de Lyra do IME-USP, responsáveis pela
estrutura de catalogação da publicação de acordo com a AACR2:
Maria Lúcia Ribeiro CRB-8/2766; Stela do Nascimento Madruga CRB 8/7534.

Lovingly dedicated to my mother,
Maria Aparecida Marchi Baldo
(in memoriam)

Acknowledgments

I'd first like to acknowledge my thesis advisor, Prof. Koichi Sameshima, for allowing me to do research in this unique field, for his endless patience, and for all the valuable discussions. Secondly, I'd like to acknowledge my co-advisor, Prof. André Fujita, for all the support. I'd also like to express my sincere gratitude to Prof. Luiz Baccalá for all the valuable insights, tips, and discussions, and to all my friends and colleagues for the friendly conversations and emotional support during this period.

Also, I'd like to thank my parents, my mother, Maria Aparecida Marchi Baldo (*in memoriam*), and my father, João Edgar Baldo, who always supported my education and always encouraged me in difficult times throughout my life.

Finally, I'd like to thank the Interunit Bioinformatics Graduate Program at the University of São Paulo, firstly for accepting me as a graduate student and, secondly, for all the support, and CAPES for the financial support.

Abstract

The development of mathematical methods for studying the structure, organization, and functioning of the brain has become increasingly important, especially in the context of brain connectivity networks studies, highlighting, in recent decades, methods associated with graph theory, network science, and computational (algebraic) topology. In particular, these methods have been used to study neurological disorders associated with abnormal structural and functional properties of brain connectivity, such as epilepsy, Alzheimer's disease, Parkinson's disease, and multiple sclerosis.

In this work, we developed new mathematical methods for analyzing network topology and we applied these methods to the analysis of brain networks. More specifically, we rigorously developed quantitative methods based on complexes constructed from digraphs (digraph-based complexes), such as path complexes and directed clique complexes (alternatively, we refer to these complexes as "higher-order structures," or "higher-order topologies," or "simplicial structures"), and, in the case of directed clique complexes, also methods based on the interrelations between the directed cliques, what we called "directed higher-order connectivities." This new quantitative theory for digraph-based complexes can be seen as a step towards the formalization of a "quantitative simplicial theory."

Subsequently, we used these new methods, such as characterization measures and similarity measures for digraph-based complexes, to analyze the topology of digraphs derived from brain connectivity estimators, specifically the estimator known as information partial directed coherence (iPDC), which is a multivariate estimator that can be considered a representation of Granger causality in the frequency-domain, particularly estimated from electroencephalography (EEG) data from patients diagnosed with left temporal lobe epilepsy, in the delta, theta and alpha frequency bands, to try to find new biomarkers based on the higher-order structures and connectivities of these digraphs. In particular, we attempted to answer the following questions: How does the higher-order topology of the brain network change from the pre-ictal to the ictal phase, from the ictal to the post-ictal phase, at each frequency band and in each cerebral hemisphere? Does the analysis of higher-order structures provide new and better biomarkers for seizure dynamics and also for the laterality of the seizure focus than the usual graph theoretical analyses?

We found that all simplicial characterization measures considered in the study showed statistically significant increases in their magnitudes from the pre-ictal phase to the ictal phase, for several higher orders, for both cerebral hemispheres, particularly in

the delta and theta bands but no statistically significant changes were observed from the ictal to the post-ictal phase, which may suggest that several topological and functional aspects of brain networks change from the pre-ictal to the ictal phase, at various levels of the higher-order topological organization. Regarding the laterality of the seizure focus, the analysis based on simplicial similarities found no statistically significant difference between the clique topology of the left and right hemispheres in the ictal phase. We conclude from this study that, despite a number of limitations, there may be evidence supporting the viability and reliability of using higher-order structures associated with digraphs to identify biomarkers associated with epileptic networks. However, further research is needed, and the applicability of the newly introduced methods to other disorders of brain connectivity networks will also depend on future studies.

Keywords: Graph Theory; Directed Clique Complexes; Path Complexes; Directed Higher-Order Connectivity; Partial Directed Coherence; Brain Connectivity; Electroencephalography; Epilepsy.

Resumo

O desenvolvimento de métodos matemáticos para o estudo da estrutura, organização e funcionamento do cérebro tem se tornado cada vez mais importante, sobretudo no âmbito de estudos das redes de conectividade cerebral, destacando-se, nas últimas décadas, os métodos associados à teoria dos grafos, ciência de redes e topologia (algébrica) computacional. Em particular, esses métodos têm sido usados para estudar desordens neurológicas associadas com propriedades estruturais e funcionais anormais da conectividade cerebral, tais como epilepsia, doença de Alzheimer, doença de Parkinson e esclerose múltipla.

Neste trabalho, desenvolvemos novos métodos matemáticos para análise da topologia de redes e a aplicação destes métodos à análise de redes cerebrais. Mais especificamente, desenvolvemos rigorosamente métodos quantitativos baseados em complexos construídos a partir de dígrafos (complexos baseados em dígrafos), como complexos de caminhos e complexos de cliques direcionados (alternativamente, referimo-nos à esses complexos por "estruturas de ordem superior", ou "topologias de ordem superior", ou "estruturas simpliciais"), e, no caso de complexos de cliques direcionados, também métodos baseados nas interrelações entre os cliques direcionados, o que chamamos de "conectividades de ordem superior direcionadas". Essa nova teoria quantitativa para complexos baseados em dígrafos pode ser vista como um passo em direção à formalização de uma "teoria quantitativa simplicial".

Subsequentemente, usamos esses novos métodos, como medidas de caracterização e de similaridade para complexos baseados em dígrafos, para analisar a topologia de dígrafos derivados de estimadores de conectividade cerebral, especificamente do estimador conhecido como coerência parcial direcionada informacional (iPDC), que é um estimador multivariado que pode ser considerado uma representação da causalidade de Granger no domínio da frequência, particularmente estimados a partir de dados de eletroencefalografia (EEG) de paciente diagnosticados com epilepsia do lobo temporal esquerdo, nas bandas de frequência delta, teta e alfa, para tentar encontrar novos biomarcadores baseados nas estruturas e conectividades de ordem superior direcionadas desses dígrafos. Em particular, tentamos responder as seguintes questões: Como a topologia de ordem superior da rede cerebral muda da fase pre-ictal para a ictal, da fase ictal para a pós-ictal, em cada banda de frequência e em cada hemisfério cerebral? A análise de estruturas de ordem superior fornece biomarcadores novos e melhores para a dinâmica das crises e também para a lateralidade do foco da crise do que as análises de grafos usuais?

Encontramos que todas as medidas de caracterização simplicial consideradas no estudo apresentaram aumentos estatisticamente significativos em suas magnitudes da fase pré-ictal para a fase ictal, para várias ordem superiores, para ambos os hemisférios cerebrais, particularmente nas bandas delta e teta, mas nenhuma mudança estatisticamente significativa foi observada da fase ictal para a fase pós-ictal, o que pode sugerir que vários aspectos topológicos e funcionais das redes cerebrais mudam da fase pré-ictal para a fase ictal, em vários níveis de organização topológica de ordem superior. Quanto à lateralidade do foco da crise, a análise baseada em similaridades simpliciais não encontrou diferença estatisticamente significativa entre a topologia de cliques dos hemisférios esquerdo e direito na fase ictal. Concluímos deste estudo que, apesar de uma série de limitações, pode haver evidências que apoiam a viabilidade e confiabilidade do uso de estruturas de ordem superior associadas à dígrafos para identificar biomarcadores associados à redes epilépticas. Entretanto, são necessários mais estudos nessa direção, e a aplicabilidade dos métodos recentemente introduzidos à outros distúrbios de redes de conectividade cerebral também dependerá de estudos futuros.

Palavras-chave: Teoria dos Grafos; Complexos de Cliques Direcionados; Complexos de Caminhos; Conectividade de Ordem Superior Direcionada; Coerência Parcial Direcionada; Conectividade Cerebral; Eletroencefalografia; Epilepsia.

List of Abbreviations

ADSC	*Abstract Directed Simplicial Complex*
AIC	*Akaike Information Criterion*
ASC	*Abstract Simplicial Complex*
DAG	*Directed Acyclic Graph*
DQC	*Directed Quasi-Clique*
DRE	*Drug-Resistent Epilepsy*
DTF	*Directed Transfer Function*
EEG	*Electroencephalography*
EZ	*Epileptogenic Zone*
fMRI	*Functional Magnetic Resonance Imaging*
GC	*Granger Causality*
GED	*Graph Edit Distance*
gPDC	*Generalized Partial Directed Coherence*
GTA	*Graph Theoretical Analysis*
ICA	*Independent Component Analysis*
ILAE	*International League Against Epilepsy*
iPDC	*Informational Partial Directed Coherence*
MRI	*Magnetic Resonance Imaging*
MTLE	*Mesial Temporal Lobe Epilepsy*
MVAR	*Multivariate Autoregressive Model*
PDC	*Partial Directed Coherence*
TDA	*Topological Data Analysis*
TLE	*Temporal Lobe Epilepsy*
VAR	*Vector Autoregressive Model*

Contents

Chapter 1

Introduction

> *Because we do not understand the brain very well we are constantly tempted to use the latest technology as a model for trying to understand it. In my childhood we were always assured that the brain was a telephone switchboard. ('What else could it be?') I was amused to see that Sherrington, the great British neuroscientist, thought that the brain worked like a telegraph system. Freud often compared the brain to hydraulic and electro-magnetic systems. Leibniz compared it to a mill, and I am told some of the ancient Greeks thought the brain functions like a catapult. At present, obviously, the metaphor is the digital computer.*
>
> — John Searle [232]

The human brain is the most complex biological system known by science, it is believed to contain between 80 and 100 billion neurons and trillions of synaptic connections [103, 140]. Essentially, the brain is a network of nervous cells with intricate connections, and the architecture of this network is the structural substrate for its functioning [241]. Understanding how the brain works is one of the greatest scientific challenges of our time, and a crucial component of this challenge is understanding brain networks.

The structural and functional properties of brain networks are the substrate of many brain processes, and the disruption of these networks may be the cause of many neurological disorders, such as Alzheimer's disease and epilepsy. Brain network studies analyze the structural and functional properties of these networks in different contexts, such as in healthy individuals performing motor, sensory, or cognitive tasks, and in individuals diagnosed with some neurological disease. Recently, these studies have used mathematical methods from graph theory, network science, and computational topology to assist in brain network analysis, and this is the scenario that we will deal with in this text.

1.1 Brain Networks

Brain activity depends on several different brain areas rather than being restricted to one or more particular brain regions, and the interaction among remote-located neuroanatomical structures produces structural and functional brain networks. These networks are characterized by the different types of structures involved in the connections and the different types of connectivity, the latter being classified into three main types [164, 239]: *structural connectivity* (or *anatomical connectivity*), which refers to a set of anatomical connections that link neural elements; *functional connectivity*, which refers to statistical relationships between different and remote-located populations of neurons; and *effective connectivity*, which refers to causal relationships between activated brain regions [108, 109, 143, 240].

The inference of the different types of brain networks is carried out through the application of mathematical and statistical methods, called *connectivity estimators*, to neurophysiological signal data that may be obtained through various different technologies, such as electroencephalography (EEG) [231], functional magnetic resonance imaging (fMRI) [145], magnetoencephalography (MEG) [199], and diffusion tensor imaging (DTI) [191]. In the literature, there are several types of connectivity estimators, for instance, there are directed and non-directed estimators, bivariate and multivariate estimators, time-domain and frequency-domain estimators [53]. Typically, effective and functional connectivity networks are estimated through methods that involve some correlation, covariance, or coherence property between the different time series measured from different cortical areas [94, 95]. For the estimation of functional connectivity, correlation, and coherence are the two most commonly used estimators. Other commonly used estimators are mutual information, transfer entropy, and phase synchronization [206]. For the estimation of effective connectivity, an important class of methods are those based on the concept of *Granger causality* (GC) [124], which is a cause-effect relation idea, where the past values of one time series can predict the current values of another. Possibly the most popular among GC-based connectivity estimators are transfer entropy, Granger causality index (GCI) [43, 116], directed coherence (DC) [225], partial directed coherence (PDC) [17], and directed transfer function (DTF) [153].

Among the estimators mentioned above, of special interest in this work is the PDC. PDC is a directed, model-based, multivariate technique that is used to simultaneously determine the directional influences and spectral properties of the interaction between any pairs of brain signals (e.g., pairs of channels in EEG) given in a multivariate ensemble (e.g., a multivariate autoregressive model (MVAR)) [95, 226]. The PDC is an estimator that is based on the coefficients of an MVAR model (or other multivariate models, such as the vector moving average (VMA) model and the vector autoregressive moving average (VARMA) model [21]), transformed in the frequency-domain, and it can be considered a representation of the concept of GC in the frequency-domain [226].

Additionally, the PDC is able to distinguish between direct and indirect causal flows in the estimated connectivity pattern [95, 226]. For instance, PDC tends not to add "erroneous" causality flows between signals recorded from one structure to another. This property makes PDC suitable to be applied to brain signals. Furthermore, a generalized form of PDC (the *generalized* PDC (gPDC)) was introduced in [22]. In an attempt to understand precisely how the PDC relates to the information flow, Takahashi et al. [259, 261] introduced the *information* PDC (iPDC), a modification of the PDC expression that formalizes the relationship between the PDC and the information flow based on the concept of mutual information rate between two time series.

Mathematically, networks can be represented by *graphs*, which are formed by a set of nodes and a set of links connecting these nodes. Commonly, the terms "networks" and "graphs" are used interchangeably. In the case of brain networks, nodes may represent brain areas or neural elements [290], and the links between them can represent, for example, anatomical connections, statistical dependencies, or causal influences. Therefore, the anatomical and functional organization of the brain can be studied from its mathematical representations as graphs, and this fact is what has led in recent years to the widespread adoption of methods from graph theory and network science in network neuroscience [46, 47, 94, 200, 223, 240, 243].

1.2 Graph Theoretical Analysis of Brain Networks

As mentioned in the previous section, there has been a significant increase in research on brain networks using methods from graph theory and network science, and, commonly, the analyses that are based on methods from these fields are called *graph theoretical analysis* (GTA). Numerous GTA studies have demonstrated that several non-trivial topological and organizational properties [289], such as hierarchical organization [28], clustering, modularity [223, 240], presence of structural and functional network motifs [240, 245], and small-world organization [3, 29, 201, 247] are exhibited by brain networks, depending on the context. Furthermore, recent works on brain network analysis have used concepts from spectral graph theory, such as eigenvectors [1, 2] and the spectrum [73] of the Laplacian matrix, the latter being able to reveal an integrative community structure of the networks.

Additionally, GTA has proven to be an important tool in the search for network-based biomarkers for many aspects of brain functioning, for instance, in identifying structural and functional anomalies in network connection patterns, such as those indicating neurological illnesses or certain cognitive processes [274]. In particular, small-world network topology has been regularly detected in various brain network studies, mainly in data from healthy patients, which has been used as a parameter to differentiate healthy and neuropathological network patterns [5, 29, 52]. The comprehensive

evaluation of GTA investigations of brain networks using fMRI data conducted by Farahani et al. [96] noted that applications of graph theory in human cognition include the identification of biomarkers (fMRI-based biomarkers) for human intelligence [265]), working memory [253], aging brain [96], and behavioral performance in natural environments [213], and applications in brain diseases lie in the discovery of biomarkers for conditions like epilepsy [268], Alzheimer's disease (AD) [71], multiple sclerosis (MS) [173], autism spectrum disorders (ASD) [157], and attention-deficit/hyperactivity disorder (ADHD) [272].

Of particular interest in the study of neuropathologies are the EEG-based biomarkers, due to the advantages of EEG over fMRI, such as portability and low cost. There are several studies that use GTA on EEG data obtained from patients diagnosed with some type of neurological disorder in the search for biomarkers [172, 251]. Among these disorders, we can mention Parkinson's disease (PD) [263], epilepsy [142] and schizophrenia (SCZ) [286].

1.3 Topological Data Analysis of Brain Networks

Together with GTA, concepts from computational topology, such as simplicial complexes, homotopy, homology, Betti numbers, and persistent homology, which, eventually, in the data analysis scenario, have been put together under the umbrella term *topological data analysis* (TDA) [51], have been used in recent investigations of brain networks.

Methods from TDA are suitable for evaluating topological characteristics of topological spaces, in particular discrete topological spaces such as simplicial complexes. Briefly, an *abstract simplicial complex* (which, for the sake of simplicity, we refer to as simplicial complex) is a finite collection of finite sets (called simplices) that is closed under subset inclusion [83], and they can be seen as a generalization of graphs. Simplicial complexes can be constructed from graphs, or directed graphs (digraphs), in several ways, e.g., by considering the cliques of a graph as the simplices of the complex (called *clique complex* or *flag complex*) [4], or by considering the directed cliques of a digraph as directed simplices of a directed simplicial complex (where directed simplices are considered to be ordered sets) (called *directed clique complex* or *directed flag complex*) [175, 181, 218]. Other complexes that can be associated with digraphs are the *path complexes* [130], which can contribute with additional insights into the substructures of digraphs.

One of the reasons for considering these complexes rather than just the network nodes is that network characterization measures cannot always provide us with relevant insights into the network topology, for example, two nodes may have the same clustering coefficient, but the topology of their neighborhoods can differ significantly

[155]. Another point that we can highlight is that methods such as persistent homology can analyze the topological features of a space at different scales and their persistence (or their "lifetime persistence").

Several examples of applications of TDA in network neuroscience are the following: characterizing functional brain networks of patients with ADHD and ASD [167, 168]; utilizing clique topology and persistent homology of clique complexes constructed out of brain networks to assess neural functions and structures [119, 120, 207, 218, 237]; and utilizing persistent homology to detect epileptic seizures [99, 208, 256, 273].

1.4 Epilepsy Studies through Brain Connectivity Networks

Although a number of brain disorders were mentioned in the previous sections, our focus in this study is epilepsy. Epilepsy is one of the most common disorders of the central nervous system (CNS), characterized by recurrent and non-induced seizures [6, 275]. It is also a brain connectivity network disorder, typified by a clear relation between pathological symptoms and aberrant network dynamics [112].

Over the years, studies have consistently shown that, compared to healthy individuals, patients diagnosed with epilepsy present changes in the topology of brain connectivity networks [96, 172, 251], and many of these discoveries come from GTA and TDA performed on these networks. Furthermore, these techniques have aided in understanding how brain network architecture changes during the ictal phase (i.e., during the seizure) and how epileptic networks can be described in terms of their topologies.

Some findings that show how epilepsy and alterations in brain network topology are related, based on different data acquisition techniques and connectivity estimation methods, are: Bernhardt et al. [32] found that patients with temporal lobe epilepsy (TLE) showed changes in the distribution of hubs (important brain regions), increased path lengths, and increased clustering coefficients compared with healthy controls; Liao et al. [169] found that patients with mesial temporal lobe epilepsy (MTLE) showed significantly increased local connectivity and decreased global connectivity compared with healthy individuals; Bonilha et al. [40] found that patients with MTLE showed increased degree, local efficiency, and clustering coefficient, in certain areas compared with healthy controls; Horstmann et al. [142] found that patients with drug-resistant epilepsy (a pharmacoresistant form of epilepsy) showed abnormally regular functional networks compared to healthy individuals.

1.5 About this Thesis

First and foremost, before proceeding further in the text, I would like to warn the reader that some relevant research may not have been discussed or cited. I apologize in advance for the research works that were overlooked.

1.5.1 Objectives and Scientific Relevance

Objectives

The main objectives of this work, in simple terms, are the development of new mathematical methods for analyzing the topology of networks and the application of these methods to the analysis of brain networks. More specifically, we are interested in building new ways of looking at the topology of digraphs; for this aim, we chose to develop quantitative methods based on complexes built out of digraphs (or *digraph-based complexes*) such as path complexes and directed cliques complexes, what we call their "higher-order structures" (or "higher-order topologies," or "simplicial structures") and, in the case of directed clique complexes, also methods based on the interrelationships between the directed cliques, what we call their "(directed) higher-order connectivities," and, ultimately, use these new methods to analyze the topology of digraphs derived from brain connectivity estimators (especially the iPDC estimator), particularly estimated from epileptic individuals, to try to find new network-based biomarkers. We can put this more clearly into two main objectives:

1. To develop rigorously a new quantitative theory for digraph-based complexes (or, as we can consider it, a step towards the formalization of a "quantitative simplicial theory"), with special emphasis on directed higher-order connectivity between directed cliques;

2. To apply the methods of the new theory to epileptic brain networks obtained through iPDC to quantitatively investigate their higher-order topologies and search for new biomarkers based on their directed higher-order structures and connectivities, thus pointing out potential applications of the theory in network neuroscience.

Clinical and Scientific Relevance

The new methods introduced in this work may be helpful to the academic community in several areas involving the study of networks, such as biology, social sciences, computer science, and, in particular, network neuroscience, especially in the study of neurological diseases associated with disorders of brain connectivity, such as epilepsy, Alzheimer's disease, and Parkinson's disease. In the case of epilepsy, for example, they may be

helpful to the community in answering fundamental questions such as: How do the brain connectivity networks change from one seizure phase to another? How do these networks change during a seizure? Is it possible to associate reliable network-based biomarkers with epileptic brain networks and with the laterality of the seizure focus? Furthermore, these methods may be useful in clinical practice, for instance, in assisting epilepsy surgeries that depend on the precise location of the epileptogenic zone.

1.5.2 Outline of the Thesis

This thesis is divided into two parts: the first part deals with the development and formalization of a quantitative theory of digraph-based complexes (objective 1), including the exposition of the necessary basic tools, such as the fundamentals of graph theory; the second part deals with the presentation of the theory of brain connectivity networks, together with its utility in the study of epilepsy, and the application of the methods developed in the first part in the analysis of epileptic brain networks (objective 2). In the following, we present a brief description of the chapters that compose each of these two parts.

Part I - Towards a Quantitative Theory of Digraph-Based Complexes

- **Chapter 2 - Fundamentals of Graph Theory:** In this chapter, we introduce the fundamental concepts of graph theory, starting with some formal definitions involving relations and orders, such as equivalence relations and ordered sets, then move on to introduce concepts related to graphs and digraphs, algebraic and spectral graph theory, graph measures, graph similarities, and finally a brief discussion on random graphs.

- **Chapter 3 - Digraph-Based Complexes and Directed Higher-Order Connectivity:** In this chapter, we present the theory of (abstract) simplicial complexes and directed clique complexes associated with digraphs, along with the case of weighted digraphs, passing through simplicial homology, persistent homology, and combinatorial Laplacians associated with these complexes. Next, we present the concept of path complexes, their homologies, and a brief discussion about combinatorial Laplacians associated with them. Finally, we introduce a new theory related to directed higher-order connectivity between directed cliques. This theory leads to the conception of new concepts such as *directed higher-order adjacencies* (upper and lower adjacencies) and *maximal q-digraphs* (defined as digraphs whose nodes are maximal directed cliques), as well as provides new concepts and formalisms for directed Q-Analysis.

- **Chapter 4 - Quantitative Approaches to Digraph-Based Complexes:** In this chapter, based on graph measures, we introduce new quantifiers for char-

acterizing q-digraphs, and, based on graph similarity comparison methods, we introduce new similarity comparison methods for directed clique complexes and path complexes. The set of all these quantifiers and methods can be seen as the formalization of a simplicial analogue of the quantitative graph theory, that is, a formalization of a "quantitative simplicial theory." Finally, we present some examples with random digraphs.

Part II - Brain Connectivity Networks and a Quantitative Graph/Simplicial Analysis of Epileptic Networks

- **Chapter 5 - Brain Connectivity Networks:** In this chapter, we study the theory behind brain connectivity networks, briefly going over the biophysical principles of brain signals and the methods for obtaining them, paying particular attention to EEG. Next, we discuss, also briefly, bivariate and multivariate connectivity estimators, focusing on PDC and its variants, particularly gPDC and iPDC. Finally, we cover the various kinds of brain connectivity networks, particularly structural, functional, and effective networks, as well as the modern uses of graph theory and computational (algebraic) topology in the analysis of these networks to explore the dynamics of brain activity in various contexts, especially in the study of neurological disorders.

- **Chapter 6 - Epilepsy as a Disorder of Brain Connectivity:** In this chapter, we look more closely at the neuropathology of epilepsy, covering its main characteristics, etiologies, epidemics, diagnoses, and treatments. We also discuss a number of studies that showed differences between the brain networks of epileptic patients and brain networks of healthy people in terms of network properties, such as clustering coefficient, characteristics path length, and degree distribution, as well as differences between the ictal and non-ictal periods.

- **Chapter 7 - Quantitative Graph/Simplicial Analysis of Epileptic Networks:** In this chapter, we perform an analysis of epileptic brain networks, estimated through iPDC from EEG data from patients with left temporal lobe epilepsy, using the new quantitative methods developed in previous chapters for directed higher-order networks (q-digraphs), to explore how certain properties of these networks change according to the seizure phases, as well as according to the cerebral hemispheres, in different frequency bands.

- **Chapter 8 - Final Considerations:** Finally, in this last chapter, we present a summary of the objectives and developments of this thesis, some relevant considerations, and the gaps left in this work that we intend to complete in future studies.

Part I

Towards a Quantitative Theory of Digraph-Based Complexes

Chapter 2

Fundamentals of Graph Theory

> *Graph theory serves as a mathematical model for any system involving a binary relation.*
>
> — Frank Harary [137]

Bearing in mind the multidisciplinary nature of this thesis and to make our discussion self-contained, we have chosen to offer enough background information so that readers with an undergraduate mathematics background may follow along. Accordingly, in this second chapter, we present the basic concepts, terminology, and notations of graph theory that will be necessary throughout the text.

2.1 Fundamental Concepts

Graphs are the central objects of this thesis because, as we made explicit in the introduction, we will deal with how different areas of the brain interact with each other, and these interactions can be represented abstractly through *graphs*, which are mathematical representations of the relationships between units of a system. Besides neuroscience, graph theory has numerous applications in a variety of other scientific fields, such as social science, biology, computer science, linguistics, and transportation planning [57, 195].

Some authors use the term "networks" to designate real-world networks and "complex networks" to designate large real-world networks, whereas the term "graphs" is used to designate their mathematical representations. However, here we will not make these distinctions, thus, henceforth, we will use these terms interchangeably.

In this section, we present the mathematical formalism of the areas of graph theory that will be necessary for the development of the text. Some of the propositions and theorems presented here do not include their respective proofs, but instead, we provide appropriate references to their proofs.

Henceforth, we will use the Bourbaki notation for the sets of natural, integer, and

real numbers, $\mathbb{N}$ ($0 \notin \mathbb{N}$), $\mathbb{Z}$, and $\mathbb{R}$, respectively. For these sets, some additional notations are adopted: $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$, and $\mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$. Also, we may use the Dirac notation to denote vectors in the Euclidian space, e.g. $|v\rangle \in \mathbb{R}^n$, and $\langle v| = |v\rangle^T$.

2.1.1 Relations and Orders

Let's start with a basic but crucial distinction between unordered and ordered pairs of elements. The primary reference for this section is [148].

The axiom of extensionality of Zermelo-Fraenkel axiomatic set theory states that if two sets have the same elements, then they are equal. From this axiom, we have $\{x, y\} = \{y, x\}$, for any set of two elements (or *unordered pair*). We define an *ordered pair* (x, y) so as to satisfy the condition: $(x, y) = (z, w)$ if and only if $x = z$ and $y = w$. In general, we define an *ordered n-tuple* $(x_1, ..., x_n)$ so as to satisfy the condition: $(x_1, ..., x_n) = (y_1, ..., y_n)$ if and only if $x_i = y_i$, for all $i = 1, ..., n$.

As a usual convention, braces $\{\}$ are used to denote sets and parentheses $()$ to denote ordered n-tuples, however, in some cases we may commit an abuse of notation and use $()$ to denote unordered pairs, as we will see in the definition of undirected graphs.

Definition 2.1.1. Let $S_1, ..., S_n$ be n sets, not necessarily distinct. A *relation* (or *n-ary relation*) on these sets is a set $R \subseteq S_1 \times ... \times S_n$ whose elements are n-tuples $(x_1, ..., x_n)$ such that $x_i \in S_i$, $\forall i = 1, .., n$. We call R a *binary relation* if $n = 2$, and, in this case, we denote a pair $(x_1, x_2) \in R$ as $x_1 R x_2$.

Definition 2.1.2. Let $\sim \subseteq S \times S$ be a binary relation on a set S. We say that $\sim$ is an *equivalence relation* on S if it satisfies the following conditions, for all $x, y, z \in S$:

1. $x \sim x$ (reflexivity);
2. $x \sim y \iff y \sim x$ (symmetry);
3. if $x \sim y$ and $y \sim z \Rightarrow x \sim z$ (transitivity).

An equivalence relation $\sim$ on S defines a *class of equivalence* for each element $x \in S$: $[x] = \{y \in S : x \sim y\}$. The classes of equivalence of S define a set called *quotient set*, and it is denoted as $S/\sim$, i.e. $S/\sim = \{[x] : x \in S\}$.

Definition 2.1.3. Given a set S, an *order* (or *partial order*) defined in S is a binary relation, denoted by $\leq$, which satisfies the following conditions, for all $x, y, z \in S$:

1. $x \leq x$ (reflexivity);
2. if $x \leq y$ and $y \leq x \Rightarrow x = y$ (antisymmetry);

3. if $x \leq y$ and $y \leq z \Rightarrow x \leq z$ (transitivity).

Moreover, if either $x \leq y$ or $y \leq x$, $\forall x, y \in S$, we said that $\leq$ is a *total order* (or *linear order*).

A set equipped with an order (partial order) is called a *ordered set* (*partially ordered set* or *poset*), and a set equipped with a total order is called a *totally ordered set* (or *linearly ordered set*).

It's important to note that if two partially/totally ordered sets have the same elements, they are identical as sets, but they are not identical as partially/totally ordered sets if the respective orders of their elements are different.

As a last consideration, since a set does not allow repetitions of its elements, i.e. $\{x, x\} = \{x\}$, and since sometimes it is necessary to take these repetitions into account, below we present the formal definition of *multiset*, which is a generalization of the concept of set [34].

Definition 2.1.4. A *multiset* is a collection of elements in which repetition of elements is allowed. The *multiplicity* of an element is the number of times it appears in the multiset. The cardinality of a multiset is equal to the total number of its elements, counting their multiplicities.

To avoid confusion, we will adopt the notation $\{\!\{\}\!\}$ to denote multisets.

Example 2.1.1. Consider the multiset $M = \{\!\{x, x, y, y, y\}\!\}$. The multiplicity of x is equal to 2, the multiplicity of y is equal to 3, and the cardinality of M is equal to 5.

2.1.2 General Concepts in Graph Theory

In this subsection, we present a brief introduction to the main concepts of graph theory that will be necessary for the development of subsequent concepts. For this first part, the basic references are [36, 79, 137, 278], however, additional references will be presented throughout the text.

Graphs: Basic Definitions

Definition 2.1.5. A *graph* is a pair $G = (V, E)$ of disjoint sets such that $E \subseteq V \times V$ is a binary relation, i.e. the elements of E are unordered pairs of elements of V. The elements of V are the *vertices* (or *nodes*) and the elements of E are the *edges* (or *links*) of the graph G. The edges are denoted as (v, u) $(= (u, v))$, where $v, u \in V$. The notations $V = V(G)$ and $E = E(G)$ are also commonly used. The cardinalities of V and E are denoted by $|V|$ and $|E|$, respectively, and $|V|$ is said to be the *order* of G. If $|V| < \infty$ and $|E| < \infty$, the graph is said to be *finite*.

Strictly speaking, in the previous definition we defined an *undirected graph*, whose edges are actually sets $(v, u) = \{v, u\}$, where we are committing an abuse of notation by using parentheses. As a convention, we will use parentheses whenever we deal with edges of graphs.

It is usual to represent graphs through diagrams of points (vertices) connected by lines (edges) (see Figure 2.1).

Definition 2.1.6. Given a graph $G = (V, E)$, a vertex $x \in V$ is said to be *incident with an edge* $e = (v, u) \in E$ if $x \in e$, i.e. if $x = v$ or $x = u$ (we can also say that the edge e is *incident to* x). Furthermore, two edges are said to be *adjacent* if they share a common vertex, and two vertices, $v, u \in V$, are said to be *adjacent* (or *neighbors*) if the edge (v, u) exists.

Some authors define the *empty graph* as the graph having at least one vertex but no edges and the *null graph* as the graph having no vertices and no edges. Here we will adopt this convention.

Definition 2.1.7. A *loop* (or *self-loop*) in a graph $G = (V, E)$ is an edge $(v, v) \in E$, i.e. it is an edge that links a vertex $v \in V$ to itself. If there is more than one edge between two vertices we say that G contains *multiple edges*. If a graph contains no loops or multiple edges, it is said to be a *simple graph*. A graph that allows multiple edges and loops is called *pseudograph*[1] and a pseudograph without loops is called *multigraph*.

We can make the vertices of a graph distinguishable from one another by associating names or *labels* with each of them. More formally, a non-null graph is said to be *labeled* if it is equipped with a bijection between the finite set of vertices and a finite set of labels. For instance, for a graph with $k + 1$ vertices, we can assign labels such as $v_0, ..., v_k$, or non-negative integers $0, 1, ..., k$, to its vertices.

Henceforth, all graphs will be considered non-empty, non-null, finite, simple, and labeled, unless said otherwise.

Definition 2.1.8. The graph $G' = (V', E')$ is a *subgraph* of $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$, and in this case we write $G' \subseteq G$. If $G' \subseteq G$ and $G' \neq G$, then G' is called a *proper subgraph* of G. If G' contains all edges $(v, u) \in E$, with $v, u \in V'$, then G' is said to be an *induced subgraph* of G, and we say that V' *induces* (or *generates*) G' into G.

Typically, in network science literature, subgraphs that appear at a significantly higher frequency in a given graph than in equivalent random graphs are called *motifs* [186].

[1] Strictly speaking, a pseudograph is not a graph since, by definition, a graph has no loops, but for practical reasons we use this terminology.

Definition 2.1.9. A *path* is a graph $P = (V, E)$, $V = \{v_0, ..., v_k\}$, such that $E = \{e_1 = (v_0, v_1), ..., e_k = (v_{k-1}, v_k)\}$, with $v_i \neq v_j$ if $i \neq j$, for all $i, j = 0, ..., k$, and $e_m \neq e_n$ if $m \neq n$, for all $m, n = 1..., k$. Also, we can denote a path simply by the sequence of its vertices, i.e. $P = v_0v_1...v_k$.

Definition 2.1.10. A *cycle* is a graph $C = (V, E)$, $V = \{v_0, ..., v_k\}$, with $k \geq 3$, such that $E = \{e_1 = (v_0, v_1), ..., e_k = (v_{k-1}, v_k), e_{k+1} = (v_k, v_0)\}$, with $v_i \neq v_j$ if $i \neq j$, for all $i, j = 0, ..., k$, and $e_m \neq e_n$ if $n \neq m$, for all $m, n = 1..., k + 1$.

Definition 2.1.11. A *walk* between two vertices v_0 and v_k (or (v_0, v_k)-*walk*) is a sequence of vertices and edges (not necessarily distinct) $v_0e_1v_1...v_{k-1}e_kv_k$ such that $e_i = (v_{i-1}, v_i)$, for all $1 \leq i \leq k$. If $v_0 = v_k$, the walk is called *closed*, and is called *open* otherwise. If the edges of a walk are all distinct, it is said to be a *trail*. If the vertices of a walk are all distinct (and consequently all of the edges), then it is said to be a *path*. A closed walk with all distinct vertices and with $k \geq 3$ is a *cycle*. The *length* of a walk is equal to its number of edges. Also, a walk of length k is said to be a *k-walk*.

It's important to highlight that, as sequences, paths, and cycles are special cases of trails, and trails are special cases of walks but paths and cycles form actually simple graphs. In a simple graph, it's common to denote a walk by the sequence of its vertices, i.e. $W = v_0v_1...v_k$, $v_i \in V$, $i = 0, 1, ..., k$.

Definition 2.1.12. Given a graph $G = (V, E)$, two vertices $v, u \in V$ are said to be *connected* if either $v = u$ or there exists a path connecting v to u. Also, if every vertex in G is connected to each other, G is said to be *connected*. If G is not connected, it is said to be *disconnected*.

Proposition 2.1.1. *The property of two vertices of a graph being connected is an equivalence relation on its vertex set.*

Proof. Let $G = (V, E)$ be a graph and let $\sim$ denote the relation "is connected to" on the vertex set V. For arbitrary vertices $x, y, z \in V$, the following properties are satisfied:

1. $x \sim x$. Indeed, by Definition 2.1.12, every vertex is connected to itself.
2. $x \sim y \Rightarrow y \sim x$. Indeed, if $x \sim y$, there is a path connecting y to x as well.
3. $x \sim y$ and $y \sim z \Rightarrow x \sim z$. Indeed, let $P_{xy} = xv_0...v_ky$ be the path connecting x to y and let $P_{yz} = yv_{k+1}...v_{k+n}z$ be the path connecting y to z, $v_i \in V$. The path $P_{xz} = xv_0...v_kyv_{k+1}...v_{k+n}z$ is a path connecting x to z, thus $x \sim z$.

Therefore, the relation $\sim$ is an equivalence relation since it is reflexive, symmetric, and transitive. □

Definition 2.1.13. A *maximal connected subgraph* of a graph G is a connected subgraph which is not a proper subgraph of any other connected subgraph of G. A *connected component* of G is a maximal connected subgraph of G. The largest connected component of G is called its *giant component*.

Definition 2.1.14. Given a graph $G = (V, E)$, with $|V| = n$ and $|E| = m$, its *density* is defined as $\text{den}(G) = 2m/n(n-1)$, since the maximum number of edges in the graph is equal to $\binom{n}{2} = n(n-1)/2$. As a convention, a graph is said to be *dense* if $\text{den}(G) > 0.5$, and it is said to be *sparse* otherwise.

Definition 2.1.15. Given a graph $G = (V, E)$, the *neighborhood* of a vertice v in G, denoted by $\mathcal{N}_G(v)$, is the set of all vertices that are adjacent to v, i.e.

$$\mathcal{N}_G(v) = \{u \in V : (v, u) \in E\}. \tag{2.1}$$

Moreover, we say that $\mathcal{N}_G(v)$ is an *open neighborhood* if $v \notin \mathcal{N}_G(v)$, and that it is a *closed neighborhood* otherwise. In the last case, we use the notation $\mathcal{N}_G[v]$.

Definition 2.1.16. Given a graph $G = (V, E)$, the *degree* of a vertex $v \in V$, denoted by $\deg_G(v) = \deg(v)$, is equal to the number of edges incident to v. If $\deg(v) = 0$, the vertex is said to be *isolated*.

In general, in network science literature, if the degree of a vertex far exceeds the average degree of the other nodes in a graph, it is called a *hub*[2].

Now that we have introduced the definition of vertex degree, let's introduce one of the most fundamental properties of a graph: its *degree distribution* [195]. Let $G = (V, E)$ be a graph with $|V| = n$. Let $\delta(k)$ be the number of vertices having degree k. The probability that a vertex chosen uniformly at random has a degree equal to k is given by

$$p(k) = \frac{\delta(k)}{n}. \tag{2.2}$$

The fractions $p(k)$ represent the *degree distribution* of the graph, and they describe how frequently a vertex with a certain degree appears in the graph. Furthermore, we can represent the degree distribution of G graphically as the plot of $p(k)$ versus k.

Definition 2.1.17. Given a graph G, if all its vertices are adjacent to each other, G is said to be *complete*. A complete graph with n vertices is commonly denoted by K_n.

The complete graph K_n has $n(n-1)/2$ edges. The graph K_3 is called *triangle*. Figure 2.1 illustrates the complete graphs K_n, for $n = 1, 2, 3, 4, 5$.

[2]Other definitions of hub were proposed either based on other node-wise measures, such as betweenness centrality and clustering coefficient [46, 244], or based on the connection with the concept of *authorities* [158].

Definition 2.1.18. Let $G = (V, E)$ be a graph. A $(k+1)$-*clique* in G is a complete induced subgraph with $k+1$ vertices, $0 \leq k \leq |V| - 1$. A clique is said to be *maximal* if it is not a proper subgraph of any other clique in G. Also, the *clique number* of G, denoted by $\omega(G)$, is the number of vertices contained in the largest clique of G.

Example 2.1.2. Figure 2.1 shows examples of $(k+1)$-cliques, for $k = 0, 1, 2, 3, 4$. From left to right: 1-clique (vertex), 2-clique (edge), 3-clique (triangle), 4-clique, 5-clique. The $(k+1)$-clique is the complete graph $K_{(k+1)}$.

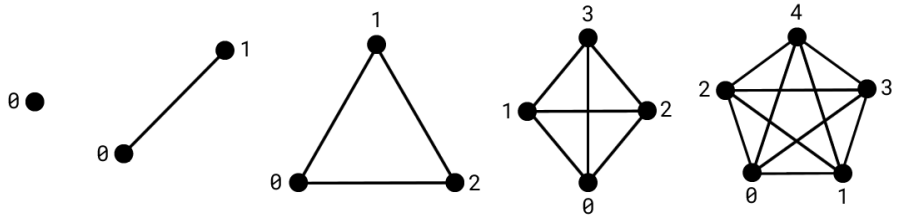


Figure 2.1: Examples of $(k+1)$-cliques, for $k = 0, 1, 2, 3, 4$.

The term "clique" originated in the study of social networks to denote the formation of a group of two or more people if the condition of being mutual friends is satisfied [174], and later was adopted to denote a complete induced subgraph. The process of clique enumeration (finding and listing all cliques in a graph) is very useful and widely used in network analysis since cliques might represent functional units within real-world networks, but the problem of finding cliques is NP-complete [64].

A natural generalization of cliques is the *quasi-cliques*, which are *almost complete* subgraphs. In the literature, there are slightly different ways to define them [44, 204], and here we will use the definition of *degree-based quasi-clique* [227].

Definition 2.1.19. Given a graph $G = (V, E)$, a subgraph $H = (V', E') \subseteq G$, with $|V'| = m$, is called a γ-*quasi-clique*, for a parameter $0 < \gamma \leq 1$, if $\deg_H(v) \geq \gamma(m-1)$, for all $v \in V'$.

Note that if $\gamma = 1$, the γ-quasi-clique is actually a clique.

Directed and Weighted Graphs

Many of the definitions and results presented here can be found in [25, 137].

Definition 2.1.20. A *directed graph* (or *digraph*) is a pair $G = (V, E)$ of disjoint sets such that the elements of E are ordered pairs of elements of V (set of vertices). The elements of E are called *directed edges*, (or *arcs*, or *arrows*), and are denoted by (v, u) $(\neq (u, v))$, $v, u \in V$. The first vertex v of an arc (v, u) is called *tail* and the second vertex u is called *head*, and we say that the *direction* of (v, u) is from v to u (or from

its tail to its head). Also, we say that a vertex u *arrives* at a vertex v if the arc (u, v) exists (equivalently, (u, v) *arrives* at v), and that a vertex w *leaves* v if the arc (v, w) exists (equivalently, (v, w) *leaves* v).

Differently from undirected edges, if we change the order of the vertices in an arc, we obtain an arc in the opposite direction. Moreover, in a digraph, we can have two arcs between the same two vertices, but with opposite directions, and when this occurs we say that the digraph has a *bidirectional edge* (or *double-edge*).

Most of the definitions made for undirected graphs in the previous section are straightforwardly extended to digraphs. For instance, we say that a digraph is *simple* if it has no directed loops (arcs in which their tails coincide with their heads) and no multiple arcs (more than one arc with the same tail and head). The Definition 2.1.8 of subgraphs in a graph is the same as for a digraph, with the difference that now we are dealing with *subdigraphs* in a digraph. The definitions of walk, trail, path, and cycle are easily extended for the directed case, as we will see in the next definition.

Henceforth, all digraphs will be considered non-empty, non-null, finite, simple, and labeled, unless said otherwise.

Definition 2.1.21. A *directed walk from* v_0 *to* v_k (or *directed* (v_0, v_k)*-walk*) is a sequence of vertices and arcs (not necessarily distinct) $v_0e_1v_1...v_{k-1}e_kv_k$ such that $e_i = (v_{i-1}, v_i)$, i.e., v_{i-1} is the tail and v_i is the head of the arc e_i, for all $1 \leq i \leq k$. If $v_0 = v_k$, the directed walk is called *closed*, and is called *open* otherwise. If the arcs of a directed walk are all distinct, it is said to be a *directed trail*. If the vertices of a directed walk are all distinct (and consequently all of the arcs), then it is said to be a *directed path*. A closed directed walk with all distinct vertices and with $k \geq 2$ is a *directed cycle*. Also, the *length* of a directed walk is equal to its number of arcs.

Note that a directed cycle of length 2 is actually a double-edge.

Definition 2.1.22. Given a digraph $G = (V, E)$, the *underlying undirected graph* of G is the undirected graph, with the same set of vertices V, formed by replacing all directed edges in E with undirected edges.

Definition 2.1.23. A *directed acyclic graph* (DAG), or *acyclic digraph*, is a digraph that has no directed cycles.

A notable property of (finite) DAGs is that they have at least one source and at least one sink (see [25], p. 32, for proof).

Definition 2.1.24. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are said to be *isomorphic* if there is a bijection $f : V_1 \to V_2$ such that if the vertices $v, u \in V_1$ are adjacent, then the vertices $f(v)$ and $f(u)$ are adjacent in V_2 and vice versa, i.e. if and only if the bijection f preserves adjacencies. Likewise, if G_1 and G_2 are digraphs, f is an isomorphism if and only if (u, v) is an arc in V_1 then $(f(u), f(v))$ is an arc in V_2.

Graph properties that are invariant under graph/digraph isomorphism are called *graph invariants*. For example, the order and the number of edges/arcs of a graph/digraph are graph invariants [79, 137].

Definition 2.1.25. Given a digraph $G = (V, E)$, a vertex $v \in V$ is said to be *reachable* from a vertex $u \in V$ if either $v = u$ or there exists a directed path from u to v in G.

A difference between undirected and directed graphs is that, for digraphs, we have two different concepts of connectivity: *weak connectivity* and *strong connectivity*.

Definition 2.1.26. Let $G = (V, E)$ be a digraph. A vertex $v \in V$ is said to be *weakly connected* to another vertex $u \in V$ if there is an undirected path between v and u in the underlying undirected graph of G. We say that G is *weakly connected* if every vertex in G is weakly connected to each other. A *weakly connected component* of G is a maximal subdigraph that is weakly connected. The largest connected component of G is called its giant component. Analogously to the undirected case, the largest weakly connected component of G is called its *giant component*.

Definition 2.1.27. Let $G = (V, E)$ be a digraph. A vertex $v \in V$ is said to be *strongly connected* to another vertex $u \in V$ if u is reachable from v and v is reachable from u in G. We say that G is *strongly connected* if every vertex in G is strongly connected to each other. A *strongly connected component* of G is a maximal subdigraph that is strongly connected.

Proposition 2.1.2. *In a digraph, weak connectivity and strong connectivity are both equivalence relations on its vertex set.*

Proof. Let $G = (V, E)$ be a digraph. Let $\sim_w$ denote the relation "is weakly connected to" and let $\sim$ denote the relation "is connected to" in the underlying undirected graph of G, on the vertex set V. For arbitrary vertices $v, u \in V$, by definition, $v \sim_w u \Leftrightarrow v \sim u$. Since $\sim$ is an equivalence relation, $\sim_w$ is also an equivalence relation on V.

Now, denote by $\sim_s$ the relation "is strongly connected to" on V. For arbitrary vertices $v, u, w \in V$, by Definition 2.1.25, every vertex is reachable from itself (reflexivity), if $v \sim_s u$, then v is reachable from u and u is reachable from $v \Rightarrow u \sim_s v$ (symmetry), and if $v \sim_s u$ and $u \sim_s w$, similarly to the argument used in the proof of Proposition 2.1.1, there exists a directed path from v to w and a directed path from w to v, thus $v \sim_s w$ (transitivity). □

Another difference between undirected and directed graphs is that the density of a digraph $G = (V, E)$, with $|V| = n$ and $|E| = m$, is $\text{den}(G) = m/n(n-1)$, since the maximum number of arcs in the digraph is equal to $n(n-1)$ (see Definition 2.1.14).

Definition 2.1.28. Given a digraph $G = (V, E)$, the *in-neighborhood* of a vertex v, denoted by $\mathcal{N}_G^{in}(v)$ or $\mathcal{N}_G^{-}(v)$, is the set of all vertices that arrive at v, and the *out-neighborhood* of v, denoted by $\mathcal{N}_G^{out}(v)$ or $\mathcal{N}_G^{+}(v)$, is the set of all vertices that leave v, i.e.

$$\mathcal{N}_G^{in}(v) = \mathcal{N}_G^{-}(v) = \{u \in V : (u, v) \in E\}, \tag{2.3}$$

$$\mathcal{N}_G^{out}(v) = \mathcal{N}_G^{+}(v) = \{u \in V : (v, u) \in E\}. \tag{2.4}$$

Note that if a digraph G have no double-edges, the neighborhood of a vertex v in the underlying undirected graph is $\mathcal{N}_G(v) = \mathcal{N}_G^{in}(v) \cup \mathcal{N}_G^{out}(v)$.

Definition 2.1.29. Let $G = (V, E)$ be a digraph. The number of arcs arriving at a vertex $v \in V$ is its *in-degree*, denoted by $\deg^{in}(v) = \deg^{-}(v)$, and the number of arcs leaving v is its *out-degree*, denoted by $\deg^{out}(v) = \deg^{+}(v)$. The *total degree* of v, denoted by $\deg^{tot}(v)$, is the sum of its in-degree and its out-degree.

Example 2.1.3. Figure 2.2a exemplifies the digraph $G = (V = \{0, 1, 2, 3\}, E = \{(1, 0), (2, 1), (0, 2), (3, 2), (1, 3)\})$ in which $\deg^{in}(1) = 1$, $\deg^{out}(1) = 2$, and $\deg^{tot}(1) = 1 + 2 = 3$.

Furthermore, since a digraph may have double-edges, let's denote by $\deg^{\pm}(v)$ the number of double-edges incident to v. It is clear that the degree of v in the underlying undirected graph is given by $\deg(v) = \deg^{-}(v) + \deg^{+}(v) - \deg^{\pm}(v)$.

Analogously to the undirected case, for a directed graph $G = (V, E)$ with $|V| = n$, if $\delta^{in}(k)$ is the number of vertices having in-degree k, the probability that a vertex chosen uniformly at random has in-degree equal to k is given by

$$p^{in}(k) = \frac{\delta^{in}(k)}{n}. \tag{2.5}$$

Similarly, the probability that a vertex chosen uniformly at random has out-degree equal to k is given by

$$p^{out}(k) = \frac{\delta^{out}(k)}{n}. \tag{2.6}$$

The fractions $p^{in}(k)$ and $p^{out}(k)$ represent the *in-degree distribution* and the *out-degree distribution*, respectively, of the digraph.

Before introducing the concept of weight, we need to introduce the concepts of *metric*, *quasi-metric*, and *pre-metric*, which will be useful in the next sections as well.

Definition 2.1.30. Let X be a set. A *metric* (or *distance*) on X is a function $d : X \times X \to \mathbb{R}_{\geq 0}$ satisfying the following conditions, for all $x, y, z \in X$:

1. $d(x, y) = 0 \iff x = y$ (identity);

2. $d(x, y) = d(y, x)$ (symmetry);

3. $d(x, y) \leq d(x, z) + d(z, y)$ (triangular inequality).

We say that d is a *quasi-metric* (or *quasi-distance*) if d does not necessarily satisfy the symmetry property; d is called a *semi-metric* (or *semi-distance*) if d does not necessarily satisfy the triangular inequality; if d does not necessarily satisfy both conditions, symmetry and triangular inequality, then d is called a *pre-metric* (or *pre-distance*). Given a (quasi/semi/pre-) metric d, the pair (X, d) is called a (quasi/semi/pre-) *metric space*.

In the following, weights in the set $\mathbb{R}_{\geq 0}$ are taken into consideration for defining weighted graphs (digraphs); alternatively, a different set of numbers may be used.

Definition 2.1.31. A *weighted graph* (*digraph*) is a triple $G^{\omega} = (V, E, \omega)$, where $G = (V, E)$ is a graph (digraph), and $\omega : V \times V \to \mathbb{R}_{\geq 0}$ is a pre-metric. The real number $\omega(v, u) = \omega_{vu}$ is the *weight* of the edge $(v, u) \in E$.

Example 2.1.4. Figure 2.2b exemplifies a weighted undirected graph with four vertices in which the edge thicknesses represent the weights that satisfy the relation $\omega_{02} < \omega_{23} < \omega_{13} < \omega_{01} < \omega_{12}$.

(a) A directed graph. (b) A weighted graph.

Figure 2.2: Examples of directed and weighted graphs (the weights ω_{ij} are visually represented by the thicknesses of the edges).

Defining the weight function ω as a pre-metric is suitable for the general case, since if G is a digraph, we may have $\omega_{vu} \neq \omega_{uv}$, for some vertices $v, u \in G$.

Definition 2.1.32. Let G^{ω} be a weighted graph. The *weighted degree* of a vertex v is the sum of all weights associated with the edges to which v is incident with, i.e. $\deg_{\omega}(v) = \sum_{u \in \mathcal{N}(v)} \omega(v, u)$. Analogously, if G^{ω} is a weighted digraph, the *weighted in-degree* of a vertex v is $\deg^{-}_{\omega}(v) = \sum_{u \in \mathcal{N}^{-}(v)} \omega(u, v)$, and its *weighted out-degree* is $\deg^{+}_{\omega}(v) = \sum_{u \in \mathcal{N}^{+}(v)} \omega(v, u)$.

To define a distance in a weighted graph, we need a *weight-to-distance* conversion function. However, for digraphs, the symmetry property is not necessarily satisfied, therefore, it would be more suitable to define a pre-distance.

Definition 2.1.33. Given a weighted graph (or digraph) $G^\omega = (V, E, \omega)$ with a (normalized) weight function $\omega : V \times V \to [0, 1]$, let $D^\omega : \mathrm{Im}(\omega) \subseteq [0, 1] \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ be a function defined by

$$D^\omega(\omega_{vu}) = \begin{cases} \omega_{vu}^{-1} - 1, \text{ if } v \neq u; \\ +\infty, \text{ if } \omega_{vu} = 0; \\ 0, \text{ if } v = u. \end{cases} \tag{2.7}$$

The function D^ω is a pre-distance.

It is clear that D^ω produces small values when applied to large weights and vice versa, thus we can say that two vertices are "closer" when the weight of the connection between them is greater. Although D^ω is actually a pre-distance, from now on we will adopt an abuse of notation and call it a distance.

Observation 2.1.1. When the weights are not normalized and take values ≥ 1, we can define a weight-to-distance function by modifying the formula of D^ω by replacing the expression $(\omega_{vu}^{-1} - 1)$ with the expression ω_{vu}^{-1}.

2.2 Algebraic and Spectral Graph Theory

Algebraic graph theory is concerned with translating properties of graphs into algebraic properties, such as matrices and groups, in such a way that concepts from abstract and linear algebra can be applied. An important branch of algebraic graph theory is the *spectral graph theory*, which studies the properties of graphs through eigenvalues, eigenvectors, and other spectra-related concepts of graph matrices.

The fundamental bibliography for this section is [31, 33, 58]. Undergraduate knowledge of matrix theory is assumed.

2.2.1 Algebraic Graph Theory

Let's start with the definition of the most common matrix representation of a graph: the *adjacency matrix*.

Definition 2.2.1. Let $G = (V, E)$ be a graph with $|V| = n$. The *adjacency matrix* of G is the $n \times n$ matrix $A = A(G) = (a_{ij})$ whose entries are given by

$$a_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E, \text{ for } v_i, v_j \in V, i \neq j; \\ 0, & \text{otherwise.} \end{cases} \tag{2.8}$$

Note that, by definition, A is a binary matrix (formed by 0's and 1's), symmetric, and with trace $\mathrm{Tr}(A) = \sum a_{ii} = 0$. In the case where G is a weighted graph, the entries

of A will be equal to the edge weights, i.e. $a_{ij} = \omega_{ij}$. On the other hand, if G is a digraph, then the matrix A might be asymmetric, since we might have $a_{ij} \neq a_{ji}$, since $a_{ij} = 1$ if and only if $(v_i, v_j) \in E$, and if there is no connection in the opposite direction, i.e. if $(v_j, v_i) \notin E$, then $a_{ji} = 0$.

Example 2.2.1. Consider the complete graph K_4, the digraph G represented in Figure 2.2a, and the weighted graph G^ω represented in Figure 2.2b. The respective adjacency matrices of these graphs are:

$$A(K_4) = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}, \quad A(G) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \text{ and } A(G^\omega) = \begin{bmatrix} 0 & \omega_{01} & \omega_{02} & 0 \\ \omega_{10} & 0 & \omega_{12} & \omega_{13} \\ \omega_{20} & \omega_{21} & 0 & \omega_{23} \\ 0 & \omega_{31} & \omega_{32} & 0 \end{bmatrix}.$$

Definition 2.2.2. Let $G = (V, E)$ be a graph with $|V| = \{v_0, ..., v_{n-1}\}$ and $|E| = \{e_0, ..., e_{m-1}\}$. The (vertex-edge) *unoriented incidence matrix* of G is the $n \times m$ matrix $B = B(G) = (b_{ij})$ whose entries are given by

$$b_{ij} = \begin{cases} 1, & \text{if } v_i \text{ is incident with } e_j; \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

If G is a directed graph, the entries b_{ij} of its (vertex-arc) *oriented incidence matrix* are given by

$$b_{ij} = \begin{cases} 1, & \text{if } v_i \text{ is the tail of the arc } e_j; \\ -1, & \text{if } v_i \text{ is the head of the arc } e_j; \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

We can associate an oriented incidence matrix (2.10) to an undirected graph by considering arbitrary directions on its edges. The following proposition shows how to obtain the number of walks of certain size between two vertices in an undirected or directed graph from the powers of its adjacency matrix.

Proposition 2.2.1. *Let $G = (V, E)$ be a graph. For a non-negative integer k, the (i, j)-entries of the k-th power of its adjacency matrix, A^k, is equal to the number of (v_i, v_j)-walks of length k in G, $v_i, v_j \in V$. Analogously, if G is a digraph, the (i, j)-entries of A^k is the number of directed (v_i, v_j)-walks of length k in the digraph.*

A simple proof (by induction in k) of the previous proposition for the undirected case (but straightforwardly extended to the directed case) can be found in [33], p. 9.

Corollary 2.2.1. *Let G be a graph with adjacency matrix A. For some non-negative integer k, the trace of the k-th power of A, $\mathrm{Tr}(A^k)$, counts the number of closed k-walks in G. Analogously, if G is a digraph, $\mathrm{Tr}(A^k)$ counts the number of closed directed k-walks in G.*

Definition 2.2.3. Given an undirected or directed graph G, let B be its oriented incidence matrix. The *Laplacian matrix* of G is defined by

$$L = BB^T. \tag{2.11}$$

Note that whether G is an undirected or directed graph, its Laplacian matrix $L(G)$ is symmetric ($BB^T = (BB^T)^T$). Be aware that there are other alternative definitions for symmetric and non-symmetric versions of the Laplacian of a digraph since the above definition is *independent* of the directions of the edges, but in this work we will adopt this definition.

Definition 2.2.4. Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix. M is said to be *positive semi-definite* if $v^T Mv \geq 0$, for all $v \in \mathbb{R}^n$.

Proposition 2.2.2. *The Laplacian matrix of a given undirected or directed graph is positive semi-definite.*

Proof. Let $G = (V, E)$ be an undirected or directed graph with $|V| = n$, and let L be its Laplacian matrix. For any real-valued vector $v \in \mathbb{R}^n$, we have

$$v^T Lv = v^T BB^T v = (B^T v)^T (B^T v) = ||B^T v||^2 \geq 0, \tag{2.12}$$

where $||.||$ is the Euclidean norm. □

2.2.2 Spectral Graph Theory

As previously stated, the spectral graph theory deals with the study of graphs through the spectra of their matrices; therefore, let's start by defining the components of these spectra: their *eigenvalues*.

Definition 2.2.5. Given an undirected or directed graph G, the *characteristic polynomial* of its adjacency matrix A is given by $p_A(\lambda) = \det(\lambda I - A)$. The zeros of $p_A(\lambda)$ are the *eigenvalues* of A. Analogously, the characteristic polynomial of its Laplacian matrix L is given by $p_L(\mu) = \det(\mu I - L)$, and its zeros are the *eigenvalues* of L (or *Laplacian eigenvalues*). The *eigenvectors* associated with the eigenvalues of A (respectively L) are the vectors v such that $Av = \lambda v$ (respectively $Lv = \mu v$).

The *spectrum* (respectively *Laplacian spectrum*) of a graph G is the set of all eigenvalues of A (respectively L), together with their multiplicities. Commonly, the eigenvalues of A are ordered in decreasing order: $\lambda_1 \geq \ldots \geq \lambda_{min}$. On the other hand, the eigenvalues of the Laplacian matrix are ordered in increasing order: $\mu_1 \leq \ldots \leq \mu_{max}$.

A notable property of the Laplacian matrix is that it is positive semi-definite (Proposition 2.2.2), and thus all of its eigenvalues are non-negative, as proved below.

Proposition 2.2.3. *All eigenvalues of the Laplacian matrix of a given undirected or directed graph are non-negative.*

Proof. Let $G = (V, E)$ be an undirected or directed graph with $|V| = n$ and with Laplacian matrix L. Since L is positive semi-definite (Proposition 2.2.2), for any eigenvector $v \in \mathbb{R}^n$ of L, with eigenvalue μ, we have

$$v^T L v = v^T \mu v = \mu v^T v = \mu ||v||^2 \geq 0, \tag{2.13}$$

but $\mu||v||^2 \geq 0$ if and only if $\mu \geq 0$, where $||.||$ is the Euclidean norm. □

From the previous proposition, we conclude that zero is always the smallest eigenvalue of the Laplacian matrix, i.e. $\mu_1 = 0$.

Definition 2.2.6. Let $M \in \mathbb{R}$ be a matrix. The square roots of the eigenvalues of the positive semi-definite matrix $M^T M$ are called the *singular values* of M. Typically, the notation $\{\sigma_i\}_{i=1}^n$ is used to denote the set of singular values (with their multiplicities).

If A is the adjacency matrix, since A is real and symmetric, all singular values of A are equal to the absolute value of its eigenvalues.

An important theorem associated with non-negative matrices is the *Perron-Frobenius theorem* [31].

Theorem 2.2.1. (**Perron-Frobenius**) Let M be a square non-negative matrix. Then M has an eigenvalue $\hat{\lambda} \geq 0$ such that $|\lambda| \leq \hat{\lambda}$, for all eigenvalue λ of M, and the eigenvector associated with $\hat{\lambda}$ is a non-negative vector.

By definition, the adjacency matrix of an undirected or directed graph is always square and non-negative, thus the Perron-Frobenius theorem guarantees that a graph or digraph has a non-negative eigenvector.

2.3 Graph Measures

Graphs in general present a wide variety of structural features, for example, they can contain certain types of subgraphs, their nodes can have specific degree distributions,

they can have specific average path lengths, their nodes can be organized into clusters or communities, etc. Accordingly, it is natural to try to define specific measures to quantify each of these structural characteristics. Once defined, they can be used to study the functionality, topology, and dynamics of real-world networks [57, 195].

Over the years, a large number of graph measures have been proposed, each with the purpose of quantifying some information about a specific characteristic of the network [65, 223]. For instance, the *global efficiency* tries to quantify how efficiently information is propagated through the graph; the *degree centrality* tries to quantify how influential or central a node is in the graph; the *closeness centrality* tries to quantify the ability of a node to transmit information [76, 77, 88, 91]. Recently, all these quantitative approaches to dealing with graphs have been brought together into a new branch within graph theory called *quantitative graph theory* (QGT) [74], which we will discuss in more detail in Chapter 4.

One can classify graph measures into two major categories:

- **Local measures:** refer to the measures that try to extract the properties of the nodes in a graph.
- **Global measures:** refer to the measures that try to extract the global properties of a graph by taking into account the graph as a whole.

In the next subsections, we present several well-known graph measures (most of them can be found in the aforementioned bibliography), and all of them represent *graph invariants*. Also, all graphs (or digraphs), $G = (V, E)$, will be considered with $|V| = n$ and $|E| = m$, and its vertices and edges (or arcs) will be denoted by $i, j \in V$ and $(i, j) \in E$, respectively, and $A = (a_{ij})$ will denote their adjacency matrix. Unless otherwise specified, the variants related to weighted graphs are directly extended to weighted digraphs.

2.3.1 Distance-Related Measures

Perhaps the most common and intuitive measures related to graphs are the node degree and node degree distribution. The next most common approaches to characterize graphs are the *distance-based measures*. These measures allow us to quantify integration at the global level of the network, that is, how each node interacts with all other nodes, and thus we can classify them as measures of *global integration* [240]. In what follows, we present some of the most relevant measures associated with distances.

Shortest Path and Distance (*global*): The *shortest path* (or *geodesic path*) between two given vertices in an undirected graph is the path with the minimum amount of edges between them. The *distance* (or *geodesic distance*) between two vertices i and j,

denoted as $d(i,j) = d_{ij}$, is the length of the shortest path between i and j, and it can be written in terms of the adjacency matrix entries as

$$d_{ij} = \sum_{x,y \in g_{i \leftrightarrow j}} a_{xy}, \tag{2.14}$$

where $g_{i \leftrightarrow j}$ is the geodesic path between i and j. Since the adjacency matrix is symmetric, we have $d_{ij} = d_{ji}$, for all i, j. Also, we define $d_{ij} = +\infty$ for every disconnected pair i, j.

Analogously, for directed graphs, the *directed distance* from a vertex i to a vertex j is the length of the shortest directed path from i to j, i.e.

$$\vec{d}_{ij} = \sum_{x,y \in \vec{g}_{i \to j}} a_{xy}, \tag{2.15}$$

where $\vec{g}_{i \to j}$ is the geodesic directed path from i to j. Notice that $\vec{d}_{ij}$ is an asymmetric function, since we might have $\vec{d}_{ij} \neq \vec{d}_{ji}$ for some i and j, therefore, strictly speaking, $\vec{d}$ is a *quasi-distance* (Definition 2.1.30). If there is no directed path from i to j, then we put $\vec{d}_{ij} = +\infty$. Also, it is worth noting that the length of the shortest path is equal to the length of the shortest walk in both undirected and directed cases.

Lastly, for weighted graphs, let f be a weight-to-distance function. The *weighted distance* between two vertices i and j is the length of the shortest path in relation to the function f, i.e.

$$d^{\omega}_{ij} = \sum_{x,y \in g_{i \leftrightarrow j}(f)} f(\omega_{xy}), \tag{2.16}$$

where $g_{i \leftrightarrow j}(f)$ is the geodesic path between i and j in relation to f (or the *weighted geodesic path*). Here, since we want a higher weight to be associated with a shorter path, if the weights are normalized, then we identify f as the weight-to-distance function defined by Equation (2.7), i.e. $f = D^{\omega}$, otherwise, we can use the modified version of D^{ω} as explained in Observation 2.1.1.

Characteristic Path Length (*global*): The *characteristic path length* (or *average path length* or *average shortest path length*) [277] is a measure that represents the path length that is most likely to occur in the graph, i.e. it is the average distance between all possible pairs of vertices. Essentially, it is a measure of how efficiently information travels through the graph. Mathematically, it is defined by

$$L(G) = \frac{1}{n} \sum_{i \in V} \frac{\sum_{j \in V, j \neq i} d_{ij}}{n-1} = \sum_{\substack{i,j \in V \\ i \neq j}} \frac{d_{ij}}{n(n-1)}. \tag{2.17}$$

For directed graphs, *directed characteristic path length*, $\vec{L}$, is obtained by replacing

the distance d_{ij} with the directed distance $\vec{d}_{ij}$, and for the weighted case, the *weighted characteristic path length*, L^{ω}, is obtained by replacing d_{ij} with d^{ω}_{ij}. Some algorithms consider $d_{ij} = 0$ (respectively $\vec{d}_{ij} = 0$) when $(i, j) \notin E$, in this case n is the order of G, otherwise n must be the order of its giant component.

Eccentricity (*global*): For a connected graph, the *eccentricity* of a vertex i is the maximum distance between i and any other vertex j in the graph, i.e.

$$\mathrm{ecc}(i) = \max_{j \in V} d(i, j). \tag{2.18}$$

Analogously, for strongly connected digraphs, the *directed eccentricity*, denoted as $\vec{\mathrm{ecc}}(i)$, is the maximum directed distance from i to any other vertex j in the digraph.

Diameter and Radius (*global*): The *diameter* of a connected graph G is the maximum eccentricity among all of its vertices, i.e.

$$\mathrm{diam}(G) = \max_{i \in V}(\mathrm{ecc}(i)). \tag{2.19}$$

If G is a strongly connected digraph, the *directed diameter* is the maximum directed eccentricity among all of its vertices. A related measure is the *radius* of a connected graph G, which is the minimum eccentricity among all of its vertices:

$$\mathrm{rad}(G) = \min_{i \in V}(\mathrm{ecc}(i)). \tag{2.20}$$

Again, for strongly connected digraphs, the *directed radius* is obtained by replacing $\mathrm{ecc}(i)$ with the directed eccentricity $\vec{\mathrm{ecc}}(i)$ in the previous formula.

Global Efficiency (*global*): Let's define by $\epsilon_{ij} = d_{ij}^{-1}$ the *efficiency* in the communication between two vertices i and j of a graph G. Latora and Marchiori [165] defined the *global efficiency* (or *average efficiency*) of G as

$$E_{glob}(G) = \frac{1}{n} \sum_{i \in V} \frac{\sum_{j \in V, j \neq i} \epsilon_{ij}}{n - 1}. \tag{2.21}$$

This measure is closely related to the characteristic path length and it is another way of trying to capture how efficiently information is propagated through the graph. The *directed global efficiency*, $\vec{E}_{glob}$, is obtained by replacing d_{ij} with the directed distance $\vec{d}_{ij}$, and the *weighted global efficiency*, E^{ω}_{glob}, is obtained by replacing d_{ij} with d^{ω}_{ij}.

Communicability (*global*): The *communicability* between two nodes [89, 91] is a measure based on the number of walks that exist between them and that assigns different importance to these walks based on their lengths. It can be expressed in terms of the

powers of the adjacency matrix as

$$CM(i,j) = \sum_{k=0}^{\infty} c_k (A^k)_{ij}, \tag{2.22}$$

since $(A^k)_{ij}$ is the number of walks of length k between i and j (Proposition 2.2.1), and the coefficients c_k must be defined in such a way that: guarantees the convergence of the series; the values of c_k increase as we decrease k and vice versa (i.e. assign greater importance to shorter walks); and produce positive values for all pairs i, j, $i \neq j$.

A convenient choice for the coefficients c_k is $c_k = 1/k!$, since in this case all of the above conditions are satisfied and we have:

$$CM(i,j) = \sum_{k=0}^{\infty} \frac{(A^k)_{ij}}{k!} = (\exp(A))_{ij}. \tag{2.23}$$

Note that the communicability between a pair of nodes increases when the number of walks between them increases.

Returnability (*global*): Returnability [90, 91] is a measure that tries to quantify the amount of information that flows through a digraph and returns to its original sources by considering the relative contribution of all closed directed walks presented in the digraph. It may be considered as a measure of "returnability of information." Formally, since $\mathrm{Tr}(A^k)$ is equal to the number of closed directed k-walks in a digraph (Corollary 2.2.1), we define its *returnability* as

$$K_r(G) = \sum_{k=2}^{\infty} c_k \mathrm{Tr}(A^k), \tag{2.24}$$

where the coefficients c_k must satisfy the same three conditions exposed previously for the communicability measure. By choosing $c_k = 1/k!$ as before, we have:

$$K_r(G) = \sum_{k=2}^{\infty} \frac{\mathrm{Tr}(A^k)}{k!} = \mathrm{Tr}(\exp(A)) - n. \tag{2.25}$$

The term n in the right-hand side of Equation (2.25) comes from our assumption that the digraph does not contain any self-loops or directed cycles of length 1, so the two first terms, $\mathrm{Tr}(A^0) = n$ and $\mathrm{Tr}(A^1) = 0$, are removed. Also, we can define the *relative returnability* as

$$K'_r(G) = \frac{\mathrm{Tr}(\exp(A)) - n}{\mathrm{Tr}(\exp(A')) - n}, \tag{2.26}$$

where A' is the adjacency matrix of the underlying undirected graph.

2.3.2 Measures of Centrality

In this part, we present the most relevant centrality measures related to a node found in the literature, such as degree centrality, closeness centrality, and betweenness centrality. Each one of these measures tries to quantify some specific property or role of a node in the graph. In general, node centrality measures try to quantify the "importance," "influence," or "centrality" of a node within a graph, in the sense of capturing the capacity that a node has to "spread" and/or "receive" information for/from other nodes, which may be characterized by the direct contact with other nodes, its closeness to a large number of other nodes, and the number of pairs of nodes that require this node as an intermediary in their interactions.

Let's start with the simplest and most straightforward measure of node centrality: the *degree centrality*.

Degree Centrality (*local*): The basic idea behind the degree centrality of a node [105] is to quantify the importance or centrality of this node based on the number of edges which are incident to it in the graph, i.e. if a node's degree is higher than another's, then that node is more influential or central than the other. As mentioned above, by "influential" or "central" we mean the capacity of spreading and/or receiving information for/from other nodes. Formally, the *degree centrality* of a node i is simply its degree divided by $(n-1)$, which is the maximum number of nodes that i can be adjacent to, i.e. it is the proportion of nodes that are adjacent to i:

$$C_{dg}(i) = \frac{\deg(i)}{n-1}. \tag{2.27}$$

For directed graphs, we have two different centralities, namely: the in-degree centrality and the out-degree centrality. The *in-degree centrality* is the proportion of nodes which arrives at i:

$$C^{-}_{dg}(i) = \frac{\deg^{-}(i)}{n-1}. \tag{2.28}$$

Similarly, the *out-degree centrality* is the proportion of nodes which leaves i:

$$C^{+}_{dg}(i) = \frac{\deg^{+}(i)}{n-1}. \tag{2.29}$$

Moreover, for a digraph without double-edges, it's clear that $C_{dg}(i) = C^{-}_{dg}(i)+C^{+}_{dg}(i)$. On the other hand, if double-edges exist in the digraph, then $C_{dg}(i) = C^{-}_{dg}(i)+C^{+}_{dg}(i)-\deg^{\pm}(i)/(n-1)$, where $\deg^{\pm}(i)$ is the number of double-edges incident to i. The weighted versions of each of the previous degree centralities are obtained by replacing the degrees with their respective weighted formulas (Definition 2.1.32).

Closeness Centrality (*local*): Closeness centrality, originally proposed by Bavelas [30], is a measure that quantifies how close a node is relatively to all other nodes in the

graph, that is, it identifies the nodes that can spread/receive information in an efficient way. Mathematically, the *closeness centrality* of a node i can be defined as the inverse of the sum of the shortest paths between i and every other node in the graph, i.e.

$$Cl(i) = \frac{1}{\sum_{\substack{j \in V \\ j \neq i}} d_{ij}}. \tag{2.30}$$

It's important to note that closeness centrality is defined for connected graphs. Thus, if N is the order of the giant component, we can define the *normalized closeness centrality* by multiplying the formula (2.30) by $(N-1)$:

$$Cl(i) = \frac{N-1}{\sum_{\substack{j \in V \\ j \neq i}} d_{ij}}. \tag{2.31}$$

For the directed case, we simply replace the distance d_{ij} with the directed distance $\vec{d}_{ij}$, and for the weighted case, we replace d_{ij} with d^{ω}_{ij}.

Harmonic Centrality (*local*): As mentioned previously, the closeness centrality is not defined for disconnected graphs. To overcome this problem, the *harmonic centrality* [35] of a node i was introduced as the sum of the inverse of all distances between i and all other nodes, i.e.

$$HC(i) = \sum_{\substack{j \in V \\ j \neq i}} \frac{1}{d(i,j)}, \tag{2.32}$$

where the convention $1/\infty = 0$ is adopted. For the directed case, we simply replace d_{ij} with $\vec{d}_{ij}$, and for the weighted case, we replace d_{ij} with d^{ω}_{ij}.

Betweenness Centrality (*local*): Betweenness centrality [105] is a measure of how much control or influence a node has over the information flow in the graph. It takes into account the proportion between all geodesic paths that pass through a specific node and all other geodesic paths between all other nodes in the graph. Formally, the *betweenness centrality* of a node i is defined as

$$B(i) = \sum_{\substack{h,j \in V \\ h \neq j, h \neq i, j \neq i}} \frac{\rho_{hj}(i)}{\rho_{hj}}, \tag{2.33}$$

where $\rho_{hj}(i)$ is the number of geodesic paths from h to j that pass through i, and ρ_{hj} is the total number of geodesic paths from h to j.

Also, note that betweenness centrality is defined for connected graphs, and we can define the *normalized betweenness centrality* by multiplying Equation (2.33) by the normalization term $2/(N-1)(N-2)$, where N is the number of nodes in the giant

component:

$$B(i) = \frac{2}{(N-1)(N-2)} \sum_{\substack{h,j \in V \\ h \neq j, h \neq i, j \neq i}} \frac{\rho_{hj}(i)}{\rho_{hj}}. \tag{2.34}$$

For directed graphs, we consider $\vec{\rho}_{hj}(i)$ as the number of directed geodesic paths from h to j that pass through i and $\vec{\rho}_{hj}$ the total number of directed geodesic paths from h to j. Moreover, the normalization term for the directed case is $1/(N-1)(N-2)$. Similarly, for weighted graphs, we define $\rho^{\omega}_{hj}(i)$ and ρ^{ω}_{hj} in an analogous way as before, but considering the weighted geodesic paths.

Over the years, several variants of betweenness centrality have been proposed, such as the flow betweenness centrality, the random walk betweenness centrality, and the communicability betweenness centrality (see [88], chap. 7).

Reaching Centrality (*local/global*): Before introducing the measure known as "reaching centrality," let's introduce the concept of hierarchy in networks [188]. In the literature, there are three main types of hierarchies, namely: *order hierarchy*, *nested hierarchy*, and *flow hierarchy*. Order hierarchy is described as the presence of an order in the elements of a set, that is, it is equivalent to an ordered set (e.g., an order in the vertex set of a network); nested hierarchy is described as the presence of higher and lower level components in the network, such that higher level components comprise of and include lower level components (e.g., cliques formed by smaller cliques); finally, flow hierarchy comprises the influence that a node has on other nodes, thus these other nodes are considered to be at a lower level than the node that influences them.

The concept of reaching centrality of a node, as proposed in [188], quantifies the concept of flow hierarchy in a digraph, i.e. how a node influences the flow of information through the digraph. Let G be a digraph and let $r_G(i)$ be the number of nodes in G that are reachable from the node i. The *local reaching centrality* of i is defined as the proportion of the nodes which are reachable from i, i.e.

$$C_R(i) = \frac{r_G(i)}{n-1}. \tag{2.35}$$

Let $C_R^{max} = \max_{i \in V}\{C_R(i)\}$ be the maximum local reaching centrality obtained in G. We define the *global reaching centrality* as the average of the difference between C_R^{max} and $C_R(i)$, over all nodes in the digraph, i.e.

$$GRC = \frac{\sum_{i \in V}[C_R^{max} - C_R(i)]}{n-1}. \tag{2.36}$$

For a weighted digraph, there are variants of the local and global reaching centralities introduced in [188]. However, here we will remain with the same formulas as we are solely considering the count of the reachable nodes.

2.3.3 Measures of Segregation

By "measures of segregation" we mean to specify measures that attempt to quantify the tendency of nodes to segregate into clusters, communities, or modules, which are different ways to refer to densely connected neighborhoods. In the literature, there are a variety of such measures, but here we will focus on three of them, namely: *clustering coefficient*, *rich-club coefficient*, and *local efficiency*.

Clustering Coefficient (*local/global*): The clustering coefficient, as introduced in [277], is a measure that tries to quantify the tendency of nodes to form clusters in the network. Formally, the *clustering coefficient* (or *agglomeration coefficient*) (local) of a node i is defined as the ratio between the number of triangles containing i, denoted by $t(i)$, and the maximum number of edges between its neighbors (equal to $\deg(i)(\deg(i)-1)/2$), i.e.

$$C(i) = \frac{2t(i)}{\deg(i)(\deg(i)-1)}. \tag{2.37}$$

The quantity $t(i)$ can be expressed in terms of the adjacency matrix entries as

$$t(i) = \frac{1}{2}\sum_{j,h\in V} a_{ij}a_{ih}a_{jh}. \tag{2.38}$$

The formula (2.37) is also known as *local clustering coefficient*. Furthermore, we can define the *average clustering coefficient* as the sum of $C(i)$ over all nodes $i \in V$ normalized by the order of the graph, i.e.

$$\bar{C} = \frac{1}{n}\sum_{i\in V} C(i) = \frac{1}{n}\sum_{i\in V} \frac{2t(i)}{\deg(i)(\deg(i)-1)}. \tag{2.39}$$

For directed graphs, Fagiolo [93] proposed a variant of the formula (2.37) by considering directed triangles and the total number of arcs between the neighbors of i (excluding the doubles-edges), i.e.

$$\vec{C}(i) = \frac{\vec{t}(i)}{\deg^{tot}(i)(\deg^{tot}(i)-1) - 2\deg^{\pm}(i)}, \tag{2.40}$$

where $\deg^{tot}(i) = \deg^{-}(i) + \deg^{+}(i)$, $\deg^{\pm}(i)$ is the number of double-edges incident to i, and $\vec{t}(i)$ is the number of directed triangles containing i as one of their nodes, that is,

$$t(\vec{i}) = \frac{1}{2}\sum_{j,h\in V} (a_{ij}+a_{ji})(a_{ih}+a_{hi})(a_{jh}+a_{hj}). \tag{2.41}$$

Analogously to the formula (2.39), we define the *average directed clustering coefficient* as the sum of $\vec{C}(i)$ over all nodes $i \in V$ normalized by the order of the graph.

A weighted version of the clustering coefficient was proposed by Onnela et al. [198]

by defining the quantity $t^{\omega}(i)$ as the sum of the geometric mean of the scaled weights of the triangles (triangle intensities):

$$t^{\omega}(i) = \frac{1}{2} \sum_{j,h \in V} (\hat{\omega}_{ij} \hat{\omega}_{ih} \hat{\omega}_{jh})^{1/3}. \tag{2.42}$$

where $\hat{\omega}_{ij} = \omega_{ij} / \max(\omega_{ij})$, and then replacing $t(i)$ with $t^{\omega}(i)$ in the formula (2.37).

Rich-Club Coefficient (*local*): The rich-club coefficient, first proposed by Zhou and Mondragon [292], is a measure that attempts to quantify the tendency of nodes with high degrees (called the "rich nodes") to be densely connected to each other. Formally, let $N_{>k}$ be the number of nodes with degree $> k$, and let $E_{>k}$ be the number of edges among the nodes with degree $> k$, the *rich-club coefficient* is defined as the ratio:

$$\phi(k) = \frac{2E_{>k}}{N_{>k}(N_{>k} - 1)}. \tag{2.43}$$

For directed graphs, Smilkov and Kocarev [238] defined the *in-degree rich-club coefficient*, which considers the in-degree instead of the undirected degree, i.e.

$$\phi^{in}(k) = \frac{E^{in}_{>k}}{N^{in}_{>k}(N^{in}_{>k} - 1)}, \tag{2.44}$$

where $N^{in}_{>k}$ is the number of nodes i having $\deg^{-}(i) > k$, and $E^{in}_{>k}$ is the number of directed edges connecting those $N^{in}_{>k}$ nodes.

The formula (2.43) can also be defined using the out-degree instead of the in-degree, and in this case, we have the *out-degree rich-club coefficient*:

$$\phi^{out}(k) = \frac{E^{out}_{>k}}{N^{out}_{>k}(N^{out}_{>k} - 1)}, \tag{2.45}$$

where $N^{out}_{>k}$ is the number of nodes i having $\deg^{+}(i) > k$, and $E^{out}_{>k}$ is the number of directed edges connecting those $N^{out}_{>k}$ nodes.

Local Efficiency (*global*): Let $G(i)$ denote the subgraph of G formed by the open neighborhood of a vertex i. Latora and Marchiori [165] defined the *local efficiency* of G as the average efficiency of the local subgraphs $G(i)$, i.e.

$$E_{loc}(G) = \frac{1}{n} \sum_{i \in V} E_{glob}(G(i)). \tag{2.46}$$

The *directed* and *weighted local efficiency* are is obtained by replacing $E_{glob}(G(i))$ with $\vec{E}_{glob}(G(i))$ and with $E^{\omega}_{glob}(G(i))$, respectively.

2.3.4 Entropy Measures

The concept of *entropy* first appeared in the study of thermodynamic systems when, in an 1865 study, Rudolf Clausius coined the term to mean "transformation content" in the sense of availability/unavailability of energy in a system, and it was further studied in other contexts [121]. In his 1948 paper [234], Claude Shannon, studying transmission of signals in communication systems, proposed an idea of "information entropy" (which came to be known as *Shannon entropy*) as a measure of *uncertainty* in the sense of how much "randomness" a signal carries or "the 'amount of surprise' a message source has for a receiver" [187].

Here we are interested in the concept of entropy associated with networks. The *graph entropy* or the *topological information content* of a graph can be interpreted as a type of *structural entropy*, and it was first proposed by Rashevsky [217] as the Shannon entropy of some probability distributions obtained from the symmetric structure of the vertices of a graph, which can also be computed in terms of the orbits of its automorphism group [190]. Nonetheless, over the years, several new approaches to defining graph entropy have been proposed [75], for example, entropies based on the degree of the nodes [48, 271], entropies based on the eigenvalues of the adjacency matrix [236, 257], and entropies based on the Laplacian eigenvalues [203, 285].

In what follows, we discuss the *entropy of the degree distribution* for undirected and directed graphs.

Entropy of the Degree Distribution (*global*): The *entropy of the degree distribution* (or *degree distribution entropy*) [271] is a measure that tries to capture the heterogeneity of the edge distribution in a given graph G. It is defined as the Shannon entropy of the node degree distributions (2.2), i.e.

$$H(G) = -\sum_{k=1}^{n-1} p(k) \log_2 p(k). \tag{2.47}$$

In the case where G is a directed graph, we have two possible definitions: the *in-degree distribution entropy* and the *out-degree distribution entropy*. These entropies are respectively defined by

$$H^{in}(G) = -\sum_{k=1}^{n-1} p^{in}(k) \log_2 p^{in}(k), \tag{2.48}$$

$$H^{out}(G) = -\sum_{k=1}^{n-1} p^{out}(k) \log_2 p^{out}(k), \tag{2.49}$$

where $p^{in}(k)$ are the node in-degree distributions (2.5) and $p^{out}(k)$ are the node out-degree distributions (2.6).

Notice that if all nodes of a graph or digraph have the same degree, or same in-degree, or same out-degree, the respective entropies reach their minimum, i.e. $H(G) = H^{in}(G) = H^{out}(G) = 0$ (with the convention $0 \log_2 0 = 0$), and reach their maximum if $p(k) = p^{in}(k) = p^{out}(k) = 1/(n-1)$, for all $k = 1, 2, ..., n-1$. Thus, we can roughly say that these entropy measures quantify the "degree of disorder" or "degree of randomness" (*in relation to* the inner or outer flux in the case of H^{in} or H^{out}, respectively) of a network since they produce higher values for networks that are closer to a random model and lower values for those that are closer to a regular model (see Section 2.5 for an overview of random graph models).

2.3.5 Spectrum-Related Measures

In this last part, we present some measures that are related to the spectra of the adjacency and Laplacian matrices of an undirected or directed graph, such as *graph energy*, *Katz centrality*, *eigenvector centrality*, and *spectral entropy*.

Graph Energy (*global*): Gutman [135] originally defined the graph energy of an undirected graph as the sum of the absolute values of the eigenvalues (with multiplicities) of its adjacency matrix. As observed by Nikiforov [196], the *trace norm* of a matrix, denoted by $||\cdot||_*$, is the sum of its singular values, and as we observed earlier, for a real symmetric matrix, its singular values are equal to the absolute value of its eigenvalues, thus the energy of a graph can be defined as the trace norm of its adjacency matrix.

In effect, the trace norm can be used to define energy for digraphs as well. Arizmendi and Arizmendi [9] defined the energy of a directed graph as the trace of the matrix $|A|^+ = (AA^T)^{1/2}$ (equivalently for $|A|^- = (A^T A)^{1/2}$), which is equal to the sum of the singular values of A (with their respective multiplicities), which in turn is equal to the trace norm as discussed previously. Therefore, for both types of graphs, undirected and directed, we define the *graph energy* as

$$\varepsilon(G) = ||A||_* = \mathrm{Tr}(|A|^+) = \mathrm{Tr}(|A|^-) = \sum_{i=1}^{n} \sigma_i. \tag{2.50}$$

In particular, if G is an undirected graph, we have $\varepsilon(G) = \sum_{i=1}^{n} |\lambda_i|$. Graph energy may be seen as a measure of graph connectivity [235].

Katz Centrality (*local*): Katz centrality, first proposed by L. Katz [156], is based on the idea that the importance of a node is not only influenced by its immediate neighbors but also by nodes that are farther away in the graph. Unlike the degree centrality of a node i, for example, which takes into account solely the influence of its nearest-neighbors (walk of length 1), the Katz centrality of i also takes into account all the other nodes that are connected to i by a walk of length > 1. Formally, we can

take these nodes into account by considering the series $A^0 + A^1 + ... + A^k + ...$, since the (i,j)-entries of A^k are the number of (i,j)-walks of length k (Proposition 2.2.1). However, this series might diverge, then we need to introduce an *attenuation factor* $\alpha \in \mathbb{R}_+$ in such a way that the series converges. This leads us to the definition of the *Katz centrality* of a node i:

$$K(i) = \left[\left(\sum_{k=0}^{\infty} \alpha^k A^k \right) |1\rangle \right]_i , \tag{2.51}$$

where $|1\rangle = (1, ..., 1)^T$ and the subscript i in the brackets represents the i-th position of the vector inside the brackets. In order to guarantee the convergence of (2.51), the attenuation factor must be $\alpha \neq 1/\lambda_1$, where λ_1 is the largest eigenvalue of A. In this case, Equation (2.51) can be written as

$$K(i) = \left[(I_n - \alpha A)^{-1} |1\rangle \right]_i , \tag{2.52}$$

where I_n is the $n \times n$ identity matrix. Typically, the attenuation factor is chosen to be $\alpha < 1/\lambda_1$. For directed graphs, we consider either the arcs arriving at i ($K^{in}(i)$) or the arcs leaving i ($K^{out}(i)$), i.e.

$$K^{in}(i) = \left[\langle 1| (I_n - \alpha A)^{-1} \right]_i , \tag{2.53}$$

$$K^{out}(i) = \left[(I_n - \alpha A)^{-1} |1\rangle \right]_i . \tag{2.54}$$

Eigenvector Centrality (*local*): The *eigenvalue centrality* [39] of a node $i \in V$ is defined as the i-th entry of the eigenvector associated with the largest eigenvalue (λ_1) of the adjacency matrix A of G, i.e.

$$C_e(i) = \left(v_1\right)_i = \left(\frac{1}{\lambda_1} A v_1 \right)_i . \tag{2.55}$$

If G is a digraph, the adjacency matrix A might be non-symmetric ($A^T \neq A$), then we consider the right eigenvector ($Av = \lambda_1 v$) or the left eigenvector ($A^T v = \lambda_1 v$) associated with λ_1 in formula (2.55), and then we can have right and left eigenvector centralities associated with the node i.

Moreover, the Perron-Frobenius theorem (Theorem 2.2.1) guarantees that the eigenvector associated with λ_1 is non-negative, thus the eigenvalue centrality of every node is non-negative.

The eigenvalue centrality can be seen as a modification of the Katz centrality, as demonstrated in [91], and it can be interpreted as a measure that tries to quantify the importance of a node according to the importance of its neighbors.

Spectral Entropy (*global*): In Subsection 2.3.4, we have already discussed the concept of entropy and presented graph entropy based on the degree distribution. Now we define the *spectral entropy* of a graph or digraph G based on its Laplacian eigenvalues as follows. Let $L(G)$ be the Laplacian matrix of G with eigenvalues $\{\mu_i\}_i$ (with multiplicities), and let

$$p(\mu_i) = \frac{\mu_i}{\sum_i \mu_i} \tag{2.56}$$

be the "eigenvalue probabilities," i.e. the contribution of μ_i in the Laplacian spectrum (Proposition 2.2.3 guarantees that $\mu_i \geq 0$, $\forall i$). Assuming the conventions $0/0 = 0$ and $0 \log_2 0 = 0$, we define the *spectral entropy* of G as the Shannon entropy of the eigenvalue probabilities, i.e.

$$S(G) = -\sum_i p(\mu_i) \log_2 p(\mu_i). \tag{2.57}$$

Different definitions of spectral entropy associated with digraphs were presented by Sun et al. [257] and Ye et al. [285].

2.4 Graph Similarity

In this section, we discuss methods and algorithms that are used to quantify how similar (or dissimilar) two graphs are, i.e. quantitative approaches to the *graph similarity comparison problem*. In the previous section, we presented several measures capable of characterizing topological aspects of graphs; now, we discuss how to use them to compare graphs and, in addition, we present distance-based algorithms that try to quantify the differences between two graphs through a "similarity" score. In what follows, all discussion applies to both directed and undirected cases.

The graph similarity comparison problem can be summarized in the following question: given two graphs, how similar are they? Despite the efforts employed in the development of methods to evaluate the similarity of graphs, there is still no one capable of satisfactorily answering this question [222]. Also, the meaning of "similarity" may vary depending on the application. For example, in different contexts, we may be interested in answering one of the following questions: are the graphs copies of each other? What changes should we apply in the graphs to transform one into the other? If we allow the graph to change over time, can we assess whether two graphs were generated from a common ancestor graph (by the successive application of evolutionary rules)? To tackle these and other questions, several methods were proposed [41, 280, 288].

Mheich et al. [183] described two major classes of methods of graph comparison, namely:

- **Statistical comparison methods:** These methods consist of applying (local

and global) measures of topological characterization (e.g., measures of centrality, segregation, integration, spectral measures, etc.) in different groups of graphs in order to compare them by using statistical tests. Furthermore, for real-world networks, this comparison can be done in two ways: either comparing them with equivalent random networks (null models), or comparing them with other correlated groups of real-world networks. Other methods include graph correlation [113] and analysis of graph variability [114].

- **Distance-based comparison algorithms:** These algorithms produce "similarity" scores as a result of comparing two graphs. Typically, these scores are normalized (i.e., they output 1 if the two graphs are totally different from each other and output 0 if they are totally similar). Among these algorithms are graph/subgraph isomorphisms, edit distances, graph kernels, and structure distances (distances based on the presence/absence of edges, cliques, quasi-cliques, or other subgraphs).

Moreover, we note that when a statistical comparison analysis is performed using local/global measures, we may use the nomenclatures *node-wise analysis*/*global-level analysis*, respectively.

In what follows, we present some distance-based comparison algorithms that are of interest in this text.

2.4.1 Distance-Based Comparison Algorithms

Graph Edit Distance: The graph edit distance (GED) is one of the most common approaches in determining the similarity between two graphs. As observed in [115], a graph can be transformed into another by a finite sequence of graph editing operations (node/edge insertion, node/edge deletion), and the GED is defined as the minimum cost of these editing operations, for some suitable cost function. Formally, we can express the GED between two graphs G_1 and G_2 as

$$d_{GED}(G_1, G_2) = \min \sum_{k=1}^{N_{op}} c(e_k), \tag{2.58}$$

where $c(e_k)$ is the cost of the k-th editing operation e_k, and N_{op} is the total number of editing operations.

Graph Kernels: The fundamental idea behind graph kernels [41] is to build feature vectors from graphs by mapping these features into a Hilbert space[3] $\mathcal{H}$ (feature space), and then define the kernel function as the inner product of this space.

[3]Hilbert space is a vector space provided with inner product such that it is complete with respect to the induced norm.

Let G_1 and G_2 be two graphs and let $\phi(G_1), \phi(G_2) \in \mathcal{H}$ be their feature vectors. Let $\langle \cdot, \cdot \rangle$ denote the inner product of $\mathcal{H}$. The graph kernel is

$$k(G_1, G_2) = \langle \phi(G_1), \phi(G_2) \rangle. \tag{2.59}$$

The value $k(G_1, G_2)$ is the similarity score used to compare the graphs. There are several ways to extract feature vectors from graphs and then define a kernel, for example, by counting the matching random walks between two graphs (random walk graph kernels) or by comparing all shortest path lengths in two graphs (shortest path kernel) [41, 267].

Structure Distance: Similarly to the case of graph kernels, the basic idea behind defining a structure distance is to extract structure vectors (feature vectors) from graphs, based, for example, on the counting of meaningful substructures (subgraphs/ motifs), and embed them into a metric space. Since we are considering all graphs finite, here we consider the n-dimensional real metric space $(\mathbb{R}^n, d_p)$ whose metric d_p is induced by the p-norm $||\cdot||_p$, $p \in [1, \infty)$.

Let $v^1 = (v_1^1, ..., v_n^1), v^2 = (v_1^2, ..., v_n^2) \in \mathbb{R}^n$ be structure vectors associated with G_1 and G_2, respectively. We can define structure distances based on the p-norm as

$$d_{str}^p(G_1, G_2) = ||v^1 - v^2||_p = \left(\sum_{i=1}^{n} |v_i^1 - v_i^2|^p \right)^{1/p}. \tag{2.60}$$

When $p = 2$, d_{str}^p represents the Euclidean distance. Also, if all entries of the structure vectors are considered non-negative, and assuming that at least one of the vectors is different from the null vector, we can normalize Equation (2.60) by dividing it by $||v^1||_p + ||v^2||_p$.

2.5 Random Graphs

This section is based on the books [195, 276], however, the fundamental concepts of random graphs presented here can be found in the classic book by Bollobás [37]. Moreover, an algorithmic approach to random graphs can be found in [152].

2.5.1 Erdős-Rényi Model

Roughly speaking, a random graph of order n is nothing more than a set of n vertices and a set of edges connecting pairs of these vertices in some random way. Almost all of random graph theory is concerned either with the analysis of the *Erdős-Rényi* $G(n, M)$ *random models* (in homage to the contributions of these authors in the study of this model [87]), or with the analysis of the *Erdős-Rényi* $G(n, p)$ *random models* (also called

binomial random models or *Gilbert models*, since they were first proposed by Edgar Gilbert [118]), and the relationship between these two models. Below, their formal definitions are presented.

The G(n, M) model: The graph $G(n, M)$ is generated as follows: given a set of n (fixed) vertices, we choose a (fixed) number M of distinct pairs of vertices uniformly at random, among the $\binom{n}{2}$ possible pairs, and we connect each pair with an edge. That is, there are $\binom{\binom{n}{2}}{M}$ ways to place the M edges, and we simply choose any of them with equal probability.

Strictly speaking, a random graph model is defined not just as a single randomly generated graph but as a set of graphs, that is, a probability distribution over the possible graphs. Therefore, the model $G(n, M)$ is correctly defined as a probability distribution $P(G)$ over all simple graphs G, with n vertices and M edges, such that

$$P(G) = \frac{1}{\binom{\binom{n}{2}}{M}}. \tag{2.61}$$

Observation 2.5.1. A directed version of the $G(n, M)$ model can be generated through a straightforward adaptation of the previously described method, i.e. given n fixed vertices and a positive integer N, we add an arc from a vertex to another vertex randomly until we get M arcs.

The G(n, p) model: $G(n, p)$ is the graph with a set of n (fixed) vertices, in which all possible $\binom{n}{2}$ edges exist with probability $0 \leq p \leq 1$. That is, the (fixed) probability of placing an edge between each distinct pair of vertices is p. In this graph, the number M of edges is not fixed. Formally, the model $G(n, p)$ is the set of simple graphs with n vertices in which each graph G appears with probability

$$P(G) = p^M (1-p)^{\binom{n}{2}-M}, \tag{2.62}$$

where M is the number of edges in the graph.

Observation 2.5.2. Similarly to the $G(n, M)$ model, a directed version of the $G(n, p)$ model can be obtained through a straightforward adaptation of the previous method: given n fixed vertices, we add an arc from a vertex to another randomly with a given probability p.

2.5.2 k-Regular Model

There are several models for generating k-regular random graphs uniformly at random, that is, random graphs whose nodes have a given fixed degree k. Perhaps the most

common approach is the *pairing model*. In what follows, we present the algorithm behind this method [255].

Given a set of n nodes, create n sets with kn elements. Choose a pair of elements at random from these sets. Then create an edge (i, j) in the graph if there is a pair composed of elements of the i'th and j'th sets. We disregard equal pairs and pairs of type the (i, i) (self-loops). The resulting random graph is k-regular.

Moreover, we can generate a k-regular random digraph by considering $k = \deg^{in}(v) + \deg^{out}(v)$, for all nodes v in the digraph, and considering (i, j) as arcs in the paring model. Also, we can use the Havel-Hakimi algorithm [159] to generate a digraph from a given sequence of in-degree and a given sequence of out-degree, and by choosing a fixed in- and out-degree k, the resulting digraph is k-regular (in this case, we have $k = \deg^{in}(v) = \deg^{out}(v)$, for all nodes v).

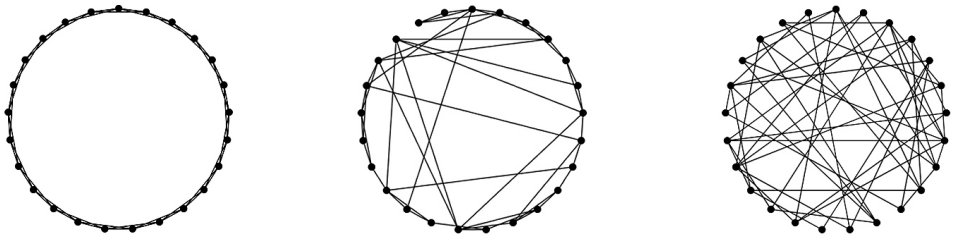
2.5.3 Watts-Strogatz Model

In this part, we present the *Watts-Strogatz model* (WS model), sometimes called *small-world network model*, proposed by Watts and Strogatz [277]. Although it is not the only small-world network model, it is the most frequently used one. The reference for this part is [195].

The WS model is a model originally created to illustrate how two features of social networks (*high clustering coefficient* (high C) and *low path lengths* (low L)) can coexist in the same network. In retrospect, the main contribution of this model is being able to show why the *small-world effect* (the existence of short paths between most vertices) is prevalent in networks of all types, especially in real-world networks.

We can define this model as follows. Let's start with a regular network (or regular grid) of some type. For example, let us arrange n vertices in a circle and connect each of them to the nearest k vertices (k an even number). Then, we randomize this network, traversing each of the edges around the circle in succession, and with probability p, we remove an edge and replace it with another that connects two randomly chosen vertices in a uniform way (i.e., we reconnect some vertices). The probability p is therefore said to be the *reconnection probability* or *rewiring probability*. For directed graphs, the process of rewiring is analogous, even though there are other processes to produce weighted and directed small-world networks [215, 283].

In a small-world network model, the rewiring probability p takes intermediate values between the value p of the initial regular circular network and the p of a random network. That is, when $p = 0$, no edges are reconnected, keeping the configuration of the initial regular network (high C and high L), when $p = 1$, all edges reconnect, producing a random network (low C and low L), and when p takes intermediate values, e.g. $p = 0.3$, only some edges are reconnected, and the generated network (small-world network) takes on properties of both networks (high C and low L), as depicted in Figure 2.3.



(a) Regular network. (b) Small-world network. (c) Random network.

Figure 2.3: Networks obtained through the NetLogo software [279], using the Watts and Strogatz model. All networks have 25 nodes. (a) Regular network: high C and high L ($p = 0, C = 0.5, L = 3.5$). (b) Small-world network: high C and low L ($p = 0.3, C = 0.357, L = 2.51$). (c) Random network: low C and low L ($p = 1, C = 0.086, L = 2.317$).

A quantitative way of defining whether a given network G has a "small-world" structure is by computing its *small-worldness* [146], which is defined as the ratio

$$S_{SW}(G) = \frac{C/C_{rand}}{L/L_{rand}}, \tag{2.63}$$

where C and L are defined as before, and C_{rand} and L_{rand} are these same quantities but computed for an equivalent Erdős-Rényi random network. We say that G is a small-world network if $S_{SW}(G) >> 1$. For directed graphs, we use the directed versions $\vec{C}$ and $\vec{L}$.

2.5.4 Barabási-Albert Model

The last random graph model that we are going to discuss is the so-called *Barabási-Albert model* (BA model) [26]. This model is based on two central rules: *growth* and *preferential attachment*[4].

Briefly, a network is generated through growth and preferential attachment as follows: given a set of n_0 initial nodes, a new node is added, at each time step, to the network and linked to r other existing nodes (where r is a parameter of the model) such that the connection probability is proportional to the degree of each node. Formally, the probability of a new node connecting to a node i is $p(i) = \deg(i)/\sum_j \deg(j)$, where the sum is performed over all other nodes j existing in the network [152].

The previous two rules produce networks whose degree distributions follow a power-law, i.e. let $p(k)$ be the probability of a vertex at chosen at random having degree k, then this probability is proportional to the power $k^{-\eta}$, $\eta \in \mathbb{R}_+$, i.e.

$$p(k) \sim k^{-\eta}. \tag{2.64}$$

[4]Preferential attachment is informally known as the "rich-get-richer" effect.

Probability distributions of the form (2.64) are called *power-laws* [194]. These networks are called "scale-free networks," since their degree distributions have the same shape at different scales.

Furthermore, Bollobás et al. [38] proposed a model which generates scale-free digraphs through the preferential attachment algorithm depending on both, in-degrees and out-degrees of the nodes. The model is based on the parameters δ_{in}, $\delta_{out}, \alpha, \beta, \gamma \in \mathbb{R}_{\geq 0}$, where

- δ_{in} and δ_{out} are bias for choosing nodes based on the in-degree/out-degree distribution, respectively;
- α is the probability of add a new node v and an arc (v, u) to an existing node u, where u is chosen with a probability proportional to $\deg^{in}(u) + \delta_{in}$;
- β is the probability of adding a new arc from an existing node v to another existing node u, where v is chosen with a probability proportional to $\deg^{out}(v) + \delta_{out}$, and u is chosen with a probability proportional to $\deg^{in}(u) + \delta_{in}$;
- γ is the probability of add a new node v and an arc (u, v) to an existing node u, where u is chosen with a probability proportional to $\deg^{out}(u) + \delta_{out}$.

These parameters must be chosen so that the equality $\alpha+\beta+\gamma = 1$ is satisfied. The scale-free digraph grows as we perform, at each time step, one of the actions associated with one of the previous probabilities.

Chapter 3

Digraph-Based Complexes and Directed Higher-Order Connectivity

> *(...) the heart of the scientific method, and of all rational study of any human activity, lies in the process of identifying sets and of understanding the structural properties of relations between sets.*
>
> — R. H. Atkin [11]

In Chapter 2 we discussed the foundations of graph theory and the elements of quantitative graph theory. In this chapter, we will extend our discussion to the field of *simplicial complexes*, which can be seen as a generalization of graphs.

Graphs can be considered hierarchically structured: "higher-level" (larger) subgraphs contain other "lower-level" (smaller) subgraphs. This is the idea behind the *clique organization* or *clique topology* of a graph, since larger cliques contain smaller cliques. This hierarchical organization allows us to build simplicial complexes from the cliques of the graphs, the so-called *clique complexes*. When we are dealing with digraphs, this same type of organization is present but, in this scenario, we may take the directionality of the edges into account and construct special types of complexes out of the digraphs, the so-called *directed clique complexes* (or *directed flag complexes*).

To be more specific, in this chapter we will deal mainly with directed flag complexes constructed from digraphs without double-edges, which constitute special types of simplicial complexes. Furthermore, we will briefly discuss more general types of complexes obtained from digraphs, the so-called *path complexes*. We use the umbrella term *digraph-based complexes* to represent these types of complexes, that is, complexes constructed from digraphs. Here we will focus not only on the development of the theory of directed flag complexes, but also on the development of the theory of homology and persistent homology for these complexes, and mainly, on the development of a *directed Q-Analysis* as a directed analogue of classical Q-Analysis that takes the directionality

of the higher-order connectivity between directed simplices into account, introducing, therefore, the concept of *directed higher-order adjacencies* (lower and upper directed adjacencies). In addition, we will discuss all these constructions for the weighted case, i.e. when these complexes are obtained from weighted digraphs.

3.1 Directed Flag Complexes of Digraphs

In this section, we present the classical theory of simplicial complexes, with considerations about directed simplicial complexes and semi-simplicial sets; subsequently we introduce the directed flag complexes and all the mathematical formalism behind these structures, including the weighted case, i.e. the case when they are obtained from weighted digraphs. Furthermore, we present several concepts from algebraic topology and computational algebraic topology for simplicial complexes constructed out of digraphs, such as simplicial homology, persistent homology, and the combinatorial Hodge Laplacians.

The basic bibliography for this section is [83, 138, 162, 192]. Also, all graphs/digraphs are considered to be non-empty, non-null, finite, simple, and labeled, and all sets are considered to be finite unless said otherwise. The notation $I_n^* = \{0, 1, ..., n\}$ is adopted for the index set.

3.1.1 Simplicial Complexes and Semi-Simplicial Sets

In this subsection, we present the formal definitions of abstract simplicial complexes and abstract directed simplicial complexes, the latter being the fundamental abstract structures behind directed flag complexes. Furthermore, we present the concept of semi-simplicial set, which can be seen as a generalization of the concept of abstract directed simplicial complex.

Abstract Simplicial Complexes

Definition 3.1.1. An *abstract simplicial complex* (ASC) is a finite collection $\mathcal{X}$ of finite sets, such that if $\sigma \in \mathcal{X}$, then for all $\tau \subseteq \sigma$ we have $\tau \in \mathcal{X}$ (closed under subset inclusion).

Let $\mathcal{X}$ be an ASC. Each set $\sigma \in \mathcal{X}$ is called a *simplex* (or *abstract simplex*), or an *n-simplex*, if $|\sigma| = n + 1$ is it cardinality; in this case we define the *dimension* of σ as $\dim \sigma = n$. The dimension of $\mathcal{X}$ is the maximum dimension of σ, $\forall \sigma \in \mathcal{X}$, i.e.

$$\dim \mathcal{X} = \max_{\sigma \in \mathcal{X}} (\dim \sigma).$$

Any element v_i of an n-simplex $\sigma = \{v_0, ..., v_n\}$ is called a *vertex of* σ. A simplex is

uniquely determined by its vertices. Sometimes we may denote an n-simplex by $\sigma^{(n)}$, where the superscript n denotes its dimension. The empty set is considered to be a subset of every simplex, therefore $\emptyset \in \mathcal{X}$. Sometimes the empty set is represented as a simplex with no vertices, i.e. a (-1)-dimensional simplex $\sigma^{(-1)} = \emptyset$ (also called *null simplex*). A *k-face* of an n-simplex σ, $0 \leq k \leq n$, is a k-simplex τ such that $\tau \subseteq \sigma$ (we'll use the notation $\tau \subseteq \sigma$ to denote that τ is a face of σ); in contrast, σ is said to be a *coface* of τ. If $\tau \subseteq \sigma$ is a face of σ such that $\tau \neq \sigma$, then τ is said to be a *proper face*, and in this case we denote $\tau \subset \sigma$. The $(n-1)$-faces of an n-simplex are said to be its *boundary*. A simplex is said to be *maximal* in $\mathcal{X}$ if it is not a face of any other simplex in $\mathcal{X}$. The set of all k-simplices in $\mathcal{X}$ is denoted by X_k; in particular, X_0 is the set of all unit sets of $\mathcal{X}$. The *vertex set* of $\mathcal{X}$ is the set of all vertices of its simplices, i.e.

$$V_{\mathcal{X}} = \bigcup_{\sigma \in \mathcal{X}} \sigma.$$

It's common to say that $\mathcal{X}$ is an ASC on the vertex set $V_{\mathcal{X}}$. A *subcomplex* of $\mathcal{X}$ is a subcollection $\mathcal{S}_{\mathcal{X}} \subseteq \mathcal{X}$ such that $\mathcal{S}_{\mathcal{X}}$ is closed under subset inclusion. If the simplices of a subcomplex $\mathcal{S}_{\mathcal{X}} \subseteq \mathcal{X}$ have dimensions at most k, $0 \leq k \leq \dim \mathcal{X}$, then $\mathcal{S}_{\mathcal{X}}$ is called the *k-skeleton* of $\mathcal{X}$, and denoted by $\mathcal{S}_{\mathcal{X}}^{(k)}$, i.e. the k-skeleton of $\mathcal{X}$ is equal to the union

$$\mathcal{S}_{\mathcal{X}}^{(k)} = \bigcup_{i=0}^{k} X_i.$$

Example 3.1.1. A simple undirected graph $G = (V, E)$ is equivalent to a 1-dimensional abstract simplicial complex on V, where the 0-simplices are the vertices of G and the 1-simplices are the edges of G. We also can say that an undirected graph is the 1-skeleton of an abstract simplicial complex.

Moreover, from the previous definitions, $\mathcal{X} \subseteq \mathcal{P}(V_{\mathcal{X}})$, where $\mathcal{P}(V_{\mathcal{X}})$ is the power-set[1] of $V_{\mathcal{X}}$ and if $\sigma, \sigma' \in \mathcal{X}$, the intersection $\sigma \cap \sigma'$ is either a face of both simplices or the empty set.

Remark 3.1.1. The power-set of the vertex set of every ASC defines a topology (called *discrete topology*) on the vertex set. Thus, every ASC defines a topological space.

A less restrictive definition is the definition of simplicial family: a *simplicial family* is a set formed by arbitrary simplices (the faces of the simplices do not need to belong to the set). Notice that every simplicial family determines an abstract simplicial complex.

As a last remark, we point out that some authors simply use the expression *simplicial complex* to refer to an abstract simplicial complex, and here we may switch between the two nomenclatures deliberately.

[1]The *power-set* $\mathcal{P}(S)$ of a given set S, also denoted by 2^S, is the set formed by all possible subsets of S, including the empty set.

Geometric Simplicial Complexes

One can associate to any abstract simplex σ a *geometric simplex*, which is the convex hull[2] of the vertices of σ in the Euclidian space. More formally:

Definition 3.1.2. Let $v_0, ..., v_k$ be $k+1$ distinct points in $\mathbb{R}^n$. We say that $\{v_0, ..., v_k\}$ is an *affinely independent set* if for real numbers $a_0, ..., a_k$ such that $\sum_{i=0}^{k} a_i v_i = 0$ and $\sum_{i=0}^{k} a_i = 0$, then $a_0 = ... = a_k = 0$. If $\{v_0, ..., v_k\}$ is an affinely independent set of points in $\mathbb{R}^n$, the *geometric simplex* spanned by it is the set

$$\sigma = \left\{ \sum_{i=0}^{k} \lambda_i v_i : \lambda_i \geq 0 \text{ and } \sum_{i=0}^{k} \lambda_i = 1 \right\}. \tag{3.1}$$

Analogously to Definition 3.1.1, a *geometric simplicial complex* (or *Euclidean simplicial complex*) $\mathcal{K}$ is defined as a finite collection of geometric simplices such that if $\sigma \in \mathcal{K}$ and $\tau \subseteq \sigma$, then $\tau \in \mathcal{K}$, and if $\sigma, \sigma' \in \mathcal{K}$, then either $\sigma \cap \sigma' \in \mathcal{K}$ or $\sigma \cap \sigma' = \emptyset$.

The *boundary* of a geometric n-simplex is constituted of all its $(n-1)$-faces, and its *interior* is constituted of all points which do not belong to its boundary. Every other definition made previously for abstract simplices is defined analogously to geometric simplices.

Example 3.1.2. Figure 3.1 presents examples of geometric k-simplices, for $k = 0, 1, 2, 3$. From left to right we have: a 0-simplex (vertex), a 1-simplex (edge), a 2-simplex (triangle with interior), and a 3-simplex (tetrahedron with interior). Note that the edge has two vertices as its faces, the triangle has three edges as its faces, and the tetrahedron has four triangles as its faces.

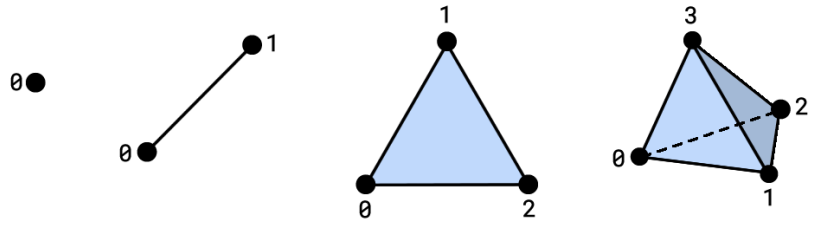


Figure 3.1: Examples of geometric simplices.

Definition 3.1.3. The *vertex scheme* of a geometric simplicial complex is the ASC built out of the sets of vertices of its geometric simplices. On the other hand, the *geometric realization* of an ASC Δ is the geometric simplicial complex whose vertex scheme is isomorphic to Δ.

Every finite ASC has a geometric realization on an Euclidean space, as stated by the geometric realization theorem: "every d-dimensional abstract simplicial complex has a geometric realization in $\mathbb{R}^{2d+1}$" (see [83], p. 53, for a proof).

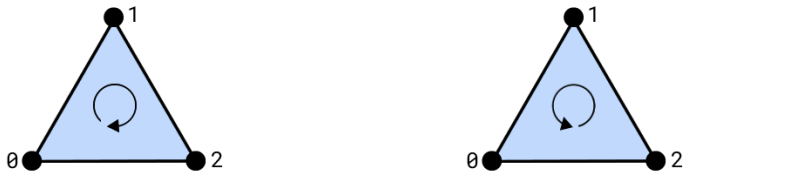
[2]The convex hull of a subset $S \subseteq \mathbb{R}^n$ is the intersection of all convex sets containing S.

Abstract Directed Simplicial Complexes

We can define an *orientation* on an n-simplex by defining a total ordering on its vertices. For instance, given an n-simplex $\{v_0, ..., v_n\}$, an orientation is obtained by ordering $v_i < v_j$, whenever $i < j$.

In the literature, it is usual to denote an *ordered* (or *oriented*) *n-simplex* by $[v_0, ..., v_n]$, which represents a totally ordered set. Just as we defined for abstract simplices, we may use the notation $\sigma^{(n)} = [v_0, ..., v_n]$, where the superscript represents its dimension. The edges of an ordered simplex inherit its ordering, i.e. $[v_i, v_j]$, $v_i < v_j$, if $i < j$. Here a caveat is necessary because we are committing an abuse of notation: i and j are actually indicating the position of the vertices in the simplex, even though we are using the same indices on the vertex labels. With that said, we'll adopt the notation $v_i = i$ for the labels, and write $[0...n] = [0, ..., n] = [v_0, ..., v_n]$, but keep in mind that the labels do not necessarily represent the position of the vertex in the ordered set.

Two orientations are equivalent if they differ by an even permutation, e.g., the ordered 2-simplex $[0, 1, 2]$ in Figure 3.2a has the same clockwise orientation as those produced by the permutations $[1, 2, 0]$ and $[2, 0, 1]$, while the permutation $[1, 0, 2]$ produces a counterclockwise orientation, as depicted in Figure 3.2b. It's important to note that the set $\{0, 1, 2\}$ corresponds to a single simplex, regardless of its orientation, and the ordered simplices $[1, 2, 0]$ and $[2, 0, 1]$ are not equal as ordered sets, despite the same orientation.



(a) Clockwise orientation. (b) Counterclockwise orientation.

Figure 3.2: Different orientations of a 2-simplex.

In summary, an ordered n-simplex is merely an n-simplex with a total ordering in its vertices. Therefore, we can formulate an analogous definition of the definition of ASC where our simplices are ordered simplices, i.e. the definition of *abstract ordered simplicial complex* or *abstract directed simplicial complex* [218].

Definition 3.1.4. An *abstract directed simplicial complex* (ADSC) is a finite collection $\mathcal{X}$ of totally ordered finite sets, such that if $\sigma \in \mathcal{X}$, then for all $\tau \subseteq \sigma$ (with an ordering inherited from σ) we have $\tau \in \mathcal{X}$ (closed under totally ordered subset inclusion).

Henceforth, we'll adopt the nomenclature *directed simplices* for the elements of an ADSC instead of ordered simplices, and we'll keep using the same notations introduced for ordered simplices. Also, all definitions and nomenclatures made previously for abstract simplicial complexes apply to abstract directed simplicial complexes as well.

It's important to point out that, as mentioned in Subsection 2.1.1, two totally ordered sets with the same elements are identical if and only if they have the same ordering, therefore, two different directed simplices may have the same set of vertices. This fact implies that, despite the name, ADSCs are not examples of ASCs in general, except in the cases where the ADSCs are built out of ASCs, since, in these cases, every ordered set in the ADSC corresponds to a single simplex.

A convenient and practical way to access the faces of a directed simplex is through *face maps*, which are defined as follows.

Definition 3.1.5. Given an abstract directed simplicial complex $\mathcal{X} = \bigcup_{k=0}^{\dim \mathcal{X}} X_k$, where X_k is the set of all directed k-simplices, we define maps $\hat{d}_i : X_{k+1} \to X_k$, for each $0 \leq k \leq \dim \mathcal{X} - 1$ and each $0 \leq i \leq k+1$, such that $\hat{d}_i([v_0, ..., v_k]) = [v_0, ..., \hat{v_i}, ..., v_k]$, where $\hat{v_i}$ denotes the deletion of the i-th vertex v_i from $[v_0, ..., v_k]$. The map $\hat{d}_i$ is called *i-th face map*.

One can easily see that indeed $\hat{d}_i$ takes a simplex on its i-th face. An important property of the face maps is that if $i < j$, then

$$\hat{d}_i \circ \hat{d}_j = \hat{d}_{j-1} \circ \hat{d}_i. \tag{3.2}$$

In fact, $\hat{d}_i(\hat{d}_j([v_0, ..., v_k])) = [v_0, ..., \hat{v_i}, ..., \hat{v_j}, ..., v_k] = \hat{d}_{j-1}(\hat{d}_i([v_0, ..., v_k]))$.

Semi-Simplicial Sets

A generalization of the concept of abstract simplicial complexes is the concept of *semi-simplicial complexes*, which was originally proposed by S. Eilenberg and J. A. Zilber [86]. The contemporary theory adopted the nomenclature *semi-simplicial sets*, and the following definition is based on Friedman's work [106].

Definition 3.1.6. A *semi-simplicial set* (or *Delta set*, or *Δ-set*) is a collection of sets $\mathcal{X} = \{X_0, X_1, X_2, ...\}$ together with maps $\hat{d}_i : X_{n+1} \to X_n$, for each $0 \leq n$ and each $0 \leq i \leq n+1$, such that $\hat{d}_i \circ \hat{d}_j = \hat{d}_{j-1} \circ \hat{d}_i$, whenever $i < j$.

Every abstract simplicial complex form a semi-simplicial set. Indeed, let X_k be the set of all k-dimensional simplices and let $d_i : X_{k+1} \to X_k$ be the face maps as defined in Definition 3.1.5. The condition $\hat{d}_i \circ \hat{d}_j = \hat{d}_{j-1} \circ \hat{d}_i$, whenever $i < j$, is satisfied as shown in Equation (3.2). Nonetheless, there exist semi-simplicial sets that cannot be built out of abstract simplicial complexes, as we can see in the following example.

Example 3.1.3. Consider the collection of sets $\mathcal{X} = \{X_0, X_1\}$, where $X_0 = \{[0], [1]\}$ and $X_1 = \{[0,1], [1,0]\}$. The face maps are $\hat{d}_0([0,1]) = \hat{d}_1([1,0]) = [1]$, $\hat{d}_1([0,1]) = \hat{d}_0([1,0]) = [0]$. Thus, $\mathcal{X}$ together with the maps $\hat{d}_i$ form a semi-simplicial set. However, it does not form an ASC, since the intersection $[0,1] \cap [1,0] = \{0,1\}$ is not a common face of both directed 1-simplices.

Similarly to an abstract simplicial complex, a semi-simplicial set has a geometric realization as described in [185].

3.1.2 Directed Flag Complexes

From this part forth, we will be mainly interested in abstract directed simplicial complexes built from directed graphs; notwithstanding, let's start by making some observations about ASCs built out of undirected graphs.

There are several ways to build an ASC from a given graph $G = (V, E)$. For instance, we can construct an ASC by considering the neighborhood of each vertex and including all its faces (*neighborhood complex*) [178]; by defining a collection of subgraphs of the power-set of V, $\mathcal{S}_G \subseteq \mathcal{P}(V)$, such that if $H \in \mathcal{S}_G$ and $e \in H$ is an edge, then $H - e \in \mathcal{S}_G$ (closed under deletion of edges) [151]; or by considering the n-simplices as the $(n+1)$-cliques of the graph, forming what is known as *clique complex* or *flag complex* [4].

Here we will focus on the study of flag complexes and their directed variants, the *directed flag complexes* [218], so let's present their formal definitions.

Definition 3.1.7. Given a graph $G = (V, E)$, its *flag complex* (or *clique complex*), denoted by $\mathrm{Fl}(G)$, is the abstract simplicial complex on the vertex set of G whose k-simplices are subsets of V which span $(k+1)$-cliques of G.

The previous definition only takes into account undirected graphs. Nevertheless, we will introduce a variation of this definition for directed graphs by using the definition of *directed cliques*.

Definition 3.1.8. A *directed* $(k+1)$*-clique* is a digraph $G = (V, E)$, $V = \{v_0, ..., v_k\}$, whose underlying undirected graph is a $(k+1)$-clique and for each $0 \leq i < j \leq k$, $v_i, v_j \in V$, and there is a directed edge from v_i to v_j, i.e. $(v_i, v_j) \in E$, for all $i < j$.

Notice that, by definition, every directed $(k+1)$-clique has a source (v_0) and a sink (v_k) and is a directed acyclic graph (DAG) (c.f. Figure 3.3).

Example 3.1.4. Figure 3.3 presents examples of directed $(k+1)$-cliques for $k = 0, 1, 2, 3, 4$. From left to right we have: a 1-clique, a directed 2-clique, a directed 3-clique, a directed 4-clique, and a directed 5-clique.

The concept of directed flag complexes [181, 218] is a variant for digraphs of the concept of flag complexes, however, although the flag complex associated with a graph is, in fact, an example of ASC, a directed flag complex associated with a digraph is actually an example of ADSC, since the digraph may contain double-edges, which implies that a subset of its vertex set may span different directed cliques; an exception occurs for digraphs *without double-edges*, in which cases the corresponding directed flag complexes are ASCs with a specific order in their simplices.

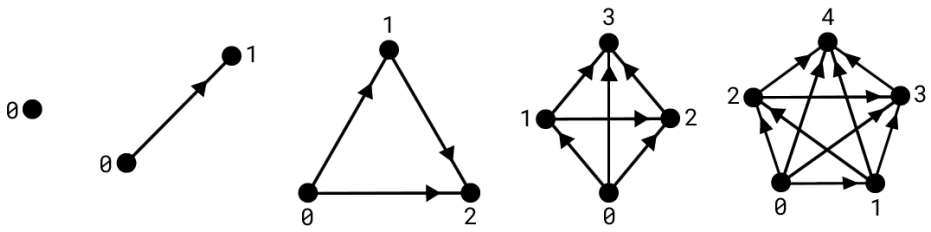


Figure 3.3: Examples of directed $(k+1)$-cliques, for $k = 0, 1, 2, 3, 4$.

Definition 3.1.9. Given a digraph $G = (V, E)$, its *directed flag complex*, denoted by $\mathrm{dFl}(G)$, is the abstract directed simplicial complex whose directed k-simplices span directed $(k+1)$-cliques of G, i.e. for every $[v_0, ..., v_k] \in \mathrm{dFl}(G)$, we have $v_i \in V$, $\forall i$, and $(v_i, v_j) \in E$, $\forall i < j$.

The directed simplices of a directed flag complex are not uniquely defined by their set of vertices, since they may differ by the ordering of their vertices. For instance, the directed 2-simplices $[v_0, v_1, v_2]$ and $[v_2, v_0, v_1]$ have the same set of vertices but they span different directed 3-cliques. Again, remember that we are committing an abuse of notation and i and j are indicating the position of the vertices in the simplex, which are independent of the vertex labels, e.g., in the case $[v_0, v_1]$ we have $(v_0, v_1) \in E$ but in the case $[v_1, v_0]$ we have $(v_1, v_0) \in E$.

Example 3.1.5. Figure 3.4a shows the digraph $G = (V, E)$ which has a double-edge between the vertices 0 and 2, i.e. $(0, 2), (2, 0) \in E$. Note that the set $\{0, 2\}$ spans two different directed 2-cliques, $[0, 2]$ and $[2, 0]$.

(a) Digraph G. (b) $\mathrm{dFl}(G)$.

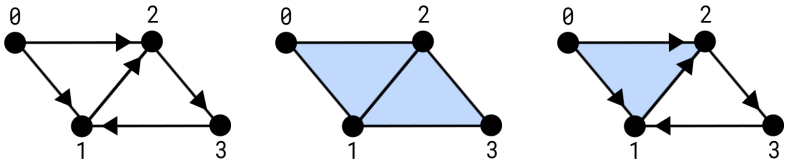
Figure 3.4: A digraph G with a double-edge and its directed flag complex $\mathrm{dFl}(G)$.

Moreover, note that directed cycles are not considered directed cliques since they do not have a source and a sink.

We say that the flag complex associated with the underlying undirected graph of a digraph is its *underlying flag complex*.

Example 3.1.6. Figure 3.5 represents a simple digraph G, its underlying flag complex Fl(G), and its directed flag complex dFl(G).

Be aware that, in this text, the colors that fill the interior of the cliques do not represent the geometric interior in the sense of Definition 3.1.2, but are used as a visual representation of the corresponding higher-dimensional directed simplices.



(a) Digraph G. (b) Fl(G). (c) dFl(G).

Figure 3.5: A digraph G along with its underlying flag complex Fl(G), and its directed flag complex dFl(G).

Remark 3.1.2. Directed cliques are combinatorial objects, thus, for instance, given a set of three vertices, we can build six directed 3-cliques by changing the directions of their edges in a proper way. They are all isomorphic because we can transform one into the other just by reordering their vertices and, since digraph isomorphism is an equivalence relation, the collection of all directed $(k+1)$-cliques form a class of equivalence. Although they are algebraically equivalent, the change in the direction of an edge can be combinatorially relevant, as it can affect the adjacent directed simplices, as we will see later in Subsection 3.3.2, when we introduce the theory of directed Q-analysis.

Observation 3.1.1. In order to build chain complexes and homology groups (see Section 3.1.4), we can associate a semi-simplicial set to a directed flag complex as follows. Given a directed flag complex of a digraph, dFl(G), let X_k be the set of all directed k-simplices of dFl(G). The collection $\{X_0, ..., X_{\omega(G)-1}\}$, where $\omega(G)$ is the clique number of G, together with face maps $d_i : X_{k+1} \to X_k$, for each $0 \leq i \leq k+1$ and each $0 \leq k \leq \omega(G) - 2$, form a simplicial set.

Govc et al. [123] proposed another example of complex built out of a digraph, called *flag tournaplex*, in which the simplices are considered to be the tournaments of the digraph. A $(k+1)$-tournament is a digraph, without double-edges, whose underlying undirected graph is a $(k+1)$-clique. Since the induced subdigraph of a tournament is also a tournament (called subtournament), in the aforementioned article, an ASC-like structure called *tournaplex* was introduced as a collection $\mathcal{X}$ of tournaments such that if $\sigma \in \mathcal{X}$ and $\tau \subseteq \sigma$ is a subtournament, then $\tau \in \mathcal{X}$; subsequently, a flag tournaplex was defined as the tournaplex built out of a given digraph. We can build a semi-simplicial set from a flag tournaplex in an analogous way as exposed in Observation 3.1.1.

3.1.3 Weighted Directed Flag Complexes

So far, we've been considering solely directed flag complexes built out of weightless digraphs; nonetheless, many real-world networks present weighted edges and/or weighted vertices, thus it would be convenient to transpose the weights associated with a digraph into its respective directed flag complex. The first formal definition of weighted simplicial complexes was proposed by Dawson [69], in which the weights were considered

to belong to the set of natural numbers. However, his definition was generalized for weights in a commutative ring (with unity) [219], thus we extended Dawson's definition to abstract directed simplicial complexes with weights on a commutative ring.

Definition 3.1.10. A *weighted abstract directed simplicial complex* is a pair $(\mathcal{X}, \widetilde{\omega})$ consisting of an abstract directed simplicial complex $\mathcal{X}$ and a function $\widetilde{\omega} : \mathcal{X} \to \mathcal{R}$, satisfying $\widetilde{\omega}(\tau)|\widetilde{\omega}(\sigma)$ whenever $\tau \subseteq \sigma$, where $\mathcal{R}$ is a commutative ring. Here, the convention $0/0 = 0$ is assumed.

Furthermore, in previous chapters, all definitions associated with weighted digraphs only considered the weights of their edges, and nothing was said about the weight of their nodes. Nevertheless, in some contexts, a function that produces an edge-weight to node-weight transformation is needed. There are several ways to define such a function, thus, in what follows, we proposed an edge-to-node weight function.

Definition 3.1.11. Let $G^{\omega} = (V, E, \omega)$ be a weighted digraph, where $\omega : E \to \mathcal{R}$ is its edge-weight function, and $\mathcal{R}$ is a commutative ring. We define a *node-weight function* $\tilde{\omega} : V \to \mathcal{R}$ based on the function ω as follows:

$$\tilde{\omega}(i) = \max\left(\deg^{-}_{\omega}(i), \deg^{+}_{\omega}(i)\right). \tag{3.3}$$

The above definition is convenient because it avoids null weights on vertices that are not isolated, and take the *in* and *out* contributions of the edges into account. However, a drawback is that it might output high values for vertices with just a few strong connections and low values for vertices with a large number of weak connections.

Definition 3.1.12. Let G^{ω} be a weighted digraph with directed flag complex $\mathrm{dFl}(G^{\omega})$, and let $\mathcal{R}$ be a commutative ring. We define the *product-weight function* $\widetilde{\omega} : \mathrm{dFl}(G^{\omega}) \to \mathcal{R}$ as

$$\widetilde{\omega}(\sigma^{(n)}) = \prod_{i=0}^{n} \tilde{\omega}(i), \tag{3.4}$$

where $\tilde{\omega}$ is an edge-to-node weight function, and $\tilde{\omega}(i)$ is the weight of the node $i \in \sigma^{(n)} = [0, 1, ..., n]$. The pair $(\mathrm{dFl}(G^{\omega}), \widetilde{\omega})$ is called *weighted directed flag complex*.

Note that the product-weight function satisfies the conditions of the Definition 3.1.10, therefore a weighted directed flag complex is indeed a weighted ADSC.

Remark 3.1.3. As defined in Chapter 2, when we use the terminology *weighted digraphs* we are referring to digraphs whose edges are weighted but the nodes aren't, thus, in order to build their weighted directed flag complexes, we must transform the edge weights into node weights, for instance, through the edge-to-node weight function proposed in Definition 3.1.11. On the other hand, if only the nodes of a digraph are weighted, then the definition of the product-weight is straightforward.

Now, we are going to make a digression and discuss briefly how we can construct simplicial complexes when considering the vertices in a metric (or pre-metric) space (see Definition 2.1.30). Different simplicial complexes can be built by establishing different criteria for the formation of simplices based on the (pre-)metric of the space, e.g., Vietoris-Rips complexes, Čech complexes, Delaunay complexes, and alpha complexes [83]. In the following, we discuss two types of complexes built in (pre-)metric spaces, namely: the Vietoris-Rips complexes and the Dowker complexes [55].

Definition 3.1.13. Let V be a set of vertices in a metric space (X, d). Given a real number $\delta > 0$, the *Vietoris-Rips complex* of V, denoted by $\mathfrak{R}_\delta(V)$, is an abstract simplicial complex formed by simplices whose diameters are at most δ, i.e.

$$\mathfrak{R}_\delta(V) = \{\sigma \subseteq V : \max_{v,w\in\sigma} d(v, w) \leq \delta\}. \tag{3.5}$$

A drawback of the Vietoris-Rips complex is that it is insensitive to asymmetry, since the metric satisfies the symmetry condition. Also, since we are dealing with weighted directed networks, and since we can define a pre-metric from a weight function (see Definition 2.1.33), we would like to extend the concept of Vietoris-Rips complex to pre-metric spaces. Chowdhury and Mémoli [55] introduced the *Dowker complexes* for weighted directed networks. In what follows, we present the definition proposed in the aforementioned article for pre-metric spaces.

Definition 3.1.14. Let V be a set of vertices in a pre-metric space (X, d) and let $R_\delta(V) \subseteq V \times V$ be the following relation:

$$R_\delta(V) = \{(v, w) : d(v, w) \leq \delta\}, \tag{3.6}$$

for any $\delta \in \mathbb{R}_+$. The *pre-metric Dowker δ-sink complex*, the *pre-metric Dowker δ-source complex* and the *pre-metric Dowker complex*, denoted, respectively, by $\mathfrak{D}^{pre}_{si}$, $\mathfrak{D}^{pre}_{so}$, and $\mathfrak{D}^{pre}$, are defined by

$$\mathfrak{D}^{pre}_{si}(V; \delta) = \{\sigma = [v_0, ..., v_n] : \exists w \in V \text{ such that } (v_i, w) \in R_\delta(V), \forall v_i \in \sigma\}, \tag{3.7}$$

$$\mathfrak{D}^{pre}_{so}(V; \delta) = \{\sigma = [v_0, ..., v_n] : \exists w \in V \text{ such that } (w, v_i) \in R_\delta(V), \forall v_i \in \sigma\}, \tag{3.8}$$

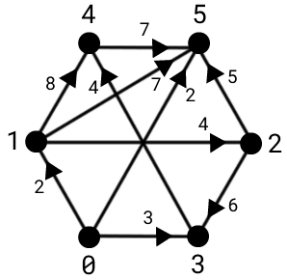
$$\mathfrak{D}^{pre}(V; \delta) = \{\sigma = [v_0, ..., v_n] : \max_{v_i,v_j\in\sigma} d(v_i, v_j) \leq \delta\}. \tag{3.9}$$

By the previous definition, if $G^\omega = (V, E, \omega)$ is a weighted digraph, and if we define the pre-metric $d = D^\omega$ such as in Definition 2.1.33, then , $\mathfrak{D}^{pre}_{si}(V; \delta), \mathfrak{D}^{pre}_{so}(V; \delta), \mathfrak{D}^{pre}(V; \delta) \subseteq \mathrm{dFl}(G^\omega)$, for any real number $\delta > 0$.

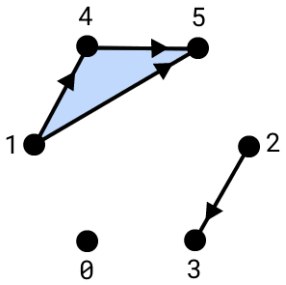
Example 3.1.7. Let $G^\omega = (V, E, \omega)$ be the weighted digraph illustrated in Figure 3.6a, whose edge weights are not normalized. Let $d = D^\omega$ be the pre-metric defined by the

formula (2.7) but replacing $\omega_{ij}^{-1} - 1$ with ω_{ij}^{-1}, as explained in Observation 2.1.1. Then, the pre-metric Dowker complex obtained by applying the pre-metric d in V with the condition $d(i,j) \leq \delta = 0.17$, $\forall i, j \in V$, is

$$\mathfrak{D}^{pre}(V;\delta) = \{[0],[1],[2],[3],[4],[5],[2,3],[1,4],[4,5],[1,5],[1,4,5]\}.$$



(a) A weighted digraph G^{ω}.



(b) The pre-metric Dowker complex built from the weighted digraph G^{ω}.

Figure 3.6: A weighted digraph and its pre-metric Dowker complex for $\delta = 0.17$.

3.1.4 Simplicial Homology

Historically, Henri Poincaré, in his 1895 paper, *Analysis situs* [209], where he quoted the work of his predecessors Riemann and Betti, was the first person to introduce the concept of homology classes for cell complexes [80]. Since then, homology theory has evolved to become one of the main branches of algebraic topology. The basic idea behind homology theory is to define algebro-topological invariants of topological spaces, which are used to distinguish between these spaces. Homology classes of a topological space are equivalence classes that represent topological invariants, and these classes form the so-called *homology groups* (or just *homologies*, when seen as vector spaces) associated with the space. The dimensions of these groups, called *Betti numbers*, roughly speaking, represent the numbers of "n-dimensional holes" in the space, and are also topological invariants [138].

A particular application of homology theory is on simplicial complexes. As already discussed in Subsection 3.1.2, we can build simplicial complexes out of the clique complexes (directed clique complexes) of graphs (digraphs without double-edges), and then use the algebro-topological framework from homology theory to study the combinatorial and topological structures associated with them.

This section is mainly based on the book [138], and we consider homology theory only for simplicial complexes. Moreover, throughout this part, $\mathbb{K}$ will denote a fixed field and $\mathcal{X}$ will denote, unless said otherwise, the directed flag complex of a given digraph *without double-edges*.

Definition 3.1.15. For a non-negative integer k, we define $C_k(\mathcal{X};\mathbb{K})$ as the $\mathbb{K}$-vector space formed by all formal $\mathbb{K}$-linear combinations of all k-dimensional elements of $\mathcal{X}$, and we call these spaces *chain spaces*. The elements of $C_k(\mathcal{X};\mathbb{K})$ are called *k-chains* and are denoted by $c = \sum a_i \sigma_i^{(k)}$, where $a_i \in \mathbb{K}$ and $\sigma_i^{(k)}$ are directed k-simplices.

Definition 3.1.16. For any integer $n \geq 1$, we define a linear map $\partial_n : C_n(\mathcal{X};\mathbb{K}) \to C_{n-1}(\mathcal{X};\mathbb{K})$, called *$n$-th boundary map*, by

$$\partial_n(\sigma^{(n)}) = \sum_{i=0}^{n} (-1)^i \hat{d}_i(\sigma^{(n)}) = \sum_{i=0}^{n} (-1)^i [v_0, ..., \hat{v}_i, ..., v_n], \tag{3.10}$$

for any directed n-simplex $\sigma^{(n)} = [v_0, ..., v_n]$, where $\hat{d}_i$ is the i-th face map. For $n = 0$, we define $C_{-1}(\mathcal{X};\mathbb{K}) := \{0\}$, and $\partial_0 = 0$ (null map).

Example 3.1.8. Consider the directed 2-simplex $\sigma^{(2)} = [0, 1, 2]$. Applying the 2-boundary map on $\sigma^{(2)}$, we obtain the following 1-chain: $\partial_2(\sigma^{(2)}) = [1, 2] - [0, 2] + [0, 1]$.

Definition 3.1.17. The *chain complex* of $\mathcal{X}$ is defined as a sequence of chain spaces connected by boundary maps:

$$\ldots \xrightarrow{\partial_{n+1}} C_n(\mathcal{X};\mathbb{K}) \xrightarrow{\partial_n} C_{n-1}(\mathcal{X};\mathbb{K}) \xrightarrow{\partial_{n-1}} \ldots \xrightarrow{\partial_2} C_1(\mathcal{X};\mathbb{K}) \xrightarrow{\partial_1} C_0(\mathcal{X};\mathbb{K}) \xrightarrow{\partial_0} \{0\}.$$

As usual, we denote the chain complex by $(C_\bullet(\mathcal{X};\mathbb{K}), \partial_\bullet)$.

Definition 3.1.18. For an integer $n \geq 0$, the elements of the image $\operatorname{Im}\partial_{n+1}$ are called *n-boundaries*, and the elements of the kernel $\ker\partial_n$ are called *n-cycles*. Since $C_{-1}(\mathcal{X};\mathbb{K}) = \{0\}$, every vertex has boundary equal to 0, thus $\ker\partial_0 = C_0(\mathcal{X};\mathbb{K})$. The notations $B_n = \operatorname{Im}\partial_{n+1}$ and $Z_n = \ker\partial_n$ are also commonly used.

In addition, if an n-cycle is not a boundary of any $(n+1)$-chain (i.e. it does not belong to $\operatorname{Im}\partial_{n+1}$), then it is called an *independent n-cycle*.

The following proposition is referred to as the *fundamental lemma of homology* and it states that the boundary of a boundary is always zero.

Proposition 3.1.1. *For any integer $n \geq 0$, the identity $\partial_n \circ \partial_{n+1} = 0$ holds.*

A proof for the previous proposition can be found in [138], p. 105. Moreover, by this proposition, we have the inclusion $\operatorname{Im}\partial_{n+1} \subseteq \ker\partial_n$, since $\partial_n(\partial_{n+1}c) = 0$, for any $(n+1)$-chain c.

Definition 3.1.19. Given an integer $n \geq 0$, the *n-th homology* of $\mathcal{X}$ (over $\mathbb{K}$) is defined as the quotient vector space

$$H_n(\mathcal{X}) = H_n(\mathcal{X};\mathbb{K}) = \ker\partial_n / \operatorname{Im}\partial_{n+1}. \tag{3.11}$$

The elements of the space $H_n(\mathcal{X})$ are called *homology classes* (equivalence classes of independent cycles), and are denoted by $[c]$, for a given chain c. Two cycles are said to be *homologous* if they belong to the same homology class. The dimension of $H_n(\mathcal{X})$, denoted by $\beta_n = \beta_n(\mathcal{X}) = \dim H_n(\mathcal{X})$, is called *$n$-th Betti number*.

Notice that $H_n(\mathcal{X}) = 0$, for all $n > \dim \mathcal{X}$. Also, the n-th Betti number is equal to $\beta_n = \dim \ker \partial_n - \dim \operatorname{Im} \partial_{n+1}$, which is equal to the number of "n-dimensional holes" in $\mathcal{X}$. In particular, as $\mathcal{X} = \mathrm{dFl}(G)$ for some given digraph G, using an argument analogous to that used for graphs in [117], the 0-th Betti number is equal to the number of weakly connected components of G.

Example 3.1.9. Consider $\mathcal{X} = \{[0], [1], [2], [0,1], [1,2], [2,0]\}$. Applying ∂_1 on the 1-chain $c = [0,1] + [1,2] + [2,0]$, we have $\partial_1(c) = (1-0) + (2-1) + (0-2) = 0$. Thus, c is a 1-cycle. However, c is not a boundary of any 2-chain, then c is an independent 1-cycle.

Let G be the digraph present in Figure 3.7a (a cycle of length 3) and let $\mathcal{X}$ be its directed flag complex. The chain formed by the formal sum of the directed 1-simplices of $\mathcal{X}$ form an independent 1-cycle (Example 3.1.9), and since $\beta_1(\mathcal{X}) = 1$, we say that there is a 1-dimensional hole associated with G. On the other hand, there is no 1-dimensional hole associated with the digraph present in Figure 3.7b (directed 3-clique).

(a) Cycle of length 3. (b) A directed 3-clique.

Figure 3.7: There is a 1-dimensional hole associated with a 3-cycle, but there is no 1-dimensional hole associated with a directed 3-clique.

An important topological invariant associated with a simplicial complex is the *Euler characteristic*, which can be defined in terms of Betti numbers as follows.

Definition 3.1.20. The *Euler characteristic* of $\mathcal{X}$ is defined as the alternating sum of its Betti numbers:

$$\chi(\mathcal{X}) = \sum_{n=0}^{\infty} (-1)^n \beta_n(\mathcal{X}) = \sum_{n=0}^{\infty} (-1)^n \dim H_n(\mathcal{X}). \tag{3.12}$$

Remark 3.1.4. As commented in Subsection 3.1.2, when a digraph has double-edges and they are taken into account, the corresponding directed flag complex is a semi-simplicial set, and then the corresponding formalism must be applied (see Observation

3.1.1). In this case, the homologies might be different from those obtained when double-edges are ignored.

Furthermore, as previously mentioned, many real-world networks are weighted, so it would be interesting to take weights into account when calculating homologies. Dawson [69] proposed a generalization of the boundary map for weighted simplicial complexes, and later other authors proposed other generalizations based on Dawson's first proposal [219, 281]. Below, we present the *weighted n-th boundary map*, as proposed by Dawson, for a weighted directed flag complex $(\mathcal{X}, \widetilde{\omega})$, which is a direct modification of the formula (3.10):

$$\partial_n^{\omega}(\sigma^{(n)}) = \sum_{i=0}^{n}(-1)^i \frac{\widetilde{\omega}(\sigma^{(n)})}{\widetilde{\omega}(\hat{d}_i(\sigma^{(n)}))}\hat{d}_i(\sigma^{(n)}). \tag{3.13}$$

All previous definitions involving the boundary map are defined analogously for the weighted boundary map, and the identity $\partial_n^{\omega} \circ \partial_{n+1}^{\omega} = 0$ also holds for any $n \geq 0$. In particular, if the weights of the directed simplices are all the same (but non-zero), then both boundary maps are identical. Moreover, since $(\mathcal{X}, \widetilde{\omega})$ is a weighted directed flag complex, where $\widetilde{\omega}$ is the product-weight (3.4), for any directed n-simplex $\sigma^{(n)} = [v_0, ..., v_n]$, we have $\widetilde{\omega}(\sigma^{(n)})/\widetilde{\omega}(\hat{d}_i(\sigma^{(n)})) = \tilde{\omega}(v_i)$. Finally, we emphasize that depending on the choice of the weights, the homologies for the weighted and the unweighted cases might be different [281].

3.1.5 Persistent Homology

Persistent homology is one of the main tools in the field of topological data analysis for computing topological features of a space at different scales and their persistence across these scales [83]. It originated with Frosini and collaborators [111] in the study of persistence of 0-dimensional homology for shape recognition (which was referred as *size theory*) and it was further developed independently by Edelsbrunner et al. [84, 293]. An important property of persistent homology is that it is a robust method with respect to small perturbations in the input dataset [61].

This section is mainly based on the book [83], and we consider persistent homology only for the case of simplicial complexes. Also, throughout this part, $\mathbb{K}$ will denote a fixed field and $\mathcal{X}$ will denote, unless said otherwise, the directed flag complex of a given digraph *without double-edges*.

Filtrations

Definition 3.1.21. Given a directed flag complex $\mathcal{X}$, a *filtration* of $\mathcal{X}$ is an indexed family of subcomplexes $\mathcal{X}_i \subseteq \mathcal{X}$, $\{\mathcal{X}_i\}_{i \in I_k^*}$, such that $\mathcal{X}_i \subseteq \mathcal{X}_j$, whenever $i \leq j$, and it

can be represented as a nested sequence of subcomplexes:

$$\emptyset = \mathcal{X}_0 \subseteq \mathcal{X}_1 \subseteq \mathcal{X}_2 \subseteq \cdots \subset \mathcal{X}_k = \mathcal{X}. \tag{3.14}$$

We say that $\mathcal{X}$ together with a filtration is a *filtered directed flag complex*.

As we advance in the sequence of a filtration (increasing the indexes), topological features of the simplicial complex, such as independent cycles, may appear and disappear.

Example 3.1.10. Consider $\mathcal{X} = \{[0], [1], [2], [0, 1], [1, 2], [0, 2], [0, 1, 2]\}$. The subcomplexes $\mathcal{X}_1 = \{0], [1], [2]\}$, $\mathcal{X}_2 = \{0], [1], [2], [0, 1]\}$, $\mathcal{X}_3 = \{0], [1], [2], [0, 1], [1, 2], [0, 2]\}$, and $\mathcal{X}_4 = \mathcal{X}$, satisfy $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \mathcal{X}_3 \subset \mathcal{X}_4$, thus they form a filtration of $\mathcal{X}$ (see Figure 3.8). We let $\mathcal{X}_0 = \emptyset$ implicit in the filtration .

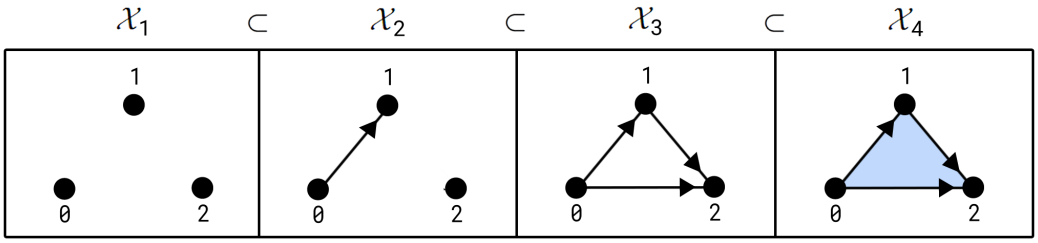


Figure 3.8: Example of filtration.

There are several different ways to create different filtrations for a (unweighted or weighted) simplicial complex [83, 219]. One way is by considering it vertices in a metric space, and then producing simplices (and thus subcomplexes) by gradually increasing the threshold on the distance between the vertices, similarly as specified in the Definition 3.1.13 of the Vietoris-Rips complex. These filtrations are called *metric filtrations*.

Analogously, in the case where $\mathcal{X}$ is associated with a weighted digraph $G = (V, E)$, we can obtain a filtration of $\mathcal{X}$ by considering the vertex set V in a pre-metric space and then compute pre-metric Dowker complexes $\mathfrak{D}^{pre}(V; \delta)$ (Definition 3.1.14) by gradually increasing the threshold δ on the pre-metric obtained from the weight function, that is, the subcomplexes $\mathcal{X}_i \subseteq \mathcal{X}$ will be $\mathcal{X}_i = \mathfrak{D}^{pre}(V; \delta_i) \subseteq \mathcal{X}_j = \mathfrak{D}^{pre}(V; \delta_j)$, such that $\delta_i \leq \delta_j$, whenever $i \leq j$. We call this filtration *pre-metric Dowker filtration*, and denote it by $\{\mathfrak{D}^{pre}(V; \delta_i) \subseteq \mathfrak{D}^{pre}(V; \delta_j)\}_{\delta_i \leq \delta_j}$. However, if we consider the pre-metric Dowker δ-sink complex or the pre-metric Dowker δ-source complex, the corresponding filtrations $\{\mathfrak{D}^{pre}_{si}(V; \delta_i) \subseteq \mathfrak{D}^{pre}_{si}(V; \delta_j)\}_{\delta_i \leq \delta_j}$ and $\{\mathfrak{D}^{pre}_{so}(V; \delta_i) \subseteq \mathfrak{D}^{pre}_{so}(V; \delta_j)\}_{\delta_i \leq \delta_j}$, respectively, are sensitive to directionality [54].

Another way to define a filtration for a weighted directed flag complex $(\mathcal{X}, \widetilde{\omega})$ based on its weights is by establishing thresholds on the product-weight function: $\mathcal{X}_i = \{\sigma \in \mathcal{X} : \widetilde{\omega}(\sigma) \leq \delta_i\}$, for given positive real numbers δ_i.

Persistent Homology

Persistent homology essentially deals with the study of the "lifetime persistence" of the topological features of a space. In the following, we present the mathematical theory behind persistent homology.

Definition 3.1.22. Let $f : \mathcal{X} \to \mathcal{X}'$ be a map between two directed flag complexes. For an integer $n \geq 0$, the map f induces a $\mathbb{K}$-linear map

$$\begin{aligned} \widetilde{f}_n : C_n(\mathcal{X}; \mathbb{K}) &\longrightarrow \quad C_n(\mathcal{X}'; \mathbb{K}), \\ \sum_i a_i \sigma_i^{(n)} &\mapsto \sum_i a_i f(\sigma_i^{(n)}), \end{aligned} \tag{3.15}$$

where $a_i \in \mathbb{K}$. The map $\widetilde{f}_n$ is called *n-th chain map*.

If $(C_\bullet(\mathcal{X}; \mathbb{K}), \partial_\bullet)$ and $(C_\bullet(\mathcal{X}'; \mathbb{K}), \partial'_\bullet)$ denote the chain complexes associated with $\mathcal{X}$ and $\mathcal{X}'$, respectively, the n-th chain map satisfies $\widetilde{f}_n \circ \partial_{n+1} = \partial'_{n+1} \circ \widetilde{f}_{n+1}$, for all $n \geq 0$, which implies that $\widetilde{f}_n(\operatorname{Im} \partial_{n+1}) \subseteq \operatorname{Im} \partial'_{n+1}$ and $\widetilde{f}_n(\ker \partial_n) \subseteq \ker \partial'_n$, i.e. it takes n-cycles to n-cycles and n-boundaries to n-boundaries. Accordingly, $\widetilde{f}_n$ induces a linear map

$$\begin{aligned} f_n : H_n(\mathcal{X}) &\to H_n(\mathcal{X}'), \\ [c] &\mapsto [\widetilde{f}_n(c)]. \end{aligned} \tag{3.16}$$

Let $\{\mathcal{X}_i\}_{i \in I_k^*}$ be a filtration of $\mathcal{X}$. For subcomplexes $\mathcal{X}_i \subseteq \mathcal{X}_j$, $i \leq j$, we have a natural inclusion map $\mathcal{X}_i \hookrightarrow \mathcal{X}_j$ that, based on the previous discussion, induces a linear map between their n-th homologies, for each $n \geq 0$:

$$f_n^{i,j} : H_n(\mathcal{X}_i) \to H_n(\mathcal{X}_j). \tag{3.17}$$

Consequently, by considering the entire filtration, for each $n \geq 0$ we have a sequence of homologies connected by linear maps:

$$\{0\} = H_n(\mathcal{X}_0) \xrightarrow{f_n^{0,1}} H_n(\mathcal{X}_1) \xrightarrow{f_n^{1,2}} \ldots \xrightarrow{f_n^{k-1,k}} H_n(\mathcal{X}_k) = H_n(\mathcal{X}). \tag{3.18}$$

Similarly as mentioned earlier for filtrations, as we advance in the above sequence of homologies, homology classes may appear and disappear or, using the usual terminology within the computational topology literature, they may be "born" and "die." The persistent homologies, as defined below, try to capture the "lifetimes" (persistence) of these homology classes.

Definition 3.1.23. Given $\mathcal{X}$ with a filtration $\{\mathcal{X}_i\}_{i \in I_k^*}$, the *$n$-th persistent homologies*,

$H_n^{i,j}$, are the images of the linear maps $f_n^{i,j}$, for $0 \leq i \leq j \leq k$, i.e.

$$H_n^{i,j} = \operatorname{Im} f_n^{i,j} = \ker \partial_n(\mathcal{X}_i)/(\operatorname{Im} \partial_{n+1}(\mathcal{X}_j) \cap \ker \partial_n(\mathcal{X}_i)). \tag{3.19}$$

The dimensions of $H_n^{i,j}$, denoted by $\beta_n^{i,j} = \dim H_n^{i,j}$, are called *n-th persistent Betti numbers*.

Note that $H_n^{i,i} = H_n(\mathcal{X}_i)$, since $\operatorname{Im} \partial_{n+1}(\mathcal{X}_i) \cap \ker \partial_n(\mathcal{X}_i) = \operatorname{Im} \partial_{n+1}(\mathcal{X}_i)$. Also, we say that a homology class $[c] \in H_n(\mathcal{X}_i)$ is *born at* $\mathcal{X}_i$ if $[c] \notin H_n^{i-1,i}(\mathcal{X}_i)$; and, if $[c]$ is born at $\mathcal{X}_i$, we say that $[c]$ *dies entering* $\mathcal{X}_j$ if $f_n^{i,j-1}([c]) \notin H_n^{i-1,j-1}$ but $f_n^{i,j}([c]) \in H_n^{i-1,j}$. The *persistence* of a homology class that is born at $\mathcal{X}_i$ and dies entering $\mathcal{X}_j$ is defined as the difference $j - i$.

Definition 3.1.24. Considering the n-th persistent Betti numbers $\beta_n^{i,j}$ of the n-th persistent homologies $H_n^{i,j}$, for $n \geq 0$ and $0 \leq i \leq j \leq k$, we define the *pairing number*, denoted by $\mu_n^{i,j}$, as the number of independent classes of dimension n that are born at $\mathcal{X}_i$ and die entering $\mathcal{X}_j$, and it is given by

$$\mu_n^{i,j} = (\beta_n^{i,j-1} - \beta_n^{i,j}) - (\beta_n^{i-1,j-1} - \beta_n^{i-1,j}), \tag{3.20}$$

where $(\beta_n^{i,j-1} - \beta_n^{i,j})$ represents the number of homology classes that are born at or before $\mathcal{X}_i$ and die entering $\mathcal{X}_j$, and $(\beta_n^{i-1,j-1} - \beta_n^{i-1,j})$ represents the number of classes that are born at or before $\mathcal{X}_{i-1}$ and die entering $\mathcal{X}_j$.

The pairing numbers can be used to visualize the *birth* and *death* of homology classes through the *persistence diagram* of a filtration, as formally defined below.

Definition 3.1.25. Given a filtration $\mathcal{F} = \{\mathcal{X}_i\}_{i \in I_k^*}$, the *n-th persistence diagram* of $\mathcal{F}$, denoted by $\operatorname{Dgm}_n(\mathcal{F})$, is a multiset of points (i, j), $0 \leq i \leq j \leq k$, with multiplicities $\mu_n^{i,j}$, in the extended plane $\bar{\mathbb{R}}^2$, where $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, such that the vertical distance of (i, j) to the diagonal is equal to $j - i$, i.e.

$$\operatorname{Dgm}_n(\mathcal{F}) = \{\!\{(i,i) \in \bar{\mathbb{R}}^2 : \mu_n^{i,i} = \infty\}\!\} \cup \{\!\{(i,j) \in \bar{\mathbb{R}}^2 : i < j, \mu_n^{i,j} < \infty\}\!\}. \tag{3.21}$$

As a consequence, the n-th persistent Betti numbers can be computed by counting the points (with multiplicities) in the respective n-th persistence diagram, as stated by the *fundamental lemma of persistent homology* (FLPH) [83].

Proposition 3.1.2. (FLPH) Given a filtration $\{\mathcal{X}_i\}_{i \in I_k^*}$, for $n \geq 0$, the n-th persistent Betti number is equal to $\beta_n^{i,j} = \sum_{r \leq i} \sum_{l > j} \mu_n^{r,l}$, for each pair i, j, with $0 \leq i \leq j \leq k$.

Moreover, from the FLPH, we can create a related function as follows.

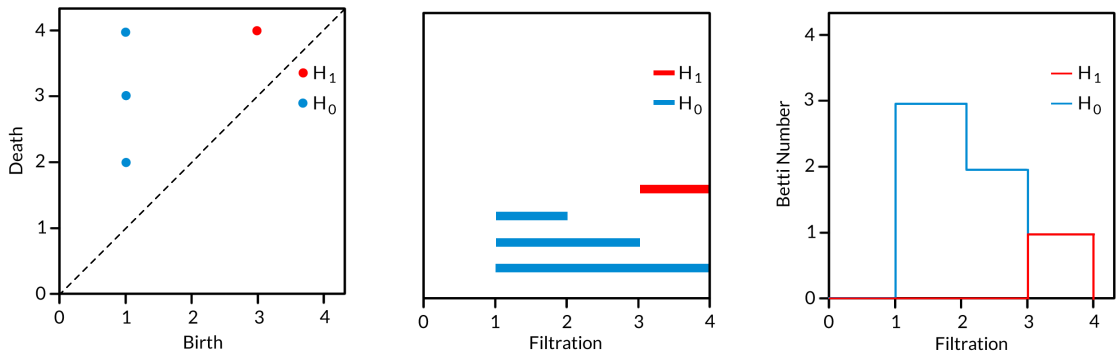
Definition 3.1.26. Given a filtration $\mathcal{F} = \{\mathcal{X}_i\}_{i \in I_k^*}$, we define the *Betti function* as

$$\mathcal{B}_n(t) = \#\{\!\{(i,j) \in \mathrm{Dgm}_n(\mathcal{F}) : i \leq t < j\}\!\}. \tag{3.22}$$

The plot of the Betti function in the plane is called *Betti curve*.

It's clear that $\mathcal{B}_n(t) = \beta_n^{i,j}$, for $i \leq t < j$. Figure 3.9a shows the persistence diagram corresponding to the filtration presented in Example 3.1.10 (Figure 3.8), and Figure 3.9c shows the corresponding Betti curves.

An alternative to persistence diagrams to visualize the birth and death of topological features in a given filtration are the *persistence barcodes* [50], in which, roughly speaking, each bar of length $j - i$ represents a topological feature that is born at $\mathcal{X}_i$ and dies entering $\mathcal{X}_j$. Figure 3.9b shows the persistence barcodes corresponding to the aforementioned example.



(a) Persistence diagram. (b) Persistence barcodes. (c) Betti curves.

Figure 3.9: The persistence diagram, persistence barcodes, and Betti curves correspond to the filtration of the Example 3.1.10 (see Figure 3.8).

Distances for Persistence Diagrams and Betti Curves

One can compare the topology of two different directed flag complexes by comparing the similarity of their persistence diagrams. The most common similarity measures between persistence diagrams are the *bottleneck* and *Wasserstein* distances.

Definition 3.1.27. Given two persistent diagrams, P and Q, let $\eta : P \to Q$ denote a bijection (perfect matching) between them, and let $||\cdot||_\infty$ denote the ∞-norm in the plane $\mathbb{R}^2$, i.e. $||\mathbf{v}||_\infty = \max_i(|v_i|)$, for $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2$. The *bottleneck distance* between P and Q is define as

$$d_{W_\infty}(P,Q) = \inf_{\eta: P \to Q} \sup_{x \in P} ||x - \eta(x)||_\infty. \tag{3.23}$$

We can verify that d_{W_∞} is indeed a distance since it satisfies the three conditions present in Definition 2.1.30 (formal definition of distance).

Definition 3.1.28. Given two persistent diagrams, P and Q, and a bijection $\eta : P \to Q$ between them, the *p-Wasserstein distance* between P and Q is defined as

$$d_{W_p}(P,Q) = \inf_{\eta:P\to Q} \left(\sum_{x\in P} ||x - \eta(x)||_\infty^p \right)^{1/p}. \tag{3.24}$$

Note that when $p \to \infty$, the p-Wasserstein distance becomes the bottleneck distance; thus, we can identify the bottleneck distance as the ∞-Wasserstein distance.

A notable feature of the persistence diagrams is that, in a space provided with a Wasserstein distance, they are *stable* (or robust) against perturbations ("noise"), i.e. small perturbations in the filtration produce small changes in the respective persistence diagram, as proved in various stability theorems [61, 83].

Furthermore, we can use the L_p-norm to define a distance between two Betti curves as follows.

Definition 3.1.29. Given two persistent diagrams, P and Q, with respective Betti functions, $\mathcal{B}_P(x)$ and $\mathcal{B}_Q(x)$, we define de *p-Betti distance* between their respective Betti curves as

$$d_{\mathcal{B}}(\mathcal{B}_P, \mathcal{B}_Q) = \left(\int_{\mathbb{R}} |\mathcal{B}_P(x) - \mathcal{B}_Q(x)|^p dx \right)^{1/p}. \tag{3.25}$$

In general, for computational purposes, the most used parameters are $p = 1$ or $p = 2$.

3.1.6 Combinatorial Hodge Laplacian

As studied in Chapter 2, the methods of spectral graph theory can capture structural properties of graphs by using eigenvalues and eigenvectors of their matrix representations, such as the adjacency and Laplacian matrices. Likewise, spectral simplicial theory, a generalization of the spectral graph theory for simplicial complexes, can reveal insightful structural information about simplicial complexes through the eigenvalues and eigenvectors associated with their *combinatorial Laplacians* (also called *Hodge Laplacians*, due to their connection with Hodge theory), which are a generalization to higher-orders of the graph Laplacian [107, 254].

Before introducing the formal definition of the Hodge Laplacian, let's remember that, given a vector space with inner product, $(X, \langle\cdot,\cdot\rangle)$, and a linear operator $f : X \to X$, the *adjoint operator* of f is a linear operator $f^* : X \to X$ satisfying $\langle f(x), y\rangle = \langle x, f^*(y)\rangle$, for all $x, y \in X$.

Just as before, throughout this part, $\mathbb{K}$ will denote a fixed field and $\mathcal{X}$ will denote the directed flag complex of a given digraph *without double-edges*.

Definition 3.1.30. Let $(C_\bullet(\mathcal{X},\mathbb{K}),\partial_\bullet)$ denote the chain complex of $\mathcal{X}$. Let ∂_n^* denote the adjoint operator of ∂_n, for $n \geq 0$. The *(combinatorial) Hodge n-Laplacian operator*, $\mathcal{L}_n : C_n(\mathcal{X},\mathbb{K}) \to C_n(\mathcal{X},\mathbb{K})$, is defined by

$$\mathcal{L}_n = \partial_{n+1} \circ \partial_{n+1}^* + \partial_n^* \circ \partial_n. \tag{3.26}$$

The higher-order boundary maps ∂_n induce matrices B_n that can be interpreted as higher-order incidence matrices between the directed simplices and their (co-)faces. For instance, the matrix B_1 is the vertex-to-arc incidence matrix (see Definition 2.2.2), and B_2 is the arc-to-(2-simplex) incidence matrix, and so on. Thus, let $[\partial_n] = B_n$ and $[\partial_n^*] = B_n^T$ be the matrix representations of the n-th boundary operator and its adjoint operator, respectively, the matrix representation of the Hodge n-Laplacian operator is given by

$$[\mathcal{L}_n] = B_{n+1}B_{n+1}^T + B_n^T B_n. \tag{3.27}$$

As commented before, the Hodge n-Laplacian is a generalization of the graph Laplacian for simplicial complexes. Indeed, for $n = 0$, we have $[\mathcal{L}_0] = BB^T$ (graph Laplacian). Also, since $\partial_n \circ \partial_{n+1} = 0$, we have $B_n B_{n+1} = 0$, for all $n \geq 0$. Occasionally, the following notations are used: $\mathcal{L}_n^u = \partial_{n+1} \circ \partial_{n+1}^*$ (*upper Laplacian*) and $\mathcal{L}_n^l = \partial_n^* \circ \partial_n$ (*lower Laplacian*), thus $\mathcal{L}_n = \mathcal{L}_n^u + \mathcal{L}_n^l$.

Observation 3.1.2. We denote the *n-Laplacian spectrum*, $n \geq 0$, of the Hodge n-Laplacian matrix by $\{\mu_i^n\}_{i=1}^{max}$ and, analogously to the graph Laplacian, we represent the eigenvalues of this spectrum in an increasing order: $\mu_1^n \leq \ldots \leq \mu_{max}^n$.

Proposition 3.1.3. *The Hodge n-Laplacian matrix is positive semi-definite and all its eigenvalues are non-negative.*

The proof of the previous proposition is analogous to the proof of the Proposition 2.2.2 together with the proof of the Proposition 2.2.3.

Furthermore, the kernel of the Hodge n-Laplacian operator associated with $\mathcal{X}$ is isomorphic to the n-th homology of $\mathcal{X}$, i.e. $ker(\mathcal{L}_n) \cong H_n(\mathcal{X})$, which implies that the number of zero-eigenvalues of $[\mathcal{L}_n]$ is equal to the n-th Betti number [107]. Accordingly, the Hodge Laplacians provide valuable topological information about the complex.

Example 3.1.11. Consider the digraph G as depicted in Figure 3.10. Its directed flag complex is $\mathcal{X} = \{[0],[1],[2],[3],[0,1],[0,2],[1,2],[1,3],[2,3],[0,1,2],[1,2,3]\}$. Therefore, the higher-order incidence matrices B_n associated with $\mathcal{X}$ are

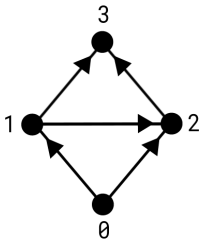


Figure 3.10: Digraph G.

$$B_1 = \begin{array}{c} \\ [0] \\ [1] \\ [2] \\ [3] \end{array} \begin{array}{c} \begin{array}{ccccc} [01] & [02] & [12] & [13] & [23] \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix} \end{array}, \quad B_2 = \begin{array}{c} \\ [01] \\ [02] \\ [12] \\ [13] \\ [23] \end{array} \begin{array}{c} \begin{array}{cc} [012] & [123] \end{array} \\ \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \\ 0 & -1 \\ 0 & 1 \end{bmatrix} \end{array}.$$

Note that $B_n = 0$ for all $n \geq 3$. Thus, by the formula 3.27, the Hodge n-Laplacians associated with $\mathcal{X}$ are:

$$[\mathcal{L}_0] = B_1 B_1^T = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix},$$

$$[\mathcal{L}_1] = B_2 B_2^T + B_1^T B_1 = \begin{bmatrix} 3 & 0 & 0 & -1 & 0 \\ 0 & 3 & 0 & 0 & -1 \\ 0 & 0 & 4 & 0 & 0 \\ -1 & 0 & 0 & 3 & 0 \\ 0 & -1 & 0 & 0 & 3 \end{bmatrix}, \quad [\mathcal{L}_2] = B_2^T B_2 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix},$$

and $[\mathcal{L}_n] = 0$ for all $n \geq 3$.

3.2 Path Complexes of Digraphs

In the literature, there are different approaches to constructing complexes and homologies from digraphs. For instance, one can build homologies from the directed clique complexes of digraphs, as shown in Subsection 3.1.4; one can also build Hochschild homologies out of the path algebras of digraphs [136]. Nevertheless, a novel type of complexes and homologies associated with digraphs that generalizes the concept of directed clique complexes, called *path complexes* and *path homologies*, was introduced and developed by Grigor'yan et al. [126, 127, 128, 129, 130, 131, 132, 133]. Some in-

teresting properties of this new formalism are that the chain complexes associated with path complexes might contain more digraph substructures than just directed cliques, and we can have non-trivial path homologies of all dimensions.

In the following, we present the formalism of path complexes and path homologies as presented in the aforementioned articles, preserving the original notation whenever possible, and also the new concept of *∂-invariant directed quasi-cliques*.

3.2.1 Path Complexes

Definition 3.2.1. Given a finite set of vertices V and an integer $p \geq 0$, an *elementary p-path* is any sequence of $p+1$ vertices $i_k \in V$, $k = 0, ..., p$ (not necessarily distinct), which will be denoted by $e_{i_0 \dots i_p}$. Given a field $\mathbb{K}$, we denote by $\Lambda_p = \Lambda_p(V)$ the $\mathbb{K}$-vector space formed by all formal $\mathbb{K}$-linear combinations of elementary p-paths, i.e. $u = \sum_{i_0,\dots,i_p \in V} u^{i_0 \dots i_p} e_{i_0 \dots i_p} \in \Lambda_p$, where $u^{i_0 \dots i_p} \in \mathbb{K}$. The elements of Λ_p are called *p-paths*. Also, denoting the empty set as an element of Λ_{-1} by e, we have $\Lambda_{-1} \cong \mathbb{K}$, since the elements of Λ_{-1} are multiples of e, and we define $\Lambda_{-2} = \{0\}$.

Definition 3.2.2. Given a digraph $G = (V, E)$, for any $p \geq 0$ we define the *p-th boundary operator* $\partial_p : \Lambda_p(V) \to \Lambda_{p-1}(V)$ by

$$\partial_p(e_{i_0 \dots i_p}) = \sum_{q=0}^{p} (-1)^q \hat{d}_q(e_{i_0 \dots i_p}), \tag{3.28}$$

where $\hat{d}_q(e_{i_0 \dots i_p}) = e_{i_0 \dots \hat{i}_q \dots i_p}$, i.e. $\hat{d}_q$ is the function that excludes the q-th node i_q of the p-path $e_{i_0 \dots i_p}$. In addition, we define $\partial_{-2} := 0$.

The operator (3.28) satisfies the property $\partial_p \circ \partial_{p+1} = 0$, for all $p \geq 0$ (see [130], p. 567, for a proof).

Definition 3.2.3. An elementary p-path $e_{i_0 \dots i_p}$ is said to be *regular* if $i_k \neq i_{k+1}$ for all $k = 0, ..., p-1$, and it is called *non-regular* otherwise. The elements of the subspace $\mathcal{R}_p(V) = \text{span}\{e_{i_0 \dots i_p} : i_k \neq i_{k+1}, \forall k = 0, ..., p-1\} \subseteq \Lambda_p(V)$ are called *regular p-paths*.

Although Definition 2.1.11 states that all vertices of a directed path must be different, here we are adopting an abuse of notation and using the expression *elementary path* as a synonym of *walk*, and the expression *regular elementary path* as a synonym of *trail*.

Definition 3.2.4. A *path complex* over a finite set V is a non-empty set $\mathcal{P} = \mathcal{P}(V)$ of elementary paths on V such that for any $n \geq 0$, if $e_{i_0 \dots i_n} \in \mathcal{P}(V)$, then the truncated paths $e_{i_0 \dots i_{n-1}}$ and $e_{i_1 \dots i_n}$ belong to $\mathcal{P}(V)$ as well.

Definition 3.2.5. Let $G = (V, E)$ be a digraph. A regular elementary p-path $e_{i_0 \dots i_p}$ on V is called *allowed* if the directed edge $(i_k, i_{k+1}) \in E$ for any $k = 0, .., p-1$, and its called *non-allowed* otherwise. We denote by $\mathcal{A}_p(V)$ the $\mathbb{K}$-vector subspace of $\Lambda_p(V)$ spanned by allowed elementary p-paths, i.e.

$$\mathcal{A}_p(V) = \text{span}\{e_{i_0 \dots i_p} : i_0 \dots i_p \text{ is allowed}\}. \tag{3.29}$$

Also, we denote by $\mathcal{P}_p(G)$ the set of all allowed p-paths and by $\mathcal{P}(G) = \bigcup_p \mathcal{P}_p(G)$ the path complex associated with the digraph $G = (V, E)$. In particular, $\mathcal{P}_0(G) = V$ and $\mathcal{P}_1(G) = E$.

Here we emphasize that, from now on, we will consider only path complexes associated with digraphs.

Example 3.2.1. Consider the digraph G shown in Figure 3.11. Its path complex is the following set of elementary p-paths, with $p = 0, 1, 2, 3$:

$$\mathcal{P}(G) = \{e_0, e_1, e_2, e_3, e_{01}, e_{02}, e_{12}, e_{13}, e_{23}, e_{012}, e_{123}, e_{013}, e_{023}, e_{0123}\}.$$

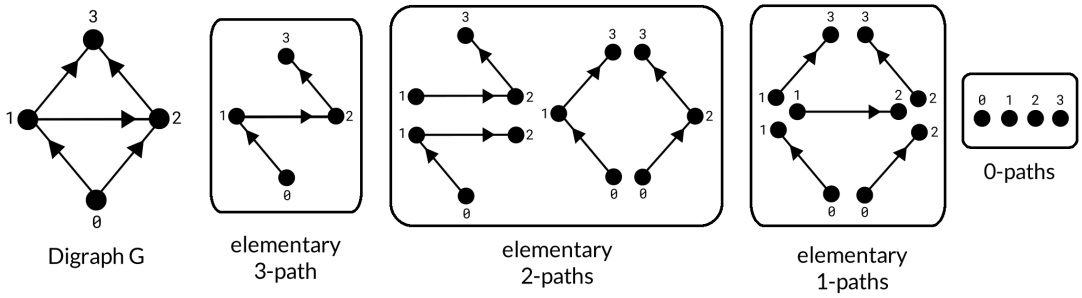


Figure 3.11: A digraph and its path complex.

Note that every p-path of a path complex is allowed. Restricting the operator ∂_p on the subspace $\mathcal{A}_p \subseteq \Lambda_p$, $p \geq 0$, we can have $\partial_p \mathcal{A}_p \not\subset \mathcal{A}_{p-1}$, but we are interested in the case where the inclusion occurs, so we define the following subspace of $\mathcal{A}_p$.

Definition 3.2.6. Given a digraph $G = (V, E)$, consider the following subspace of $\mathcal{A}_p(V)$, $p \geq 0$:

$$\Omega_p = \Omega_p(G) := \{u \in \mathcal{A}_p : \partial_p u \in \mathcal{A}_{p-1}\}. \tag{3.30}$$

The elements of Ω_p are called *∂-invariant p-paths.*

Notice that $\partial_p \Omega_p \subseteq \Omega_{p-1}$, for all $p \geq 0$. In fact, by definition, $\partial_p u \in \mathcal{A}_{p-1}$ for all $u \in \Omega_p$, and since $\partial_{p-1}(\partial_p u) = 0 \in \mathcal{A}_{p-2}$, we have $\partial_p u \in \Omega_{p-1}$.

Example 3.2.2. A triangle is a sequence of three vertices $0, 1, 2$ such that the directed edges $(0, 1)$, $(1, 2)$, and $(0, 2)$ exist (this coincides with the definition of a directed 3-clique, but, as already commented, in this part we'll use the original nomenclature as exposed in [127]). A triangle determines a ∂-invariant 2-path, $e_{012} \in \Omega_2$, because $e_{012} \in \mathcal{A}_2$ and $\partial e_{012} = e_{12} - e_{02} + e_{01} \in \mathcal{A}_1$. Other examples of digraphs that determine ∂-invariant 2-paths are double-edges and squares (see Figure 3.12).

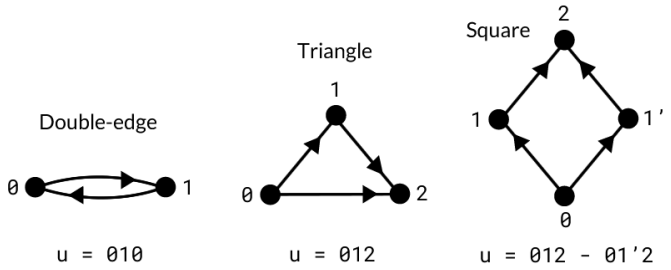


Figure 3.12: Examples of digraphs containing ∂-invariant 2-paths.

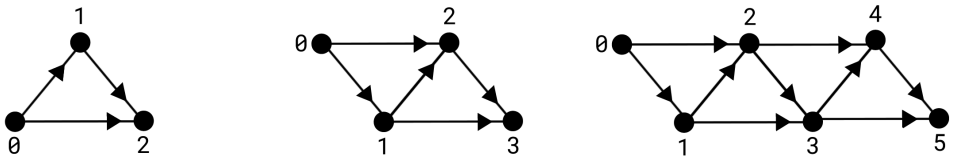
Moreover, Grigor'yan et al. [127] proved that the elements of Ω_2 are linear combinations of triangles, squares, and double-edges.

Snakes and ∂-Invariant Directed Quasi-Cliques

The presence or absence of ∂-invariant paths in a digraph can characterize some of its topological properties, since these paths are related to the *path homology* of the digraph, as we will see in the next section. In view of this, it's worth looking for certain types of subdigraphs that contain ∂-invariant paths, such as *snakes* and *∂-invariant directed quasi-cliques*.

Definition 3.2.7. For a given integer $p \geq 0$, a *p-snake* is a digraph $G = (V, E)$, with $V = \{v_0, ..., v_p\}$, such that its arcs are (v_i, v_{i+1}) for all $i = 0, ..., p-1$, and (v_j, v_{j+2}), for all $j = 0, ..., p-2$.

Example 3.2.3. Figure 3.13 presents examples of p-snakes for $p = 2, 3, 4$.



(a) 2-snake. (b) 3-snake. (c) 4-snake.

Figure 3.13: Examples of p-snakes for $p = 2, 3, 4$.

In the next definition, we present a directed version of the definition of γ-quasi-clique (Definition 2.1.19).

Definition 3.2.8. Given a digraph $G = (V, E)$, a subdigraph $H = (V', E') \subseteq G$, with $|V'| = m$, is called a *directed γ-quasi-clique* (or γ-DQC), for a parameter $0 < \gamma \leq 1$, if $\deg_H^{tot}(v) \geq \gamma(m-1)$, for all $v \in V'$.

Note that the previous definition is equivalent to saying that a digraph is a directed γ-quasi-clique if its underlying undirected graph is a γ-quasi-clique.

Definition 3.2.9. Let $u \in \Omega_p$ be a ∂-invariant p-path. For a parameter $0 < \gamma \leq 1$, the *∂-invariant (u, γ)-DQC* (or just *(u, γ)-DQC*), is the directed γ-quasi-clique with the minimum amount of arcs which contains all the paths necessary to make u ∂-invariant. In particular, if $u = e_{0\ldots p}$ is an elementary ∂-invariant p-path, we denote its (u, γ)-DQC simply by (p, γ)-DQC.

It is important to point out that every directed $(p+1)$-clique determines an elementary p-path (the concept of *simplex-digraph* as proposed by [127] coincides with the definition of directed clique), thus every (p, γ)-DQC is a subdigraph of a directed $(p+1)$-clique. Moreover, if the number of edges in the digraph (p, γ)-DQC is equal to $2p - 1$, then (p, γ)-DQC coincides with the p-snake, and since the total degree of each node of a p-snake is greater than or equal to 2, we can adopt $\gamma = 1/p$ to every (p, γ)-DQC. Table 3.1 summarizes the three types of digraphs discussed here, which contain ∂-invariant paths.

Table 3.1: Digraphs containing ∂-invariant paths.

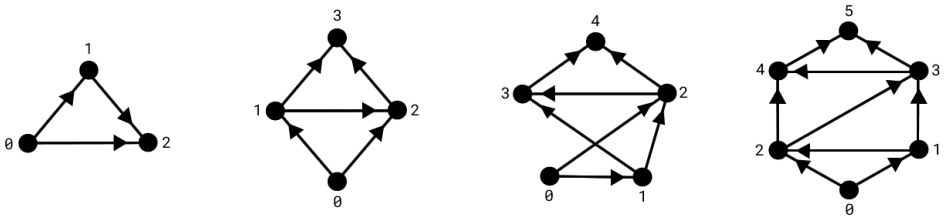
Directed $(p+1)$-Clique	**(u, γ)-DQC**	**p-Snake**
Contains a ∂-invariant elementary p-path	Contains an arbitrary ∂-invariant p-path u	Contains a ∂-invariant elementary p-path

Example 3.2.4. Figures 3.14 and 3.15 present some examples of (u, γ)-DQC associated with some given ∂-invariant paths u. Note that the digraphs (a), (b), and (d) in Figure 3.14 coincides with the snakes (a), (b), and (c) present in Figure 3.13, respectively.

3.2.2 Path Homology

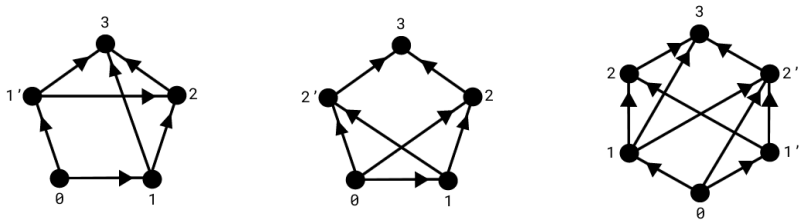
In this part, we extend the concept of simplicial homology introduced for simplicial complexes in Subsection 3.1.4 to path complexes.

We begin by pointing out that the boundary operators (3.28) restricted to the spaces $\Omega_\bullet$ satisfy the same properties as the boundary operators (3.10). Accordingly, by restricting ∂_p to Ω_p, $p \geq 0$, we have $\partial_p \circ \partial_{p+1} = 0$, and $\operatorname{Im} \partial_{p+1} \subseteq \ker \partial_p$, thus we can define a chain complex $(\Omega_\bullet, \partial_\bullet)$:



(a) (u_1, γ)-DQC (b) (u_2, γ)-DQC (c) (u_3, γ)-DQC (d) (u_4, γ)-DQC

Figure 3.14: The (u_i, γ)-DQCs associated with elementary ∂-invariant p-paths, with $\gamma = 1/p$. (a) $u_1 = e_{012}$. (b)$u_2 = e_{0123}$. (c) $u_3 = e_{01234}$ (d) $u_4 = e_{012345}$.



(a) (u_1, γ)-DQC (b) (u_2, γ)-DQC (c) (u_3, γ)-DQC

Figure 3.15: Examples of (u, γ)-DQCs associated with ∂-invariant 3-paths. (a) $u = e_{0123} - e_{01'23}$. (b) $u_2 = e_{0123} - e_{012'3}$. (c) $u_3 = e_{0123} - e_{012'3} + e_{01'2'3}$.

$$\ldots \xrightarrow{\partial_{p+1}} \Omega_p \xrightarrow{\partial_p} \Omega_{p-1} \xrightarrow{\partial_{p-1}} \ldots \xrightarrow{\partial_2} \Omega_1 \xrightarrow{\partial_1} \Omega_0 \to \{0\}.$$

Definition 3.2.10. Given a digraph G and boundary operators $\partial_p : \Omega_p(G) \to \Omega_{p-1}(G)$, $p \geq 0$, the *p-th path homology* of G is defined as the quotient space

$$H_p(G) = \ker \partial_p / \operatorname{Im} \partial_{p+1}. \tag{3.31}$$

Here we adopt the same nomenclature of Subsection 3.1.4 by saying that the p-th Betti number corresponds to the dimension of the p-th path homology, and we denote $\beta_p(G) = \dim H_p(G)$. Furthermore, the Euler characteristic of G is defined as the alternating sum of its Betti numbers:

$$\chi(G) = \sum_{p=0}^{\infty} (-1)^p \beta_p(G) = \sum_{p=0}^{\infty} (-1)^p \dim H_p(G). \tag{3.32}$$

Observation 3.2.1. A version of persistent homology for path complexes, the *persistent path homology* (PPH), was introduced in [56]. Lin et al. [171] generalized the path homology of digraphs for weighted digraphs by defining weighted path homologies through weighted boundary operators defined in an analogous way to the formula (3.13), and, in addition, they proved that the corresponding persistent weighted path

homology, in the case where the coefficients of the homologies belong to a field, is independent on the weights, but it will depend on the weights if the coefficients belong to a general ring.

3.2.3 Combinatorial Hodge Laplacian of Path Complexes

In Subsection 3.1.6 we have introduced the Hodge Laplacian operators for simplicial complexes. Nonetheless, one can build Hodge Laplacian operators for path complexes associated with digraphs in an analogous way [125].

Consider an inner product in the spaces Λ_p, $p \geq 0$. Since Ω_p is a subspace of Λ_p, it inherits the inner product. Let $\partial_p : \Omega_p \to \Omega_{p-1}$ be the p-th boundary operator and $\partial_p^* : \Omega_{p-1} \to \Omega_p$ be its adjoint operator. The *Hodge p-Laplacian operator*, $\mathcal{L}_p : \Omega_p \to \Omega_p$, is defined by

$$\mathcal{L}_p = \partial_{p+1} \circ \partial_{p+1}^* + \partial_p^* \circ \partial_p. \tag{3.33}$$

The matrix representation of the Hodge p-Laplacian operator is given by

$$[\mathcal{L}_p] = B_{p+1} B_{p+1}^T + B_p^T B_p, \tag{3.34}$$

where $B_p = [\partial_p]$ is the matrix representation of the operator ∂_p.

Observation 3.2.2. We point out that all additional observations from Subsection 3.1.6 made for the Hodge Laplacians associated with simplicial complexes are equally valid for the case when they are obtained from path complexes.

Example 3.2.5. Consider the digraph G as depicted in Figure 3.10. The higher-order incidence matrices B_p are:

$$B_1 = \begin{array}{c} \\ e_0 \\ e_1 \\ e_2 \\ e_3 \end{array} \begin{array}{c} \begin{array}{ccccc} e_{01} & e_{02} & e_{12} & e_{13} & e_{23} \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix} \end{array}, \quad B_2 = \begin{array}{c} \\ e_{01} \\ e_{02} \\ e_{12} \\ e_{13} \\ e_{23} \end{array} \begin{array}{c} \begin{array}{ccc} e_{012} & e_{123} & (e_{013} - e_{023}) \end{array} \\ \begin{bmatrix} 1 & 0 & 1 \\ -1 & 0 & -1 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 1 & -1 \end{bmatrix} \end{array},$$

$$B_3 = \begin{array}{c} \\ e_{012} \\ e_{123} \\ (e_{013} - e_{023}) \end{array} \begin{array}{c} \begin{array}{c} e_{0123} \end{array} \\ \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \end{array}.$$

Note that $B_p = 0$ for all $p \geq 4$. Thus, by Equation (3.34), the matrix representations

of the Hodge p-Laplacians of G are:

$$[\mathcal{L}_0] = B_1 B_1^T = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}, \quad [\mathcal{L}_1] = B_2 B_2^T + B_1^T B_1 = \begin{bmatrix} 4 & -1 & 0 & 0 & -1 \\ -1 & 4 & 0 & -1 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & -1 & 0 & 4 & -1 \\ -1 & 0 & 0 & -1 & 4 \end{bmatrix},$$

$$[\mathcal{L}_2] = B_3 B_3^T + B_2^T B_2 = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 4 & -1 \\ 1 & -1 & 5 \end{bmatrix}, \quad [\mathcal{L}_3] = B_3^T B_3 = \begin{bmatrix} 3 \end{bmatrix},$$

and $[\mathcal{L}_p] = 0$ for all $p \geq 4$.

Comparing Example 3.1.11 with Example 3.2.5, we can verify that the Hodge Laplacians obtained from the directed flag complex of a given digraph may differ from the Hodge Laplacians obtained from the path complex of the same digraph.

3.3 Directed Q-Analysis and Directed Higher-Order Adjacencies

In this section, we present briefly the main concepts of the classical Q-Analysis and, subsequently, we present a directed version of Q-Analysis by introducing new concepts to deal with directed higher-order connectivity between directed simplices.

3.3.1 A Brief Introduction to Q-Analysis

The set of ideas that later came to be known as Q-Analysis was first introduced by the physicist R. H. Atkin [10, 11, 12]. Atkin's initial proposal was to develop mathematical tools to analyze structures associated with relations, specially relations within social systems [13]. His idea was to model social networks via simplicial complexes and study the connections among their simplices, highlighting, therefore, the importance of the concept of topological connectivity. Sometimes Q-Analysis is referred to as a "language of structure." Since Atkin's seminal work, several extensions and applications of Q-Analysis have been proposed [27, 63, 163], and many other new concepts were introduced [150].

One of the most important concepts coming from Q-Analysis is the concept of *q-connectivity* in a simplicial complex, which is, essentially, a notion of *higher-order connectivity*, i.e. connectivity in different levels of organization (or dimensional levels) of the complex. It is important to highlight that classical Q-Analysis was developed to deal

solely with the higher-order connectivity without considering any kind of directionality between the connections.

This section is based mainly on the references [13, 82, 150]. Also, throughout this part, $\mathcal{X}$ denotes an arbitrary simplicial complex.

Definition 3.3.1. Given two simplices $\sigma^{(n)}, \tau^{(m)} \in \mathcal{X}$, they are called *q-near*, with $0 \leq q \leq \min(n, m)$, if they share a q-face, and in this case we denote $\sigma^{(n)} \sim_q \tau^{(m)}$. In particular, we say that an n-simplex is n-near to itself.

Definition 3.3.2. Given two simplices $\sigma^{(n)}, \tau^{(m)} \in \mathcal{X}$, they are called *q-connected* if there exists a finite number of simplices $\alpha_i^{(n_i)} \in \mathcal{X}$, let's put $i = 1, ..., l$, with $0 \leq q \leq \min(n, m, n_1, ..., n_l)$, such that

$$\sigma^{(n)} \sim_{q_0} \alpha_1^{(n_1)} \sim_{q_1} ... \sim_{q_{l-1}} \alpha_l^{(n_l)} \sim_{q_l} \tau^{(m)}, \tag{3.35}$$

where $q \leq q_j$, for all $j = 0, ..., l$, and in this case we denote[3] $\sigma^{(n)} \boldsymbol{\sim_q} \tau^{(m)}$. We call this sequence a *chain of q-connection*. Also, we say that $\sigma^{(n)}$ and $\tau^{(m)}$ are q-connected by a *chain* of length l. In particular, an n-simplex is said to be n-connected to itself by a chain of length 0.

Example 3.3.1. Consider the simplicial complex shown in Figure 3.16 (which corresponds to the flag complex of the underlying graph). The simplices σ and τ are 0-connected by a chain of length 4: $\sigma \sim_1 \alpha_1 \sim_1 \alpha_2 \sim_1 \alpha_3 \sim_1 \alpha_4 \sim_0 \tau$. At the same time, they are 1-connected by a chain of length 5: $\sigma \sim_1 \alpha_1 \sim_1 \alpha_5 \sim_1 \alpha_6 \sim_1 \alpha_7 \sim_1 \alpha_8 \sim_1 \tau$.

Notice that q-connectivity does not imply q-nearness but the converse is true: two q-near simplices are q-connected (by a chain of length 0) but two q-connected simplices may not be q-near. That is, q-nearness is a particular case of q-connectivity. Also, by definition, if two simplices are q-connected, then they are q'-connected for all $q' < q$.

Definition 3.3.3. Given a simplicial complex $\mathcal{X}$, we denote by $\mathcal{X}_q$ the set of simplices in $\mathcal{X}$ with dimension $\geq q$, i.e.

$$\mathcal{X}_q = \{\sigma^{(n)} \in \mathcal{X} : q \leq n\}. \tag{3.36}$$

A notable property of q-connectivity is that it is an *equivalence relation* on $\mathcal{X}_q$.

Proposition 3.3.1. *Given a simplicial complex $\mathcal{X}$, the relation "is q-connected to" ($\boldsymbol{\sim_q}$) is an equivalence relation on the set $\mathcal{X}_q$.*

[3] As a convention, here we use the symbol $\sim_q$ to denote q-nearness and its bold version $\boldsymbol{\sim_q}$ to denote q-connectivity.

The proof of the previous proposition is analogous to the proof of the Proposition 2.1.1. Moreover, we define the *q-connected components* of $\mathcal{X}$ as the elements of the quotient set $\mathcal{X}_q/\sim_{\boldsymbol{q}}$, i.e. the equivalence classes (or *q-connectivity classes*).

As a matter of fact, to perform a *Q-analysis* on a simplicial complex means to compute its q-connected components for $0 \leq q \leq \dim \mathcal{X}$, and then summarize the number of these components existing at each level q into a *structure vector*.

Definition 3.3.4. Let f_q be the number of q-connected components of $\mathcal{X}$ for $0 \leq q \leq \dim \mathcal{X}$. The *structure vector* (or *first structure vector*) of $\mathcal{X}$ is the tuple

$$\mathbf{f}(\mathcal{X}) = (f_0, ..., f_{\dim \mathcal{X}}). \tag{3.37}$$

We point out that the nomenclature "first structure vector" comes from the fact that there are other structure vectors defined for simplicial complexes, such as the "second and third structure vectors" [7], but they'll be discussed in the next section. Also, since f_0 denotes the number of 0-connected components of $\mathcal{X}$, it is equal to the 0-th Betti number, i.e. $f_0 = \beta_0$.

Example 3.3.2. Let $\mathcal{X}$ be the simplicial complex shown in Figure 3.16. The corresponding structure vector is $\mathbf{f}(\mathcal{X}) = (2, 2, 11, 1)$.

One way to summarize the connectivity between the simplices of a simplicial complex at a given level q is through the idea of *q-graph*.

Definition 3.3.5. The *q-graph* of a simplicial complex $\mathcal{X}$, $0 \leq q \leq \dim \mathcal{X}$, is a graph in which every vertex corresponds to a simplex in $\mathcal{X}_q$ and there exists an edge between two vertices if and only if the simplices corresponding to these vertices are q-near.

Now we present some concepts that are not only related to q-connectivity but also related with the complex itself.

Definition 3.3.6. The *q-star* of a simplex $\sigma^{(n)} \in \mathcal{X}$, $0 \leq q \leq n$, is defined as the set of all simplices that are q-near with $\sigma^{(n)}$, i.e.

$$\mathrm{st}_q(\sigma^{(n)}) = \{\tau^{(m)} \in \mathcal{X} : \sigma^{(n)} \sim_q \tau^{(m)}\}. \tag{3.38}$$

When $q = n$, the n-star is the set of all the simplices having $\sigma^{(n)}$ as a face, and in this case we use the notation $\mathrm{st}^*(\sigma^{(n)}) = \mathrm{st}_n(\sigma^{(n)})$. Also, if $\mathcal{F}$ is a simplicial family obtained from $\mathcal{X}$, the set of all simplices that have at least one face in $\mathcal{F}$ is given by

$$\mathrm{st}^*(\mathcal{F}) = \bigcup_{\sigma \in \mathcal{F}} \mathrm{st}^*(\sigma). \tag{3.39}$$

Definition 3.3.7. Given a simplicial family $\mathcal{F}$, the *hub* of $\mathcal{F}$ is the set formed by all the simplices that are common faces of the elements of $\mathcal{F}$, i.e.

$$\text{hub}(\mathcal{F}) = \bigcap_{\sigma \in \mathcal{F}} \sigma. \tag{3.40}$$

A generalization for simplicial complexes of the idea of neighborhood of a node in a graph is the concept of *link* of a simplex.

Definition 3.3.8. The *link* of a simplex $\sigma \in \mathcal{X}$ is defined as the set of all simplices τ such that σ and τ are disjoint faces of the simplex $\sigma \cup \tau$, i.e.

$$\text{lk}(\sigma) = \{\tau \in \mathcal{X} : \sigma \cap \tau = \emptyset \text{ and } \sigma \cup \tau \in \mathcal{X}\}. \tag{3.41}$$

Example 3.3.3. Considering the simplicial complex presented in Figure 3.16, we have the following examples: the 1-star of the simplex α_1 is $\text{st}_1(\alpha_1) = \{\sigma, \alpha_2, \alpha_5\}$; the hub of the simplicial family $\mathcal{F} = \{\alpha_1, \alpha_2, \alpha_3, \alpha_5, \alpha_6\}$ is $\text{hub}(\mathcal{F}) = \{3\}$; the link of the 1-simplex $\{12, 13\}$ is the 1-simplex $\{10, 11\}$.

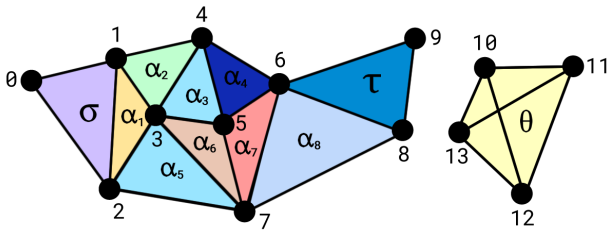


Figure 3.16: A simplicial complex.

3.3.2 Directed Q-Analysis and Directed Higher-Order Adjacencies

As we have seen in the previous section, Atkin's Q-Analysis defines q-connectivity between two simplices based solely on the face shared by them and does not say anything about the directionality of this connection. Recently, however, H. Riihimaki [220] introduced a formalism to treat the q-connectivity between directed simplices taking directionality into account, thus creating a directed analogue of Q-Analysis for directed flag complexes.

Moreover, unlike adjacencies between vertices in a graph, when we are dealing with simplices, we can distinguish between two types of adjacencies: *lower* and *upper* adjacencies. Lower adjacencies compare how two simplices share their faces, and upper adjacencies tell us how they are nested in other higher-dimensional simplices [92, 122, 233].

Accordingly, we need appropriate definitions of lower and upper adjacencies for directed simplices, so that the directionality of the connection between them is taken into account, and this is the aim of this section.

In what follows, we present the concept of $(q, \hat{d}_i, \hat{d}_j)$-*connectivity* as introduced in [220], and then we extend the concepts of lower, upper, and general adjacencies as presented in [233] to directed simplices. Throughout this part, dFl(G) will denote the directed flag complex of a given simple digraph G *without double-edges*.

$(q, \hat{d}_i, \hat{d}_j)$-Connectivity Between Directed Simplices

The direction of the connection between two directed simplices is based on a slightly modified face map as defined below.

Definition 3.3.9. Given a directed simplex $\sigma^{(n)} = [v_0, ..., v_n] \in \text{dFl}(G)$, the i-th face map $\hat{d}_i$ is defined as

$$\hat{d}_i(\sigma^{(n)}) = \begin{cases} [v_0, ..., \hat{v}_i, ..., v_n], & \text{if } i < n, \\ [v_0, ..., v_{n-1}, \hat{v}_n], & \text{if } i \geq n. \end{cases} \tag{3.42}$$

Definition 3.3.10. Let $\sigma^{(n)}, \tau^{(m)} \in \text{dFl}(G)$ be two directed simplices. For $0 \leq q \leq \min(n, m)$, we say that $\sigma^{(n)}$ is $(q, \hat{d}_i, \hat{d}_j)$-*near* to $\tau^{(m)}$ if either of the following conditions is satisfied:

1. $\sigma^{(n)} \subseteq \tau^{(m)}$;
2. $\hat{d}_i(\sigma^{(n)}) \supseteq \alpha^{(q)} \subseteq \hat{d}_j(\tau^{(m)})$, for some $\alpha^{(q)} \in \text{dFl}(G)$ (i.e. if they share a q-face).

The previous definition established a concept of directionality in the connection between two directed simplices based on *how* their faces are shared. However, a more precise and descriptive definition/notation is needed to make analogies, for the higher-order setting, with some definitions made for directed networks. Accordingly, we propose the following definitions/notations.

Definition 3.3.11. For two directed simplices $\sigma^{(n)}, \tau^{(m)} \in \text{dFl}(G)$ and for $0 \leq q \leq \min(n, m)$, we have the following definitions:

1. $\sigma^{(n)}$ is said to be *in-q-near* (or $(-)$-*q-near*) to $\tau^{(m)}$ if they are $(q, d_i(\sigma^{(n)}), d_j(\tau^{(m)}))$-near with $i \geq j$, for at least one pair i, j. In this case, we denote $\sigma^{(n)} \sim_q^- \tau^{(m)}$.
2. $\sigma^{(n)}$ is said to be *out-q-near* (or $(+)$-*q-near*) to $\tau^{(m)}$ if they are $(q, d_i(\sigma^{(n)}), d_j(\tau^{(m)}))$-near with $i \leq j$, for at least one pair i, j. In this case, we denote $\sigma^{(n)} \sim_q^+ \tau^{(m)}$.

3. $\sigma^{(n)}$ is said to be *bidirectionally q-near* (or $(\pm)$-*q-near*) to $\tau^{(m)}$ if $\sigma^{(n)} \sim_q^- \tau^{(m)}$ and $\sigma^{(n)} \sim_q^+ \tau^{(m)}$. In this case, we denote $\sigma^{(n)} \sim_q^\pm \tau^{(m)} = \tau^{(m)} \sim_q^\pm \sigma^{(n)}$.

It's clear from the definitions that $\sigma^{(n)} \sim_q^- \tau^{(m)} = \tau^{(m)} \sim_q^+ \sigma^{(n)}$ and $\sigma^{(n)} \sim_q^+ \tau^{(m)} = \tau^{(m)} \sim_q^- \sigma^{(n)}$. Also, if $\sigma^{(n)} \subseteq \tau^{(m)}$, by definition, $\sigma^{(n)} \sim_q^\pm \tau^{(m)}$.

Notation: We use the notation $\sigma^{(n)} \sim_q^\bullet \tau^{(m)}$, where $\bullet \in \{-, +, \pm\}$, to represent the respective connectivity relation by replacing $\bullet$ with its respective symbol.

Now that we have introduced the concept of $(q, \hat{d}_i, \hat{d}_j)$-nearness and established the notations, let's introduce the concept of $(q, \hat{d}_i, \hat{d}_j)$-*connectivity* using the new notations.

Definition 3.3.12. Given two directed simplices $\sigma^{(n)}, \tau^{(m)} \in \mathrm{dFl}(G)$, we say that $\sigma^{(n)}$ is $(\bullet)$-q-*connected* to $\tau^{(m)}$, where $\bullet \in \{-, +\}$, if there exists a finite number of simplices $\alpha_i^{(n_i)} \in \mathrm{dFl}(G)$, let's put $i = 1, ..., l$, with $0 \leq q \leq \min(n, m, n_1, ..., n_l)$, such that

$$\sigma^{(n)} \sim_{q_0}^\bullet \alpha_1^{(n_1)} \sim_{q_1}^\bullet ... \sim_{q_{l-1}}^\bullet \alpha_l^{(n_l)} \sim_{q_l}^\bullet \tau^{(m)}, \tag{3.43}$$

where $q \leq q_j$, for all $j = 0, ..., l$, and in this case we denote $\sigma^{(n)} \boldsymbol{\sim}_{\boldsymbol{q}}^\bullet \tau^{(m)}$. We say that $\sigma^{(n)}$ is $(\bullet)$-q-connected to $\tau^{(m)}$ by a *directed* $(\bullet)$-q-*chain* of length l. We denote $\sigma^{(n)} \boldsymbol{\sim}_{\boldsymbol{q}}^\pm \tau^{(m)}$ if $\sigma^{(n)} \boldsymbol{\sim}_{\boldsymbol{q}}^+ \tau^{(m)}$ and $\sigma^{(n)} \boldsymbol{\sim}_{\boldsymbol{q}}^- \tau^{(m)}$ and we say they are $(\pm)$-q-*connected*. In particular, a directed n-simplex is said to be $(\pm)$-q-connected to itself by a directed $(\pm)$-q-chain of length 0. Finally, for $\bullet \in \{-, +\}$, if $\sigma^{(n)}$ is not $(\bullet)$-q-connected to $\tau^{(m)}$ for all $\bullet$, we say that they are q-*disconnected*.

Notice that, just as in the case of q-connectivity, if two directed simplices are $(\bullet)$-q-near, $\bullet \in \{-, +\}$, then they are $(\bullet)$-q-connected by a directed $(\bullet)$-chain of length 0. Moreover, if two directed simplices are $(\bullet)$-q-connected, then they are $(\bullet)$-q'-connected for all $q' < q$.

Example 3.3.4. Consider the directed flag complex shown in Figure 3.17. We have the following relations:

1) $\theta \sim_0^\pm \tau$, since $\hat{d}_0(\theta) \supseteq [2] \subseteq \hat{d}_2(\tau)$ and $\hat{d}_2(\theta) \supseteq [2] \subseteq \hat{d}_1(\tau)$.
2) $\sigma \sim_0^\pm \tau$, since $\hat{d}_1(\sigma) \supseteq [2] \subseteq \hat{d}_2(\tau)$ and $\hat{d}_2(\sigma) \supseteq [2] \subseteq \hat{d}_1(\tau)$.
3) $\theta \sim_1^+ \sigma$, since $\hat{d}_0(\theta) \supseteq [2, 6] \subseteq \hat{d}_2(\sigma)$.
4) $\alpha \sim_1^+ \theta$, since $\hat{d}_0(\alpha) \supseteq [1, 6] \subseteq \hat{d}_1(\theta)$.
5) $\alpha \boldsymbol{\sim}_{\boldsymbol{0}}^+ \tau$, since $\alpha \sim_1^+ \theta \sim_0^+ \tau$.

$\hat{d}_i(\alpha)$	$\hat{d}_i(\theta)$	$\hat{d}_i(\sigma)$	$\hat{d}_i(\tau)$
$\hat{d}_0(\alpha) = [1, 6]$	$\hat{d}_0(\theta) = [2, 6]$	$\hat{d}_0(\sigma) = [6, 7]$	$\hat{d}_0(\tau) = [3, 4, 5]$
$\hat{d}_1(\alpha) = [0, 6]$	$\hat{d}_1(\theta) = [1, 6]$	$\hat{d}_1(\sigma) = [2, 7]$	$\hat{d}_1(\tau) = [2, 4, 5]$
$\hat{d}_2(\alpha) = [0, 1]$	$\hat{d}_2(\theta) = [1, 2]$	$\hat{d}_2(\sigma) = [2, 6]$	$\hat{d}_2(\tau) = [2, 3, 5]$
			$\hat{d}_3(\tau) = [2, 3, 4]$

$(q, \hat{d}_i(\theta), \hat{d}_j(\tau))$
$(0, \hat{d}_0, \hat{d}_1)$
$(0, \hat{d}_0, \hat{d}_2)$
$(0, \hat{d}_0, \hat{d}_3)$
$(0, \hat{d}_2, \hat{d}_1)$
$(0, \hat{d}_2, \hat{d}_2)$
$(0, \hat{d}_2, \hat{d}_3)$

$(q, \hat{d}_i(\sigma), \hat{d}_j(\tau))$
$(0, \hat{d}_1, \hat{d}_1)$
$(0, \hat{d}_1, \hat{d}_2)$
$(0, \hat{d}_1, \hat{d}_3)$
$(0, \hat{d}_2, \hat{d}_1)$
$(0, \hat{d}_2, \hat{d}_2)$
$(0, \hat{d}_2, \hat{d}_3)$

$(q, \hat{d}_i(\theta), \hat{d}_j(\sigma))$
$(0, \hat{d}_0, \hat{d}_0)$
$(0, \hat{d}_0, \hat{d}_1)$
$(0, \hat{d}_0, \hat{d}_2)$
$(0, \hat{d}_1, \hat{d}_0)$
$(0, \hat{d}_1, \hat{d}_2)$
$(0, \hat{d}_2, \hat{d}_1)$
$(0, \hat{d}_2, \hat{d}_2)$
$(1, \hat{d}_0, \hat{d}_2)$

$(q, \hat{d}_i(\alpha), \hat{d}_j(\theta))$
$(0, \hat{d}_0, \hat{d}_0)$
$(0, \hat{d}_0, \hat{d}_1)$
$(0, \hat{d}_0, \hat{d}_2)$
$(0, \hat{d}_1, \hat{d}_0)$
$(0, \hat{d}_1, \hat{d}_1)$
$(0, \hat{d}_2, \hat{d}_1)$
$(0, \hat{d}_2, \hat{d}_2)$
$(1, \hat{d}_0, \hat{d}_1)$

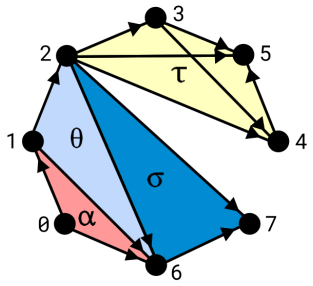


Figure 3.17: A directed flag complex.

Lower, Upper, and General Adjacencies

In what follows, we extend the definitions of lower, upper, and general adjacencies as exposed in [233] to directed simplices.

Definition 3.3.13. For two directed simplices $\sigma^{(n)}, \tau^{(m)} \in \mathrm{dFl}(G)$ and for $0 \leq q \leq \min(n, m)$, we have the following definitions:

1. $\sigma^{(n)}$ is *lower* $(\bullet)$-q-*adjacent* to $\tau^{(m)}$, where $\bullet \in \{-, +, \pm\}$, if and only if $\sigma^{(n)}$ is $(\bullet)$-q-near to $\tau^{(m)}$, i.e.

$$\sigma^{(n)} \sim^{\bullet}_{L_q} \tau^{(m)} \iff \sigma^{(n)} \sim^{\bullet}_{q} \tau^{(m)}.$$

2. $\sigma^{(n)}$ is *strictly lower* $(\bullet)$-q-*adjacent* to $\tau^{(m)}$, where $\bullet \in \{-, +, \pm\}$, if and only if $\sigma^{(n)} \sim^{\bullet}_{L_q} \tau^{(m)}$ and $\sigma^{(n)}$ is not $(\star)$-$(q+1)$-near to $\tau^{(m)}$, for all $\star \in \{-, +, \pm\}$, i.e.

$$\sigma^{(n)} \sim^{\bullet}_{L_{q^*}} \tau^{(m)} \iff \sigma^{(n)} \sim^{\bullet}_{L_q} \tau^{(m)} \text{ and } \sigma^{(n)} \not\sim^{\star}_{L_{q+1}} \tau^{(m)}, \ \forall \star .$$

We point out that lower $(\bullet)$-q-adjacency and $(\bullet)$-q-nearness are exactly the same definitions, thus we propose *lower* $(\bullet)$-q-*nearness* as an alternative nomenclature of $(\bullet)$-q-nearness. Also, note that for vertices they are only lower $(\bullet)$-0-adjacent to themselves.

Besides the lower adjacency, that is, how two simplices share their faces, we could also think about how simplices are nested in other simplices of greater dimensions, and for this, we have the idea of *upper adjacency*. To extend upper adjacency to directed simplices, we need to define precisely how the directionality is taken into account between two directed simplices that are faces of other directed simplices of greater dimensions, thus, based on Definition 3.3.10, we propose the definition of *upper* $(p, \hat{d}_i, \hat{d}_j)$*-nearness* as follows.

Definition 3.3.14. Let $\sigma^{(n)}, \tau^{(m)} \in \mathrm{dFl}(G)$ be two directed simplices. For $n, m < p \leq \dim \mathrm{dFl}(G)$, $\sigma^{(n)}$ is said to be *upper* $(p, \hat{d}_i, \hat{d}_j)$*-near* to $\tau^{(m)}$ if the following condition is true:

$$\sigma^{(n)} = \hat{d}_i(\Theta^{(n+1)}) \subseteq \Theta^{(p)} \supseteq \hat{d}_j(\Theta^{(m+1)}) = \tau^{(m)}, \text{ for some } \Theta^{(n+1)}, \Theta^{(m+1)} \subseteq \Theta^{(p)} \in \mathrm{dFl}(G).$$

Definition 3.3.15. For two directed simplices $\sigma^{(n)}, \tau^{(m)} \in \mathrm{dFl}(G)$ and for $n, m < p \leq \dim \mathrm{dFl}(G)$, we have the following definitions:

1. $\sigma^{(n)}$ is *upper* $(-)$*-p-adjacent* to $\tau^{(m)}$ if and only if $\sigma^{(n)}$ is upper $(p, \hat{d}_i, \hat{d}_j)$-near to $\tau^{(m)}$ and $i \geq j$, for at least one pair i, j, i.e.

$$\sigma^{(n)} \sim^{-}_{U_p} \tau^{(m)} \iff \sigma^{(n)} \text{ is upper } (p, \hat{d}_i, \hat{d}_j)\text{-near to } \tau^{(m)} \text{ with } i \geq j.$$

2. $\sigma^{(n)}$ is *upper* $(+)$*-p-adjacent* to $\tau^{(m)}$ if and only if $\sigma^{(n)}$ is upper $(p, \hat{d}_i, \hat{d}_j)$-near to $\tau^{(m)}$ and $i \leq j$, for at least one pair i, j, i.e.

$$\sigma^{(n)} \sim^{+}_{U_p} \tau^{(m)} \iff \sigma^{(n)} \text{ is upper } (p, \hat{d}_i, \hat{d}_j)\text{-near to } \tau^{(m)} \text{ with } i \leq j.$$

3. $\sigma^{(n)}$ is *upper* $(\pm)$*-p-adjacent* to $\tau^{(m)}$ if and only if $\sigma^{(n)}$ is upper $(-)$-p-adjacent and upper $(+)$-p-adjacent to $\tau^{(m)}$, i.e.

$$\sigma^{(n)} \sim^{\pm}_{U_p} \tau^{(m)} \iff \sigma^{(n)} \sim^{-}_{U_p} \tau^{(m)} \text{ and } \sigma^{(n)} \sim^{+}_{U_p} \tau^{(m)}.$$

4. $\sigma^{(n)}$ is *strictly* $(\bullet)$*-p-upper adjacent* to $\tau^{(m)}$, where $\bullet \in \{-, +, \pm\}$, if and only if $\sigma^{(n)} \sim^{\bullet}_{U_p} \tau^{(m)}$ and $\sigma^{(n)}$ is not upper $(\star)$-$(p+1)$-adjacent to $\tau^{(m)}$, for all $\star \in \{-, +, \pm\}$, i.e.

$$\sigma^{(n)} \sim^{\bullet}_{U_{p^*}} \tau^{(m)} \iff \sigma^{(n)} \sim^{\bullet}_{U_p} \tau^{(m)} \text{ and } \sigma^{(n)} \not\sim^{\star}_{U_{p+1}} \tau^{(m)}, \ \forall \star .$$

Note that if (v, u) is an arc in the graph G, then $v \sim^{+}_{U_1} u$. On the other hand, if (u, v) is an arc in G, then $v \sim^{-}_{U_1} u$.

Now that we have introduced the lower and upper adjacencies, let's define the *general adjacencies*, which take both lower and upper adjacencies into account.

Definition 3.3.16. For two directed simplices $\sigma^{(n)}, \tau^{(m)} \in \mathrm{dFl}(G)$, we have the following definitions:

1. $\sigma^{(n)}$ is $(\bullet)$-q-*adjacent* to $\tau^{(m)}$, where $\bullet \in \{-, +, \pm\}$, if and only if $\sigma^{(n)}$ is strictly lower $(\bullet)$-q-adjacent to $\tau^{(m)}$ and $\sigma^{(n)}$ is not upper $(\star)$-p-adjacent to $\tau^{(m)}$, with $p = n + m - q$, for all $\star \in \{-, +, \pm\}$, i.e.

$$\sigma^{(n)} \sim^{\bullet}_{A_q} \tau^{(m)} \iff \sigma^{(n)} \sim^{\bullet}_{L_{q^*}} \tau^{(m)} \text{ and } \sigma^{(n)} \not\sim^{\star}_{U_p} \tau^{(m)}, \ \forall \star .$$

2. $\sigma^{(n)}$ is *maximal* $(\bullet)$-q-*adjacent* to $\tau^{(m)}$, where $\bullet \in \{-, +, \pm\}$, if and only if $\sigma^{(n)}$ is $(\bullet)$-q-adjacent to $\tau^{(m)}$ and $\sigma^{(n)}$ is not a face of any other directed simplex which is $(\star)$-q-adjacent to $\tau^{(m)}$, for all $\star \in \{-, +, \pm\}$, i.e.

$$\sigma^{(n)} \sim^{\bullet}_{A_{q^*}} \tau^{(m)} \iff \sigma^{(n)} \sim^{\bullet}_{A_q} \tau^{(m)} \text{ and } \sigma^{(n)} \not\subset \sigma^{(r)}, \ \forall \sigma^{(r)} : \sigma^{(r)} \sim^{\star}_{A_q} \tau^{(m)}, \ \forall \star .$$

Observation 3.3.1. The quantity $p = m + n - q$ of the previous definition comes from the fact that if $\sigma^{(n)} \sim^{\bullet}_{L_{q^*}} \tau^{(m)}$, then they share $(q + 1)$ vertices and thus the smallest directed simplex which might contain $\sigma^{(n)}$ and $\tau^{(m)}$ as faces must have $(n + 1) + (m + 1) - (q + 1)$ vertices, i.e. must have a dimension equal to $n + m - q$.

As will be proved in the next proposition, if a maximal directed simplex is strictly lower $(\bullet)$-q-adjacent to another maximal directed simplex, then this adjacency is actually a maximal $(\bullet)$-q-adjacency. Let's first introduce some useful definitions.

Definition 3.3.17. The set of the maximal directed simplices of $\mathrm{dFl}(G)$ is defined by

$$\mathrm{dFl}^*(G) = \{\sigma^{(n)} \in \mathrm{dFl}(G) : \sigma^{(n)} \text{ is maximal } \}. \tag{3.44}$$

The set of the maximal directed simplices of $\mathrm{dFl}(G)$ whose dimensions are greater than or equal to $0 \leq q \leq \dim \mathrm{dFl}(G)$ is defined by

$$\mathrm{dFl}^*_q(G) = \{\sigma^{(n)} \in \mathrm{dFl}^*(G) : q \leq n\}. \tag{3.45}$$

Proposition 3.3.2. *For two maximal directed simplices $\sigma^{(n)}, \tau^{(m)} \in \mathrm{dFl}^*(G)$ and for $\bullet \in \{-, +, \pm\}$, we have the following equivalence:*

$$\sigma^{(n)} \sim^{\bullet}_{L_{q^*}} \tau^{(m)} \iff \sigma^{(n)} \sim^{\bullet}_{A_{q^*}} \tau^{(m)}.$$

Proof. Suppose $\sigma^{(n)} \sim^{\bullet}_{L_{q^*}} \tau^{(m)}$. Since $\sigma^{(n)}, \tau^{(m)} \in \mathrm{dFl}^*(G)$, both simplices are not faces of any other simplices in $\mathrm{dFl}(G)$, then $\sigma^{(n)} \sim^{\bullet}_{A_{q^*}} \tau^{(m)}$. On the other hand, if $\sigma^{(n)} \sim^{\bullet}_{A_{q^*}} \tau^{(m)}$, by definition, $\sigma^{(n)} \sim^{\bullet}_{L_{q^*}} \tau^{(m)}$. □

Directed Simplicial q-Walks and q-Distances

In Definition 3.3.12, we implicitly defined $(\bullet)$-q-connectivity in terms of *lower* $(\bullet)$-q-adjacency. In what follows, we introduce the concept of *maximal* $(\bullet)$-*q-connectivity*, which is the basis for defining the idea of *directed simplicial q-walk*.

Definition 3.3.18. Given two directed simplices $\sigma^{(n)}, \tau^{(m)} \in \mathrm{dFl}(G)$, we say that $\sigma^{(n)}$ is *maximal* $(\bullet)$-*q-connected* to $\tau^{(m)}$, where $\bullet \in \{-, +\}$, if there exists a finite number of simplices $\alpha_i^{(n_i)} \in \mathrm{dFl}(G)$, let's put $i = 1, ..., l$, with $0 \leq q \leq \min(n, m, n_1, ..., n_l)$, such that

$$\sigma^{(n)} \sim^{\bullet}_{A_{q_0^*}} \alpha_1^{(n_1)} \sim^{\bullet}_{A_{q_1^*}} ... \sim^{\bullet}_{A_{q_{l-1}^*}} \alpha_l^{(n_l)} \sim^{\bullet}_{A_{q_l^*}} \tau^{(m)}, \tag{3.46}$$

where $q \leq q_j$, for all $j = 0, ..., l$, and in this case we denote $\sigma^{(n)} \sim^{\bullet}_{\boldsymbol{A_{q^*}}} \tau^{(m)}$. We say that there is a *directed simplicial q-walk* of length l from $\sigma^{(n)}$ to $\tau^{(m)}$ if $\sigma^{(n)} \sim^{+}_{\boldsymbol{A_{q^*}}} \tau^{(m)}$ or from $\tau^{(m)}$ to $\sigma^{(n)}$ if $\sigma^{(n)} \sim^{-}_{\boldsymbol{A_{q^*}}} \tau^{(m)}$. We denote $\sigma^{(n)} \sim^{\pm}_{\boldsymbol{A_{q^*}}} \tau^{(m)}$ if $\sigma^{(n)} \sim^{+}_{\boldsymbol{A_{q^*}}} \tau^{(m)}$ and $\sigma^{(n)} \sim^{-}_{\boldsymbol{A_{q^*}}} \tau^{(m)}$. Moreover, for $\bullet \in \{-, +, \pm\}$, if $\sigma^{(n)}$ is not maximal $(\bullet)$-q-connected to $\tau^{(m)}$ for all $\bullet$, we say that they are *maximal q-disconnected*.

Notice that if $\sigma \sim^{+}_{\boldsymbol{A_{q^*}}} \tau$ with $\sigma = \tau$, then we have a *directed simplicial q-cycle*. Also, q-disconnectedness implies maximal q-disconnectedness, but the converse might not be true.

Definition 3.3.19. Given two directed simplices $\sigma, \tau \in \mathrm{dFl}(G)$, the *directed simplicial q-distance* from σ to τ, denoted by $\vec{d}_q(\sigma, \tau)$, is equal to the length of the shortest directed simplicial q-walk from σ to τ. If σ and τ are q-disconnected, then we define $\vec{d}_q(\sigma, \tau) = \infty$.

As a matter of fact, $\vec{d}_q$ is a quasi-distance (see Definition 2.1.30), since the property of symmetry is not necessarily satisfied.

Weakly and Strongly q-Connected Components and Structure Vectors

As already commented in Subsection 3.3.1, performing a Q-analysis on a simplicial complex consists of calculating its q-connected components and then constructing its structure vector. However, similarly when we are dealing with digraphs, for directed flag complexes we have two types of "q-connected components," namely: *weakly q-connected components* and *strongly q-connected components*. Before defining these kinds of "q-connected components," let's define one more subset of $\mathrm{dFl}(G)$.

Definition 3.3.20. The set of the directed simplices of $\mathrm{dFl}(G)$ whose dimension is greater or equal to $0 \leq q \leq \dim \mathrm{dFl}(G)$ is defined by

$$\mathrm{dFl}_q(G) = \{\sigma^{(n)} \in \mathrm{dFl}(G) : q \leq n\}. \tag{3.47}$$

In analogy with what was exposed in [233], a directed simplex is not maximal $(\bullet)$-q-connected to itself, $\bullet \in \{-, +, \pm\}$, thus this relation is not reflexive. Accordingly, in order to obtain an equivalence relation, we introduce the following relation:

$$(\sigma^{(n)}, \tau^{(m)}) \in S_q^s \iff \begin{cases} \sigma^{(n)} \sim_{\boldsymbol{A}_{q^*}}^{\pm} \tau^{(m)}, \\ \text{or } \sigma^{(n)} = \tau^{(m)}. \end{cases} \tag{3.48}$$

One can verify that (3.48) is indeed an equivalence relation by following the steps of the proof of Proposition 2.1.2.

Definition 3.3.21. The *maximal strongly q-connected components* of $\mathrm{dFl}(G)$ are the equivalence classes of the quotient set $\mathcal{K}_q^s = \mathrm{dFl}_q(G)/S_q^s$, which are the equivalence classes of maximal $(+)$-q-connected directed simplices.

Furthermore, if we disregard the directionality of the connections between the directed simplices in the relation (3.48), we obtain exactly the maximal q-connectivity as defined in [233] for undirected simplices; thus, we define the following equivalence relation:

$$(\sigma^{(n)}, \tau^{(m)}) \in S_q^w \iff \begin{cases} \sigma^{(n)} \sim_{\boldsymbol{A}_{q^*}} \tau^{(m)}, \\ \text{or } \sigma^{(n)} = \tau^{(m)}, \end{cases} \tag{3.49}$$

where $\sigma^{(n)} \sim_{\boldsymbol{A}_{q^*}} \tau^{(m)}$ denotes the maximal q-connectivity between $\sigma^{(n)}$ and $\tau^{(m)}$. One can verify that (3.49) is indeed an equivalence relation by following the steps of the proof of Proposition 2.1.1.

Definition 3.3.22. The *maximal weakly q-connected components* of $\mathrm{dFl}(G)$ are the equivalence classes of the quotient set $\mathcal{K}_q^w = \mathrm{dFl}_q(G)/S_q^w$, which are the equivalence classes of maximal q-connected directed simplices.

In Subsection 3.3.1 we have defined the (first) structure vector (Definition 3.3.4) based on the number of q-connected components of a simplicial complex $\mathcal{X}$, nonetheless, Andjelkovic et al. [7] described two additional different structure vectors, the *second* and the *third* structure vectors, which are based, respectively, on the number of simplices in the set $\mathcal{X}_q$ and the "degree of connectedness" among the simplices at level q. In what follows, we extend these vectors for the directed case.

Definition 3.3.23. Let $\mathrm{dFl}(G)$ be a directed flag complex with $N = \dim \mathrm{dFl}(G)$. We define the following structure vectors associated with $\mathrm{dFl}(G)$:

1. The *first weak/strong structure vector* is $F^1 = (F_0^1, ..., F_N^1)$, where F_q^1 is the number of maximal weakly/strongly q-connected components.

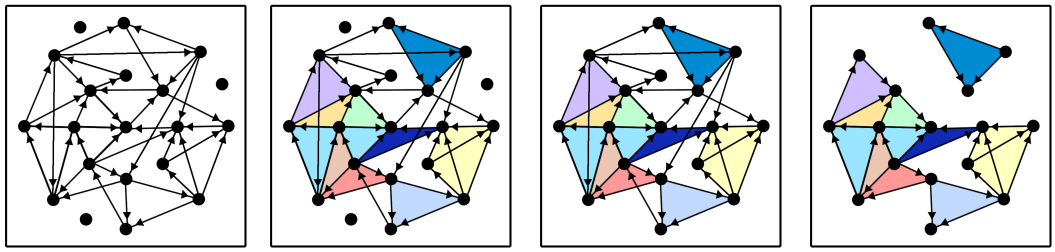
2. The *second structure vector* is $F^2 = (F_0^2, ..., F_N^2)$, where F_q^2 is the number of directed simplices in the set $\mathrm{dFl}_q(G)$.

3. The *third weak/strong structure vector* is $F^3 = (F_0^3, ..., F_N^3)$, where $F_q^3 = 1 - F_q^1/F_q^2$, with F_q^1 being the q-th element of the first weak or strong structure vector. The quantity F_q^3 can be interpreted as the "degree of directed connectedness" among the directed simplices at level q.

Notice that we have two different "first" structure vectors, one for the maximal weakly q-connected components and one for the maximal strongly q-connected components, and the same occurs for the "third" structure vector, but we have only one "second" structure vector.

Maximal q-Digraphs

Previously, we have defined $\mathrm{dFl}_q^*(G)$ as the set of all maximal directed simplices with dimensions greater than or equal to q, so when we advance in the level q, that is, *change the level of organization* of the complex (see Figure 3.18), we obtain a new perspective on its higher-order topology, and, consequently, we may gain insights about the higher-order topology (or the clique organization) of the underlying network.



(a) Digraph G. (b) $\mathrm{dFl}_0^*(G)$. (c) $\mathrm{dFl}_1^*(G)$. (d) $\mathrm{dFl}_2^*(G)$.

Figure 3.18: Graphical representation of the maximal directed simplices for each level $q = 0, 1, 2$.

In view of this, we can go further and consider the directed higher-order connectivity among the simplices in each level q, and then define a "higher-order digraph" for each of these levels. A definition of "higher-order digraph" was already introduced in [49] with the concept of *q-digraph*, however, since we want to deal solely with maximal directed simplices, here we propose a slightly different definition: the *maximal q-digraph*.

Definition 3.3.24. The *maximal q-digraph* of $\mathrm{dFl}(G)$, denoted by $\mathcal{G}_q$, is the digraph whose vertices are the simplices of $\mathrm{dFl}_q^*(G)$ and for each pair $\sigma, \tau \in \mathrm{dFl}_q^*(G)$ there is a directed edge from σ to τ if $\sigma \sim^+_{A_{q^*}} \tau$, with $0 \leq q \leq \dim \mathrm{dFl}(G)$.

In addition, when $\sigma \sim^{+}_{A_{q^*}} \tau$, for $\sigma, \tau \in \mathrm{dFl}^*_q(G)$, we say that there is a *q-arc* from σ to τ and in this case we denote (σ, τ). Also, we may use the notation $\mathcal{G}_q = (\mathcal{V}_q, \mathcal{E}_q)$, where $\mathcal{V}_q = \mathrm{dFl}^*(G)$ and $\mathcal{E}_q$ is the set of all q-arcs (σ, τ).

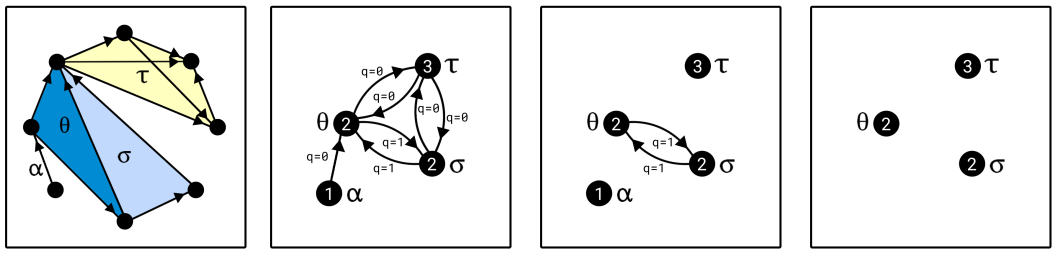
In analogy with graph theory, the maximal q-digraph can be represented in terms of a *q-adjacency matrix*, which is defined as follows.

Definition 3.3.25. Let $\mathcal{G}_q$ be the maximal q-digraph of $\mathrm{dFl}(G)$. The *maximal q-adjacency matrix* of $\mathcal{G}_q$, denoted by $\mathcal{H}_q = \mathcal{H}_q(\mathcal{G}_q)$, is a real square matrix whose entries are given by

$$\big(\mathcal{H}_q\big)_{ij} = \begin{cases} 1, \text{ if } \sigma_i \sim^{+}_{A_{q^*}} \sigma_j, \\ 0, \text{ if } i = j \text{ or } \sigma_i \not\sim^{+}_{A_{q^*}} \sigma_j. \end{cases} \tag{3.50}$$

Note that, by Proposition 3.3.2, we can replace the maximal q-adjacency in the expression (3.50) with the strictly lower q-adjacency.

Example 3.3.5. Consider the directed flag complex shown in Figure 3.19a. Figures 3.19b, 3.19c, and 3.19d represent its respective maximal q-digraphs $\mathcal{G}_q$ for $q = 0, 1, 2$.



(a) $\mathrm{dFl}(G)$. (b) $\mathcal{G}_0$. (c) $\mathcal{G}_1$. (d) $\mathcal{G}_2$.

Figure 3.19: A directed flag complex and its respective maximal q-digraphs, for $q = 0, 1, 2$. The numbers inside the nodes represent the dimensions of the respective directed simplices.

Observation 3.3.2. The weakly and strongly connected components of a maximal q-digraph $\mathcal{G}_q$ are equivalent to the maximal weakly and strongly q-connected components as defined in Definition 3.3.22 and Definition 3.3.21, respectively, where the set $\mathrm{dFl}(G)$ is replaced by $\mathrm{dFl}^*(G)$. Moreover, the largest weakly q-connected component of $\mathcal{G}_q$ is called its *giant q-component*.

In the case where we have a weighted directed flag complex obtained from a weighted digraph, by definition, the corresponding maximal q-digraphs will be node-weighted digraphs, since their nodes represent directed simplices. Accordingly, in order to obtain edge-weighted digraphs, we need a node-to-edge weight function. In the literature, there is a myriad of methods to transform a node-weighted digraph into an edge-weighted digraph [72], however, since the relations in a digraph can be non-symmetric, we would

like a non-symmetric transformation function, thus, for a given $(\mathrm{dFl}(G), \widetilde{\omega})$ and for a given pair $\sigma_i \sim^{+}_{A_{q^*}} \sigma_j$, we consider the following node-to-edge weight functions:

$$f(\widetilde{\omega}(\sigma_i), \widetilde{\omega}(\sigma_j)) = \widetilde{\omega}(\sigma_i), \tag{3.51}$$

$$f(\widetilde{\omega}(\sigma_i), \widetilde{\omega}(\sigma_j)) = \widetilde{\omega}(\sigma_j). \tag{3.52}$$

This leads us to extend our definition of maximal q-digraphs to the weighted case as follows.

Definition 3.3.26. Given a weighted digraph G^{ω}, the *weighted maximal q-digraph* of $(\mathrm{dFl}(G^{\omega}), \widetilde{\omega})$, denoted by $\mathcal{G}_q^{\widetilde{\omega}}$, is the digraph whose vertices are the simplices of $\mathrm{dFl}_q^*(G^{\omega})$ and for each pair $\sigma, \tau \in \mathrm{dFl}_q^*(G^{\omega})$ there is a weighted arc from σ to τ if $\sigma \sim^{+}_{A_{q^*}} \tau$, with $0 \leq q \leq \dim \mathrm{dFl}(G^{\omega})$, such that the weight of the arc is given by a node-to-edge weight function.

In addition, we may use the notation $\mathcal{G}_q^{\widetilde{\omega}} = (\mathcal{V}_q, \mathcal{E}_q, \widetilde{\omega})$, where $\mathcal{V}_q = \mathrm{dFl}_q^*(G^{\omega})$, $\mathcal{E}_q$ is the set of all q-arcs (σ, τ), and $\widetilde{\omega}$ is the product-weight function.

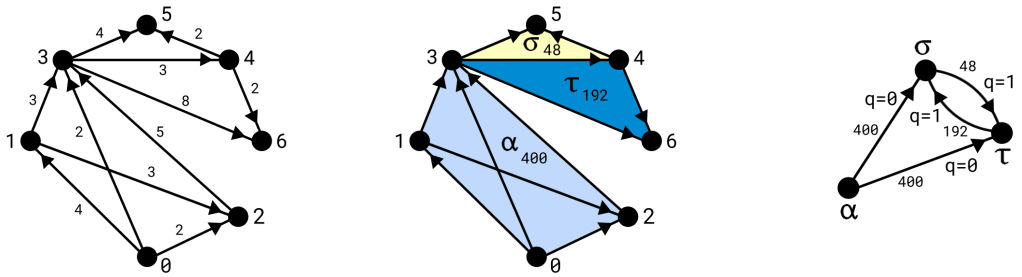
Definition 3.3.27. Let $\mathcal{G}_q^{\widetilde{\omega}}$ be the weighted maximal q-digraph of $(\mathrm{dFl}(G^{\omega}), \widetilde{\omega})$. The *weighted maximal q-adjacency matrix* of $\mathcal{G}_q^{\widetilde{\omega}}$, denoted by $\mathcal{H}_q^{\widetilde{\omega}} = \mathcal{H}_q^{\widetilde{\omega}}(\mathcal{G}_q^{\widetilde{\omega}})$, is a real square matrix whose entries are given by

$$\left(\mathcal{H}_q^{\widetilde{\omega}}\right)_{ij} = \begin{cases} f(\widetilde{\omega}(\sigma_i), \widetilde{\omega}(\sigma_j)), & \text{if } \sigma_i \sim^{+}_{A_{q^*}} \sigma_j, \\ 0, & \text{if } i = j \text{ or } \sigma_i \not\sim^{+}_{A_{q^*}} \sigma_j, \end{cases} \tag{3.53}$$

where f is some non-symmetric node-to-edge weight function.

Example 3.3.6. Consider the (non-normalized) weighted digraph G^{ω} as depicted in Figure 3.20a. Figure 3.20b depicts its weighted directed flag complex $(\mathrm{dFl}(G^{\omega}), \widetilde{\omega})$ in which the simplices are shown with their respective (non-normalized) weights. The weight function $\widetilde{\omega}$ is the product-weight function (3.4) such that the edge-to-node function is given by Definition 3.1.11. Thus, considering the node-to-edge weight function (3.51), the corresponding weighted maximal 0-digraph $\mathcal{G}_0^{\widetilde{\omega}}$ is the weighted digraph presented in Figure 3.20c. Also, both maximal 0-adjacency matrices of $\mathcal{G}_0$ and $\mathcal{G}_0^{\widetilde{\omega}}$ were computed in (3.54).

$$\mathcal{H}_0 = \begin{array}{c} \\ \sigma \\ \tau \\ \alpha \end{array} \begin{array}{c} \begin{array}{ccc} \sigma & \tau & \alpha \end{array} \\ \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \end{array}, \quad \mathcal{H}_0^{\widetilde{\omega}} = \begin{array}{c} \\ \sigma \\ \tau \\ \alpha \end{array} \begin{array}{c} \begin{array}{ccc} \sigma & \tau & \alpha \end{array} \\ \begin{bmatrix} 0 & 48 & 0 \\ 192 & 0 & 0 \\ 400 & 400 & 0 \end{bmatrix} \end{array}. \tag{3.54}$$



(a) G^{ω}. (b) $(\mathrm{dFl}(G^{\omega}), \widetilde{\omega})$. (c) $\mathcal{G}_0^{\widetilde{\omega}}$.

Figure 3.20: A weighted digraph together with its weighted directed flag complex and its weighted maximal 0-digraph.

Furthermore, the directed simplicial q-distance between two vertices of a maximal q-digraph can be written in terms of the entries of its maximal q-adjacency matrix, $\mathcal{H}_q = (h_{\sigma\tau})$:

$$\vec{d}_q(\sigma, \tau) = \sum_{\sigma', \tau' \in s^q_{\sigma \to \tau}} h_{\sigma'\tau'}, \tag{3.55}$$

where $s^q_{\sigma\to\tau}$ is the shortest directed simplicial q-walk from σ to τ. Similarly, for a weighted maximal q-digraph, we can define the *weighted directed simplicial q-distance* in term of the entries of its weighted maximal q-adjancecy matrix:

$$\vec{d}^{\,\widetilde{\omega}}_q(\sigma, \tau) = \sum_{\sigma', \tau' \in s^q_{\sigma \to \tau}(F)} F\big(f(\widetilde{\omega}(\sigma'), \widetilde{\omega}(\tau')\big), \tag{3.56}$$

where f is some non-symmetric node-to-edge weight function, F is a weight-to-distance function, and $s^q_{\sigma\to\tau}(F)$ is the shortest directed simplicial q-walk from σ to τ with respect to F.

Stars, Hubs, and Links

In this part, we extend the definitions of stars, hubs, and links introduced in Subsection 3.3.1 to directed simplices.

Definition 3.3.28. Given a directed simplex $\sigma^{(n)} \in \mathrm{dFl}(G)$, for $0 \leq q \leq n$ and for $\bullet \in \{-, +, \pm\}$, we have the following definitions:

1. The *lower* $(\bullet)$*-q-star* of $\sigma^{(n)}$ is the set defined by

$$\mathrm{st}^{\bullet}_{L_q}(\sigma^{(n)}) = \{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{L_q} \tau^{(m)}\}. \tag{3.57}$$

2. The *strictly lower* $(\bullet)$*-q-star* of $\sigma^{(n)}$ is the set defined by

$$\mathrm{st}^{\bullet}_{L_{q^*}}(\sigma^{(n)}) = \{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{L_{q^*}} \tau^{(m)}\}. \tag{3.58}$$

3. The *upper* $(\bullet)$-*q-star* of $\sigma^{(n)}$ is the set defined by

$$\mathrm{st}^{\bullet}_{U_q}(\sigma^{(n)}) = \{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{U_q} \tau^{(m)}\}. \tag{3.59}$$

4. The *strictly upper* $(\bullet)$-*q-star* of $\sigma^{(n)}$ is the set defined by

$$\mathrm{st}^{\bullet}_{U_{q^*}}(\sigma^{(n)}) = \{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{U_{q^*}} \tau^{(m)}\}. \tag{3.60}$$

5. The $(\bullet)$-*q-star* of $\sigma^{(n)}$ is the set defined by

$$\mathrm{st}^{\bullet}_{A_q}(\sigma^{(n)}) = \{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{A_q} \tau^{(m)}\}. \tag{3.61}$$

6. The *maximal* $(\bullet)$-*q-star* of $\sigma^{(n)}$ is the set defined by

$$\mathrm{st}^{\bullet}_{A_{q^*}}(\sigma^{(n)}) = \{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{A_{q^*}} \tau^{(m)}\}. \tag{3.62}$$

Notice that $\mathrm{st}^{\pm}(\sigma) = \mathrm{st}^{+}(\sigma) \cap \mathrm{st}^{-}(\sigma)$, for any of the q-stars defined above.

The definition of the hub of a simplicial family of directed simplices obtained from $\mathrm{dFl}(G)$ is exactly the same as the Definition 3.3.7, since it is defined as the set of faces that are shared by the elements of the family, regardless of the direction of the connection between them. We formalize this fact as follows.

Definition 3.3.29. Let $\mathcal{F}(G)$ denote a simplicial family of directed simplices obtained from $\mathrm{dFl}(G)$. The *hub* of $\mathcal{F}(G)$ is the set formed by all directed simplices that are common faces of the elements of $\mathcal{F}(G)$, i.e.

$$\mathrm{hub}(\mathcal{F}(G)) = \bigcap_{\sigma \in \mathcal{F}(G)} \sigma. \tag{3.63}$$

Analogously to Definition 3.3.8 (link of a simplex), we can generalize the idea of in- and out-neighborhood of a node in a digraph to directed simplices through the concepts of *in-* and *out-link*.

Definition 3.3.30. The *in-link* and *out-link* of a directed simplex $\sigma^{(n)} \in \mathrm{dFl}(G)$ are defined, respectively, by

$$\mathrm{lk}^{-}(\sigma^{(n)}) = \{\tau^{(m)} \in \mathrm{dFl}(G) | \sigma^{(n)} \cap \tau^{(m)} = \emptyset, \sigma^{(n)} \sim^{-}_{U_p} \tau^{(m)}\}, \tag{3.64}$$

$$\mathrm{lk}^{+}(\sigma^{(n)}) = \{\tau^{(m)} \in \mathrm{dFl}(G) | \sigma^{(n)} \cap \tau^{(m)} = \emptyset, \sigma^{(n)} \sim^{+}_{U_p} \tau^{(m)}\}, \tag{3.65}$$

where $p = n + m + 1$.

Notice that if σ is a simplex in the underlying flag complex of $\mathrm{dFl}(G)$, then $\mathrm{lk}(\sigma) = \mathrm{lk}^-(\sigma) \cup \mathrm{lk}^+(\sigma) - (\mathrm{lk}^-(\sigma) \cap \mathrm{lk}^+(\sigma))$.

Example 3.3.7. Considering the directed flag complex depicted in Figure 3.21, we have the following examples: the in- and out-link of the arc $[0,3]$ are $\mathrm{lk}^-([0,3]) = \{[4],[1,9]\}$ and $\mathrm{lk}^+([0,3]) = \{[1,9]\}$; the maximal $(\bullet)$-1-star of σ_1 are $\mathrm{st}^-_{A_{1^*}}(\sigma_1) = \{\sigma_2\}$ and $\mathrm{st}^+_{A_{1^*}}(\sigma_1) = \emptyset$; the maximal $(\bullet)$-1-star of σ_6 are $\mathrm{st}^-_{A_{1^*}}(\sigma_6) = \{\sigma_5, \sigma_7\}$ and $\mathrm{st}^+_{A_{1^*}}(\sigma_6) = \{\sigma_5, \sigma_3\}$.

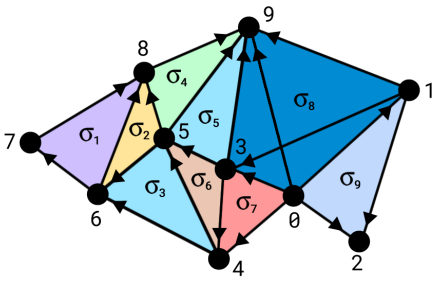


Figure 3.21: A directed flag complex.

Lower, Upper, and General Degrees

Based on the previous definitions of q-stars, in this last part we extend the definitions of lower, upper, and general degrees to directed simplices.

Definition 3.3.31. Given a directed simplex $\sigma^{(n)} \in \mathrm{dFl}(G)$, for $0 \leq q \leq n$ and for $\bullet \in \{-,+,\pm\}$, we have the following definitions:

1. The *lower* $(\bullet)$-*q*-*degree* of $\sigma^{(n)}$ is defined by

$$\deg^{\bullet}_{L_q}(\sigma^{(n)}) = |\mathrm{st}^{\bullet}_{L_q}(\sigma^{(n)})| = \#\{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{L_q} \tau^{(m)}\}. \tag{3.66}$$

2. The *strictly lower* $(\bullet)$-*q*-*degree* of $\sigma^{(n)}$ is defined by

$$\deg^{\bullet}_{L_{q^*}}(\sigma^{(n)}) = |\mathrm{st}^{\bullet}_{L_{q^*}}(\sigma^{(n)})| = \#\{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{L_{q^*}} \tau^{(m)}\}. \tag{3.67}$$

3. The *upper* $(\bullet)$-*q*-*degree* of $\sigma^{(n)}$ is defined by

$$\deg^{\bullet}_{U_q}(\sigma^{(n)}) = |\mathrm{st}^{\bullet}_{U_q}(\sigma^{(n)})| = \#\{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{U_q} \tau^{(m)}\}. \tag{3.68}$$

4. The *strictly upper* $(\bullet)$-*q*-*degree* of $\sigma^{(n)}$ is defined by

$$\deg^{\bullet}_{U_{q^*}}(\sigma^{(n)}) = |\mathrm{st}^{\bullet}_{U_{q^*}}(\sigma^{(n)})| = \#\{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{U_{q^*}} \tau^{(m)}\}. \tag{3.69}$$

5. The $(\bullet)$-*q-degree* of $\sigma^{(n)}$ is defined by

$$\deg^{\bullet}_{A_q}(\sigma^{(n)}) = |\mathrm{st}^{\bullet}_{A_q}(\sigma^{(n)})| = \#\{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{A_q} \tau^{(m)}\}. \tag{3.70}$$

6. The *maximal* $(\bullet)$-*q-degree* of $\sigma^{(n)}$ is defined by

$$\deg^{\bullet}_{A_{q^*}}(\sigma^{(n)}) = |\mathrm{st}^{\bullet}_{A_{q^*}}(\sigma^{(n)})| = \#\{\tau^{(m)} \in \mathrm{dFl}(G) : \sigma^{(n)} \sim^{\bullet}_{A_{q^*}} \tau^{(m)}\}. \tag{3.71}$$

Notice that if σ is a simplex in the underlying flag complex of $\mathrm{dFl}(G)$, then its q-degree, for any of the lower, upper, and general adjacencies, is equal to

$$\deg(\sigma) = \deg^{-}(\sigma) + \deg^{+}(\sigma) - \deg^{\pm}(\sigma). \tag{3.72}$$

Example 3.3.8. Considering the directed flag complex depicted in Figure 3.21, we have the following examples: $\deg^{-}_{A_{1^*}}(\sigma_1) = 1$ and $\deg^{+}_{A_{1^*}}(\sigma_1) = 0$; $\deg^{-}_{A_{1^*}}(\sigma_6) = 2$ and $\deg^{+}_{A_{1^*}}(\sigma_6) = 2$.

It's important to note that if we are considering solely the elements of $\mathrm{dFl}^{*}_{q}(G)$, i.e. the maximal directed simplices, then the maximal $(\bullet)$-q-degrees can be written in terms of the entries of the q-adjacency matrix of the maximal q-digraph, $\mathcal{H}_q = (h_{\sigma\tau})$:

$$\deg^{+}_{A_{q^*}}(\sigma) = \sum_{\tau \in \mathrm{st}^{+}_{A_{q^*}}(\sigma)} h_{\sigma\tau}, \tag{3.73}$$

$$\deg^{-}_{A_{q^*}}(\sigma) = \sum_{\tau \in \mathrm{st}^{-}_{A_{q^*}}(\sigma)} h_{\tau\sigma}. \tag{3.74}$$

On the other hand, in the case where we have a weighted directed flag complex $(\mathrm{dFl}(G^{\omega}), \widetilde{\omega})$, the *weighted maximal* $(\bullet)$-*q-degrees* can be written as:

$$\deg^{\omega,+}_{A_{q^*}}(\sigma) = \sum_{\tau \in \mathrm{st}^{+}_{A_{q^*}}(\sigma)} f(\widetilde{\omega}(\sigma), \widetilde{\omega}(\tau)), \tag{3.75}$$

$$\deg^{\omega,-}_{A_{q^*}}(\sigma) = \sum_{\tau \in \mathrm{st}^{-}_{A_{q^*}}(\sigma)} f(\widetilde{\omega}(\tau), \widetilde{\omega}(\sigma)), \tag{3.76}$$

where f is some non-symmetric node-to-edge weight function.

Chapter 4

Quantitative Approaches to Digraph-Based Complexes

> *En un mot, pour tirer la loi de l'expérience, if faut généraliser; c'est une nécessité qui s'impose à l'observateur le plus circonspect. (In one word, to draw the rule from experience, one must generalize; this is a necessity that imposes itself on the most circumspect observer.)*
>
> — Henri Poincaré [210]

In this chapter, we extend several quantifiers and similarity comparison methods defined for (weighted and unweighted) graphs in Chapter 2, such as distance-related measures, measures of centrality, segregation, spectrum-related measures, graph kernels, etc., which take into account their abstract, algebraic, and topological properties, to (weighted and unweighted) digraph-based complexes, specially for directed flag complexes, and, in addition, we introduce some new measures.

Specifically, we extend all of the quantitative methods defined previously for graphs to maximal q-digraphs obtained from directed flag complexes, so we can quantify the higher-order connectivity and topological properties of the directed network at each level q, i.e. we can analyze quantitatively each level of organization of the directed cliques in the network taking into account their directed higher-order connectivities.

4.1 Quantitative Graph Theory and Beyond

In this first section, we present some conceptual descriptions and examples related to quantitative graph theory; subsequently, we discuss some recently introduced quantitative methods for simplicial complexes and their importance as a basis for a quantitative theory of digraph-based complexes.

4.1.1 Quantitative Graph Theory

According to Dehmer et al. [74], the "quantitative graph theory (QGT) deals with the quantification of structural aspects of graphs, instead of characterizing graphs only descriptively." The QGT can be considered a new branch within graph theory [74, 76], and the reason why its introduction was necessary is that most of the methods from classical graph theory for graph analysis are descriptive.

Moreover, we can divide the QGT into two broad categories, namely: *graph characterization* and *comparative graph analysis* [76]. In what follows, we present a description of these categories and provide some examples.

Graph Characterization: Graph characterization is concerned with describing some property of the network through a local or global numerical graph invariant. Examples of numerical graph invariants are: distance-based measures (characteristic path length (2.17), eccentricity (2.18)); centrality measures (closeness centrality (2.30), betweenness centrality (2.33)); complexity/information-theoretic measures (graph entropy (2.47)); and spectrum-based measures (graph energy (2.50)).

Comparative Graph Analysis: Comparative graph analysis is concerned with methods to compare the structural similarity or structural distance between two or more graphs. As already commented in Section 2.4, there are two classes of methods of graph similarity comparison: *statistical comparison methods* and *distance-based comparison algorithms*. Examples of the first class are: applying numerical invariants in different groups of graphs and comparing them using statistical tests; and examples of the second class are: graph edit distances; graph kernels; and structure distances.

In the past two paragraphs, we presented just a few examples among a myriad of measures and methods used to characterize and compare graphs in several different scientific disciplines, and it is up to the reader to look over the references if they want to learn more [74, 76].

4.1.2 Beyond QGT: Quantitative Simplicial Theory

In view of the recent developments of quantitative methods for the analysis of simplicial complexes, we can draw an analogy with QGT and propose a "quantitative simplicial theory," with the same conceptual characteristics described in the previous section for QGT, i.e. as the field which "deals with the quantification of structural aspects of simplicial complexes, instead of characterizing simplicial complexes only descriptively."

Making another analogy with QGT, we can divide the "quantitative simplicial theory" into two categories: *simplicial characterization* and *simplicial similarity comparison mehtods*. The description of these categories is analogous to those of the QGT;

thus, below we present some examples and the respective references of the methods belonging to each of these categories:

Simplicial Characterization: simplical distances and simplical eccentricities [150, 233]; simplicial centralities [92, 233]; simplicial clustering coefficients [178, 233]; simplicial entropies [23, 68, 177]; discrete curvatures [284]; simplicial energies [161].

Simplicial Similarity Comparison Methods: distances between persistent diagrams [83]; distances between vectorized persistence summaries [97]; simplicial kernels [180, 291]; distances between structure vectors.

Most of these simplicial quantitative methods can be extended to directed flag complexes, and some of them can also be extended to path complexes, as we will see in the next sections.

4.2 Simplicial Characterization Measures

In this section, we introduce novel simplicial analogues of the digraph measures (graph invariants) presented in Section 2.3 to maximal q-digraphs (henceforth simply referred to as q-digraphs) associated with directed flag complexes. We use a similar textual organizational structure as used in the aforementioned section, that is, we start by presenting the distance-based simplicial measures, followed by the simplicial centralities, simplicial segregation measures, simplicial entropies, discrete curvatures, and finally the spectrum-related simplicial measures. It is worth to point out that all these measures are based on the directed high-order connectivity of the complex in a certain way.

Throughout the next subsections, $\mathrm{dFl}(G)$ will denote the directed flag complex of a given simple digraph G *without double-edges*, and $\mathcal{G}_q = (\mathcal{V}_q, \mathcal{E}_q)$ will denote its q-digraph with maximal q-adjacency matrix (henceforth simply referred to as q-adjacency matrix) $\mathcal{H}_q = (h_{\sigma\tau})$, for $0 \leq q \leq \dim \mathrm{dFl}(G)$.

4.2.1 Distance-Based Simplicial Measures

In this part, we extend the distance-based measures as defined for digraphs in Subsection 2.3.1 to q-digraphs. These new simplicial measures can be seen as measures of higher-order global integration of a directed network, that is, how the network is integrated at various levels of organization.

Average Shortest Directed Simplicial q-Walk Length

The *average shortest directed simplicial q-walk length* of $\mathcal{G}_q$ is defined as the simplicial analogue of the directed version of the average shortest path length (2.17), i.e.

$$\vec{L}_q(\mathcal{G}_q) = \frac{1}{|\mathcal{V}_q|}\sum_{\sigma\in\mathcal{V}_q}\frac{\sum_{\tau\in\mathcal{V}_q,\tau\neq\sigma}\vec{d}_q(\sigma,\tau)}{|\mathcal{V}_q|-1} = \sum_{\substack{\sigma,\tau\in\mathcal{V}_q\\ \sigma\neq\tau}}\frac{\vec{d}_q(\sigma,\tau)}{|\mathcal{V}_q|(|\mathcal{V}_q|-1)}. \tag{4.1}$$

For a weighted q-digraph $\mathcal{G}_q^{\widetilde{\omega}}$, the weighted version of the formula (4.1) is obtained by replacing $\vec{d}_q$ with $\vec{d}_q^{\,\omega}$. Also, for computational purposes, we may consider $\vec{d}_q(\sigma,\tau) = 0$ instead of $\vec{d}_q(\sigma,\tau) = \infty$ when $(\sigma,\tau) \notin \mathcal{E}_q$, otherwise $|\mathcal{V}_q|$ must be replaced with the order of the giant q-component.

Directed Simplicial q-Eccentricity

In the literature, there are different ways to define the eccentricity of a simplex [13, 150]. Here, however, we extend the definition of simplicial eccentricity as proposed in [233] to directed simplices. Following the formula (2.18), for a strongly q-connected q-digraph $\mathcal{G}_q$, we define the *directed simplicial q-eccentricity* of $\sigma \in \mathcal{V}_q$ as the maximum directed simplicial q-distance from σ to any other $\tau \in \mathcal{V}_q$, i.e.

$$\vec{\mathrm{ecc}}_q(\sigma) = \max_{\tau\in\mathcal{V}_q}\vec{d}_q(\sigma,\tau). \tag{4.2}$$

Considering the formula (4.2), the *directed simplicial q-diameter* and the *directed simplicial q-radius* of $\mathcal{G}_q$ are defined in an analogous way to the formulas (2.19) and (2.20), respectively.

Directed Simplicial Global q-Efficiency

The *directed simplicial global q-efficiency* of $\mathcal{G}_q$ is defined as the simplicial analogue of the directed version of the global efficiency (2.21), i.e.

$$\vec{E}^q_{glob}(\mathcal{G}_q) = \frac{1}{|\mathcal{V}_q|}\sum_{\sigma\in\mathcal{V}_q}\frac{\sum_{\tau\in\mathcal{V}_q,\tau\neq\sigma}\vec{d}_q^{-1}(\sigma,\tau)}{|\mathcal{V}_q|-1}. \tag{4.3}$$

For a weighted q-digraph $\mathcal{G}_q^{\widetilde{\omega}}$, the weighted version of the formula (4.3) is obtained by replacing $\vec{d}_q$ with $\vec{d}_q^{\,\omega}$.

Simplicial q-Communicability

The *simplicial q-communicability* between two directed simplices $\sigma, \tau \in \mathcal{V}_q$ is defined as the simplicial analogue of the communicability between two vertices in a digraph

(2.23), and therefore it can be written in terms of the powers of the q-adjacency matrix as

$$CM_q(\sigma,\tau) = \sum_{k=0}^{\infty} \frac{(\mathcal{H}_q^k)_{\sigma\tau}}{k!} = (\exp(\mathcal{H}_q))_{\sigma\tau}. \tag{4.4}$$

Be aware that the entire discussion on the properties of the adjacency matrices of digraphs made in Subsection 2.2.1 are equally valid for the q-adjacency matrices, therefore $(\mathcal{H}_q^k)_{\sigma\tau}$ represents the number of directed simplicial q-walks of length k from σ to τ.

Simplicial q-Returnability

The *simplicial q-returnability* of $\mathcal{G}_q$ is defined as the simplicial analogue of the returnability of a digraph (2.23), and therefore it can be written in terms of the powers of the q-adjacency matrix as

$$K_{r,q}(\mathcal{G}_q) = \sum_{k=2}^{\infty} \frac{\mathrm{Tr}(\mathcal{H}_q^k)}{k!} = \mathrm{Tr}(\exp(\mathcal{H}_q)) - |\mathcal{V}_q|. \tag{4.5}$$

Moreover, we can define the *relative simplicial q-returnability* as the simplicial analogue of the relative returnability (4.6), i.e.

$$K'_{r,q}(\mathcal{G}_q) = \frac{\mathrm{Tr}(\exp(\mathcal{H}_q)) - |\mathcal{V}_q|}{\mathrm{Tr}(\exp(\mathcal{H}'_q)) - |\mathcal{V}_q|}, \tag{4.6}$$

where $\mathcal{H}'_q$ is the q-adjacency matrix of the underlying q-graph.

4.2.2 Simplicial Centrality Measures

In this part, we extend the measures of centrality as defined for digraphs in Subsection 2.3.2 to q-digraphs. These new simplicial measures can be interpreted as measures that try to quantify the "importance," "influence," or "centrality" of a directed clique within the higher-order settings, that is, its "centrality" at various levels of organization of the network.

Simplicial Degree Centralities

The *simplicial in-q-degree centrality* of $\sigma \in \mathcal{V}_q$ is defined as the simplicial analogue of the in-degree centrality of a node (2.28), i.e.

$$C^{-}_{\deg_q}(\sigma) = \frac{\deg^{-}_{A_{q^*}}(\sigma)}{|\mathcal{V}_q| - 1}. \tag{4.7}$$

In the same way, the *simplicial out-q-degree centrality* of σ is defined as the simplicial

analogue of the out-degree centrality of a node (2.29), i.e.

$$C^{+}_{\deg_q}(\sigma) = \frac{\deg^{+}_{A_{q^*}}(\sigma)}{|\mathcal{V}_q| - 1}. \tag{4.8}$$

Also, if σ is a simplex in the corresponding underlying q-graph, then its q-degree centrality can be written as

$$C_{\deg_q}(\sigma) = C^{-}_{\deg_q}(\sigma) + C^{+}_{\deg_q}(\sigma) - \frac{\deg^{\pm}_{A_{q^*}}}{|\mathcal{V}_q| - 1}. \tag{4.9}$$

To obtain the *weighted simplicial in-q-degree centrality* we simply replace $\deg^{-}_{A_{q^*}}$ with $\deg^{-,\omega}_{A_{q^*}}$ (see formula (3.75)), and to obtain the *weighted simplicial out-q-degree centrality* we replace $\deg^{+}_{A_{q^*}}$ with $\deg^{+,\omega}_{A_{q^*}}$ (see formula (3.76)).

Remark 4.2.1. Since $\sigma \in \mathrm{dFl}^{*}_{q}(G)$, the maximal q-degree in the formulas (4.7) and (4.8) can be replace by the strictly lower q-degree.

Directed Simplicial Closeness Centrality

The *directed simplicial q-closeness centrality* of $\sigma \in \mathcal{V}_q$ is defined as the simplicial analogue of the directed version of the closeness centrality (2.30), i.e.

$$\vec{Cl}_q(\sigma) = \frac{1}{\sum_{\substack{\tau \in \mathcal{V}_q \\ \tau \neq \sigma}} \vec{d}_q(\sigma, \tau)}. \tag{4.10}$$

The formula (4.10) is defined for weakly q-connected q-digraphs, thus if N_q is the order of the giant q-component of $\mathcal{G}_q$, we define the *normalized directed simplicial q-closeness centrality* by

$$\vec{Cl}_q(\sigma) = \frac{N_q - 1}{\sum_{\substack{\tau \in \mathcal{V}_q \\ \tau \neq \sigma}} \vec{d}_q(\sigma, \tau)}. \tag{4.11}$$

For the weighted case, we simply replace $\vec{d}_q$ with $\vec{d}^{\omega}_q$.

Remark 4.2.2. For the corresponding underlying q-graph of $\mathcal{G}_q$, the formula (4.10) corresponds to the q-closeness centrality as introduced in [233] for simplicial complexes.

Directed Simplicial Harmonic Centrality

The *directed simplicial q-harmonic centrality* of $\sigma \in \mathcal{V}_q$ is defined as the simplicial analogue of the directed version of the harmonic centrality (2.32), i.e.

$$\vec{HC}_q(\sigma) = \sum_{\substack{\tau \in \mathcal{V}_q \\ \tau \neq \sigma}} \frac{1}{\vec{d}_q(\sigma, \tau)}, \tag{4.12}$$

where the convention $1/\infty = 0$ is adopted. Unlike the directed simplicial q-closeness centrality, the directed simplicial q-harmonic centrality can be computed for disconnected q-digraphs. Also, for the weighted case, we simply replace $\vec{d}_q$ with $\vec{d}_q^{\omega}$.

Directed Simplicial Betweenness Centrality

The *directed simplicial q-betweenness centrality* of $\sigma \in \mathcal{V}_q$ is defined as the simplicial analogue of the directed version of the between centrality (2.33), i.e.

$$\vec{B}_q(\sigma) = \sum_{\substack{\tau,\tau' \in \mathcal{V}_q \\ \tau' \neq \tau \neq \sigma}} \frac{\vec{l}^{\,q}_{\tau'\tau}(\sigma)}{\vec{l}^{\,q}_{\tau'\tau}}, \tag{4.13}$$

where $\vec{l}^{\,q}_{\tau'\tau}(\sigma)$ is the number of shortest directed simplicial q-walks from τ' to τ passing through σ, and $\vec{l}^{\,q}_{\tau'\tau}$ is the total number of shortest directed simplicial q-walks from τ' to τ.

The formula (4.13) is defined for weakly q-connected q-digraphs, thus if N_q is the order of the giant q-component of $\mathcal{G}_q$, we define the *normalized directed simplicial q-betweenness centrality* of σ by

$$\vec{B}_q(\sigma) = \frac{1}{(N_q - 1)(N_q - 2)} \sum_{\substack{\tau,\tau' \in \mathcal{V}_q \\ \tau' \neq \tau \neq \sigma}} \frac{\vec{l}^{\,q}_{\tau'\tau}(\sigma)}{\vec{l}^{\,q}_{\tau'\tau}}. \tag{4.14}$$

For the weighted case, we simply consider the weighted version $\vec{l}^{\,\omega,q}_{\tau'\tau}(\sigma)$, where the shortest directed simplicial q-walks are computed in relation to a weight-to-distance function.

Remark 4.2.3. For the corresponding underlying q-graph of $\mathcal{G}_q$, the formula (4.13) corresponds to the q-betweenness centrality as introduced in [233] for simplicial complexes.

Simplicial Reaching Centrality

The *simplicial local q-reaching centrality* of $\sigma \in \mathcal{V}_q$ is defined as the simplicial analogue of the local reaching centrality (2.35), i.e.

$$C_{R,q}(\sigma) = \frac{r_{\mathcal{G}_q}(\sigma)}{|\mathcal{V}_q| - 1}, \tag{4.15}$$

where $r_{\mathcal{G}_q}(\sigma)$ is the number of vertices in $\mathcal{G}_q$ which are reachable from σ. Define $C_{R,q}^{max} = \max_{\sigma \in \mathcal{V}_q} C_{R,q}(\sigma)$. The *simplicial global q-reaching centrality* is defined as the simplicial

analogue of the global reaching centrality (2.36), i.e.

$$GRC_q(\mathcal{G}_q) = \frac{\sum_{\sigma \in \mathcal{V}_q} [C_{R,q}^{max} - C_{R,q}(\sigma)]}{|\mathcal{V}_q| - 1}. \tag{4.16}$$

Example 4.2.1. Consider the directed flag complex presented in Figure 4.1. Table 4.1 presents the values of the following simplicial centralities for all maximal simplices of this complex, for $q = 0, 1$: in/out-q-degree centrality, (non-normalized) directed simplicial q-betweenness centrality, and directed simplicial q-harmonic centrality. Note that at the level $q = 1$ we only have directed q-connectivity between σ_1 and σ_2, σ_1 and θ_2, and θ_1 and θ_2. Also, note that θ_1 has the largest $\vec{B}_0$ and the largest $C^+_{deg_0}$, suggesting that it may be the most central simplex in the complex at the level $q = 0$, but at the level $q = 1$, θ_2 may be the most central, since it has the largest $\vec{B}_1$ and the largest $C^-_{deg_1}$.

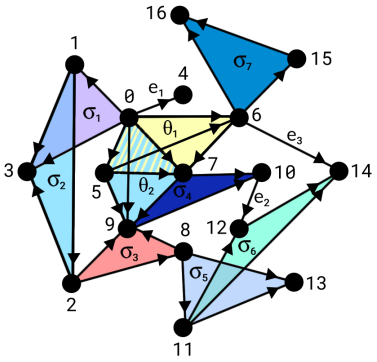


Figure 4.1: A directed flag complex.

Table 4.1: Directed simplicial centralities for $q = 0, 1$.

	$\mathbf{C}^-_{\mathbf{deg_0}}$	$\mathbf{C}^+_{\mathbf{deg_0}}$	$\mathbf{C}^-_{\mathbf{deg_1}}$	$\mathbf{C}^+_{\mathbf{deg_1}}$	$\vec{\mathbf{B}}_0$	$\vec{\mathbf{B}}_1$	$\vec{\mathbf{HC}}_0$	$\vec{\mathbf{HC}}_1$
e_1	0.27	0.27	0.0	0.0	1.0	0.0	6.25	0.0
e_2	0.18	0.09	0.0	0.0	1.5	0.0	5.83	0.0
e_3	0.27	0.27	0.0	0.0	17.34	0.0	6.42	0.0
σ_1	0.27	0.36	0.0	0.09	5.5	0.0	6.25	0.0
σ_2	0.18	0.09	0.09	0.0	1.0	0.0	5.75	1.0
σ_3	0.36	0.36	0.0	0.0	26.17	0.0	6.99	0.0
σ_4	0.27	0.36	0.09	0.09	15.17	0.0	6.58	1.5
σ_5	0.18	0.18	0.0	0.0	9.8	0.0	5.83	0.0
σ_6	0.27	0.27	0.0	0.0	18.67	0.0	6.34	0.0
σ_7	0.18	0.18	0.0	0.0	0.0	0.0	5.58	0.0
θ_1	0.45	0.54	0.0	0.09	34.17	0.0	7.49	0.0
θ_2	0.45	0.36	0.18	0.09	13.67	1.0	7.58	2.0

4.2.3 Simplicial Segregation Measures

In this part, we extend the measures of segregation as defined for digraphs in Subsection 2.3.3 to q-digraphs. These new simplicial measures can be interpreted as measures that attempt to quantify the tendency of directed cliques to segregate into higher-order clusters or higher-order communities in a directed network.

Directed Simplicial Clustering Coefficients

The *average directed simplicial q-clustering coefficient* of $\mathcal{G}_q$ is defined as the simplicial analogue of the average directed q-clustering coefficient (2.40), i.e.

$$\vec{\bar{C}}_q(\mathcal{G}_q) = \frac{1}{|\mathcal{V}_q|} \sum_{\sigma \in \mathcal{V}_q} \frac{\vec{T}_q(\sigma)}{\deg^{tot}_{A_{q^*}}(\sigma)(\deg^{tot}_{A_{q^*}}(\sigma) - 1) - 2\deg^{\pm}_{A_{q^*}}(\sigma)}, \tag{4.17}$$

where $\deg^{tot}_{A_{q^*}}(\sigma) = \deg^{-}_{A_{q^*}}(\sigma) + \deg^{+}_{A_{q^*}}(\sigma)$ and $\vec{T}_q(\sigma)$ is the number of directed triangles at the level q containing σ, which can be written in terms of the q-adjacency matrix entries in an analogous way to the formula (2.41):

$$\vec{T}(\sigma) = \frac{1}{2} \sum_{\tau', \tau \in \mathcal{V}_q} (h_{\sigma\tau'} + h_{\tau'\sigma})(h_{\sigma\tau} + h_{\tau\sigma})(h_{\tau'\tau} + h_{\tau\tau'}). \tag{4.18}$$

Moreover, Maletic et al. [178] proposed a simplicial clustering coefficient of a simplex based on its dimension and on the dimensions of the simplices of its neighborhood; we can adapt this coefficient for a simplex $\sigma^{(n)}$ in the underlying q-graph of $\mathcal{G}_q$ as

$$C_q(\sigma^{(n)}) = \sum_{\tau^{(m)} \in \mathrm{st}_{A_{q^*}}(\sigma)} \frac{2^{1+f_{\sigma\tau}} - 1}{2^n + 2^m - 1}, \tag{4.19}$$

where $f_{\sigma\tau}$ is the dimension of the face shared between σ and τ, and $\mathrm{st}_{A_{q^*}}(\sigma) = \mathrm{st}^{-}_{A_{q^*}}(\sigma) \cup \mathrm{st}^{+}_{A_{q^*}}(\sigma) - (\mathrm{st}^{-}_{A_{q^*}}(\sigma) \cap \mathrm{st}^{+}_{A_{q^*}}(\sigma))$ (the directed stars are computed considering the corresponding σ in $\mathcal{G}_q$).

Accordingly, in what follows, we propose directed variants of this simplicial coefficient for a directed simplex $\sigma^{(n)}$ in $\mathcal{G}_q$. The *simplicial $(\bullet)$-q-clustering coefficient*, with $\bullet \in \{-, +, \pm\}$, is defined by

$$\vec{C}^{\bullet}_q(\sigma^{(n)}) = \sum_{\tau^{(m)} \in \mathrm{st}^{\bullet}_{A_{q^*}}(\sigma)} \frac{2^{1+f_{\sigma\tau}} - 1}{2^n + 2^m - 1}. \tag{4.20}$$

Also, the simplicial clustering coefficient (4.19) can be written in terms of the sim-

plicial (•)-q-clustering coefficients as

$$C_q(\sigma) = \vec{C}_q^{-}(\sigma) + \vec{C}_q^{+}(\sigma) - \vec{C}_q^{\pm}(\sigma). \tag{4.21}$$

Directed Simplicial Rich-Club Coefficients

For an integer $k \geq 0$, the *simplicial in-q-degree rich-club coefficient* of $\mathcal{G}_q$ is defined as the simplicial analogue of the in-degree rich-club coefficient (2.44), i.e.

$$\phi_{q,k}^{-}(\mathcal{G}_q) = \frac{E_{>k}^{-}(q)}{F_{>k}^{-}(q)(F_{>k}^{-}(q)-1)}, \tag{4.22}$$

where $F_{>k}^{-}(q)$ is the number of vertices $\sigma \in \mathcal{V}_q$ having $\deg_{A_{q^*}}^{-}(\sigma) > k$, and $E_{>k}^{-}(q)$ is the number of q-arcs connecting those $F_{>k}^{-}(q)$ vertices.

Analogously, the *simplicial out-q-degree rich-club coefficient* of $\mathcal{G}_q$ is defined as the simplicial analogue of the out-degree rich-club coefficient (2.45), i.e.

$$\phi_{q,k}^{+}(\mathcal{G}_q) = \frac{E_{>k}^{+}(q)}{F_{>k}^{+}(q)(F_{>k}^{+}(q)-1)}, \tag{4.23}$$

where $F_{>k}^{+}(q)$ is the number of vertices $\sigma \in \mathcal{V}_q$ having $\deg_{A_{q^*}}^{+}(\sigma) > k$, and $E_{>k}^{+}(q)$ is the number of q-arcs connecting those $F_{>k}^{+}(q)$ vertices.

Directed Simplicial Local Efficiency

The *directed simplicial local q-efficiency* of $\sigma \in \mathcal{V}_q$ is defined as the simplicial analogue of the directed version of the local efficiency (2.46), i.e.

$$\vec{E}_{loc}^{q}(\sigma) = \frac{1}{|\mathcal{V}_q|} \sum_{\sigma \in \mathcal{V}_q} \vec{E}_{glob}^{q}(\mathcal{G}_q(\sigma)), \tag{4.24}$$

where $\mathcal{G}_q(\sigma)$ is the induced subdigraph of $\mathcal{G}_q$ formed by elements of the set $\mathrm{st}_{A_{q^*}}^{-}(\sigma) \cup \mathrm{st}_{A_{q^*}}^{+}(\sigma) - \mathrm{st}_{A_{q^*}}^{\pm}(\sigma) \cap \mathrm{st}_{A_{q^*}}^{+}(\sigma))$, excluding σ. For the weighted case, we simply replace $\vec{E}_{glob}^{q}$ with its weighted version.

4.2.4 Simplicial Entropies

In the literature, there are different ways to define entropies for simplicial complexes [23, 68, 177]. However, here we propose novel simplicial entropies associated with q-digraphs, namely: the *simplicial q-structural entropy* and the *simplicial in/out-q-degree distribution entropies*.

Simplicial Structural Entropy

Let us start by introducing a new concept. The *relative q-communicability* of a directed simplex $\sigma \in \mathcal{V}_q$, denoted by $RCM_q(\sigma)$, is defined as the fraction of all directed simplicial q-walks in $\mathcal{G}_q$ that starts in σ, i.e.

$$RCM_q(\sigma) = \frac{\sum_{\tau \in \mathcal{V}_q} CM_q(\sigma, \tau)}{\sum_{\tau' \in \mathcal{V}_q} \sum_{\tau'' \in \mathcal{V}_q} CM_q(\tau', \tau'')}. \tag{4.25}$$

The *simplicial q-structural entropy* of $\mathcal{G}_q$ is defined as the Shannon entropy of the relative q-communicabilities:

$$H_q^{str}(\mathcal{G}_q) = -\sum_{\sigma \in \mathcal{V}_q} RCM_q(\sigma) \log_2 RCM_q(\sigma). \tag{4.26}$$

This measure can be roughly interpreted as the "degree of higher-order structural disorder" in the network for each level q.

Simplicial Degree Distribution Entropy

Let $\delta_q^-(k)$ be the number of directed simplices having in-q-degree k and $\delta_q^+(k)$ be the number of directed simplices having out-q-degree k. The *in-q-degree distribution* and the *out-q-degree distribution* are defined, respectively, by

$$p_q^-(k) = \frac{\delta_q^-(k)}{|\mathcal{V}_q|}, \tag{4.27}$$

$$p_q^+(k) = \frac{\delta_q^+(k)}{|\mathcal{V}_q|}. \tag{4.28}$$

We define the *simplicial in-q-degree distribution entropy* of $\mathcal{G}_q$ as the simplicial analogue of the in-degree distribution entropy (2.48), i.e.

$$H_q^-(\mathcal{G}_q) = -\sum_{k=1}^{n} p_q^-(k) \log_2 p_q^-(k). \tag{4.29}$$

Similarly, the *simplicial out-q-degree distribution entropy* of $\mathcal{G}_q$ is defined as the simplicial analogue of the out-degree distribution entropy (2.49), i.e.

$$H_q^+(\mathcal{G}_q) = -\sum_{k=1}^{n} p_q^+(k) \log_2 p_q^+(k). \tag{4.30}$$

Roughly speaking, these entropies can be interpreted as measures of the "degree of higher-order disorder" (*in relation to* the inner or outer higher-order flux in the case of H^- or H^+, respectively) in the network for each level q. Also, similarly to the in/out-

degree distributions, H_q^- and H_q^+ reach their minimum when all directed simplices have the same in/out-q-degree, and reach their maximum when $p_q^-(k) = p_q^+(k) = 1/(|\mathcal{V}_q| - 1)$, for all k.

4.2.5 Forman-Ricci Curvature

In Riemannian geometry, there are several notions of curvature associated with Riemannian manifolds [81]. One of these curvatures is the Ricci curvature, which tries to quantify the "non-flatness" of a Riemannian manifold, or in which "degree" it deviates from being locally Euclidean [81, 85]. Robin Forman [102] was the first person to propose a discrete notion of Ricci curvature for cell complexes (known as *Forman-Ricci curvature*). Since then, a myriad of generalized Ricci curvatures for discrete structures have been proposed, such as discrete curvatures for graphs [228, 249], digraphs [248], hypergraphs [85, 166], and simplicial complexes [284].

Sreejith et al. [249] proposed a version of Forman's discrete version of Ricci curvature for (unweighted and weighted) undirected graphs, and it was later extended to (unweighted and weighted) directed graphs [248]. These generalized curvatures are edge-centric local measures of geometrical characterization that can help us gain insights into the network organization.

In what follows, we present the mathematical formalism of the Forman-Ricci curvature for directed graphs as it was introduced in [248]. Let $e = (v_1, v_2)$ be an arc of a digraph G, and let $\omega(e)$, $\omega(v_1)$, and $\omega(v_1)$ denote the weights associated with e, v_1, and v_2, respectively. The *Forman-Ricci curvature* of the arc e is defined as

$$F(e) = \omega(e)\left(\frac{\omega(v_1)}{\omega(e)} - \sum_{e_{v_1} \sim e} \frac{\omega(v_1)}{\sqrt{\omega(e)\omega(e_{v_1})}}\right) + \omega(e)\left(\frac{\omega(v_2)}{\omega(e)} - \sum_{e_{v_2} \sim e} \frac{\omega(v_2)}{\sqrt{\omega(e)\omega(e_{v_2})}}\right), \quad (4.31)$$

where $e_{v_1} \sim e$ represents the arcs whose heads coincide with v_1 (i.e., the incoming arcs at node v_1) and $e_{v_2} \sim e$ represents the arcs whose tails coincide with v_2 (i.e. the outgoing arcs at node v_2), excluding the arc e.

In view of this, we define the *q-Forman-Ricci curvature* for a q-arc $E = (\sigma_1, \sigma_2)$ in a q-digraph $\mathcal{G}_q$ as a straightforward extension of the formula (4.31), i.e.

$$F_q(E) = \omega(E)\left(\frac{\omega(\sigma_1)}{\omega(E)} - \sum_{E_{\sigma_1} \sim E} \frac{\omega(\sigma_1)}{\sqrt{\omega(E)\omega(E_{\sigma_1})}}\right) + \omega(E)\left(\frac{\omega(\sigma_2)}{\omega(E)} - \sum_{E_{\sigma_2} \sim E} \frac{\omega(\sigma_2)}{\sqrt{\omega(E)\omega(E_{\sigma_2})}}\right), \quad (4.32)$$

where $\omega(\sigma_1)$, $\omega(\sigma_2)$, and $\omega(\sigma)$ are the weights associated with σ_1, σ_1, and $\omega(E)$, respectively, and $E_{\sigma_1} \sim E$ represents the q-arcs whose heads coincide with σ_1 (i.e. the incoming q-arcs at node σ_1) and $E_{\sigma_2} \sim E$ represents the q-arcs whose tails coincide with σ_2 (i.e. the outgoing q-arcs at node σ_2), excluding the q-arc E.

Now we can construct a straightforward extension of the *in* and *out* Forman-Ricci curvatures of a node in a digraph, as proposed in [248], to a node in a q-digraph as follows. Let $E_{I,\sigma}$ and $E_{O,\sigma}$ denote the set of all incoming q-arcs of σ (i.e. the q-arcs whose heads coincide with σ) and the set of all outgoing q-arcs of σ (i.e. the q-arcs whose tails coincide with σ), respectively. The *in-q-Forman-Ricci curvature* and the *out-q-Forman-Ricci curvature* of a directed simplex σ are defined, respectively, by

$$F_q^-(\sigma) = \sum_{E \in E_{I,\sigma}} F_q(E), \tag{4.33}$$

$$F_q^+(\sigma) = \sum_{E \in E_{O,\sigma}} F_q(E). \tag{4.34}$$

In addition, we define the *total q-flow* through σ as

$$F_q(\sigma) = F_q^-(\sigma) + F_q^+(\sigma). \tag{4.35}$$

It's important to note that for a weighted q-digraph the weights associated with the nodes are given by the product-weight function (3.4), and the weights associated with the q-arcs are given by some non-symmetric node-to-edge function, such as the functions (3.51) and (3.52). On the other hand, for an unweighted q-digraph, there are several ways to assign weights to its nodes and q-arcs, for instance, we can identify the weight of a node with its in- or out-degree, and use the function (3.51) to transform the node weights into q-arc weights.

Example 4.2.2. Figure 4.2 presents a directed flag complex and its q-digraphs, for $q = 0, 1$. We computed the q-Forman-Ricci curvature for the q-arcs $E_{12} = (\sigma_1, \sigma_2)$ and $E_{23} = (\sigma_2, \sigma_3)$, and the in- and out-$q$-Forman-Ricci curvatures for the nodes σ_1, σ_2, and σ_3. The node weights were obtained through the function (3.3), where we considered the arc weights of the underlying digraph equal to 1, and the q-arc weights were obtained through the node-to-edge function (3.51).

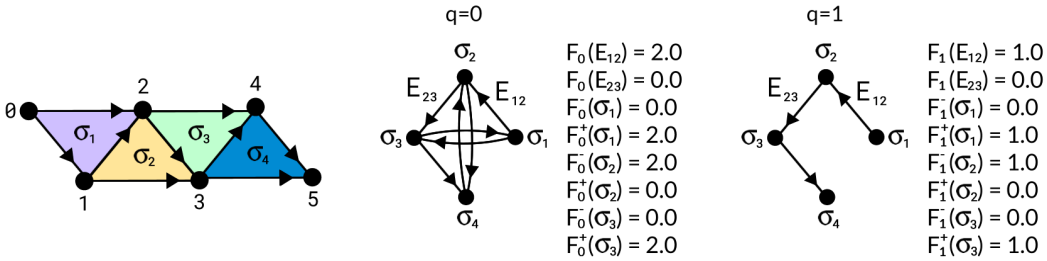


Figure 4.2: q-Forman-Ricci curvatures for q-arcs and in- and out-q-Forman-Ricci curvatures for nodes in the q-digraphs associated with the directed flag complex shown on the left side, for $q = 0, 1$.

Furthermore, from formula (4.32), we can infer that q-arcs connecting nodes with high in-q-degrees to nodes with high out-q-degrees have highly negative curvature values.

4.2.6 Spectrum-Related Simplicial Measures

In this last part, we extend the spectrum-based measures as defined for digraphs in Subsection 2.3.5 to directed flag complexes and some of them to path complexes. These new simplicial measures are related to the spectrum of the q-adjacency matrices or the spectrum of the Hodge Laplacian matrices, and they can be interpreted as measures that try to quantify the global "higher-order structures" of the network through its "higher-order spectra."

Simplicial Energy

Let $\{\varsigma_k^q\}_k$ be the *singular q-values* of $\mathcal{H}_q$ (i.e., the square roots of the eigenvalues of the matrix $\mathcal{H}_q^T\mathcal{H}_q$). We define the *simplicial q-energy* of $\mathcal{G}_q$ as the simplicial analogue of the graph energy (2.50), i.e. as the trace norm of the matrix $(\mathcal{H}_q\mathcal{H}_q^T)^{1/2}$:

$$\varepsilon_q(\mathcal{G}_q) = ||(\mathcal{H}_q\mathcal{H}_q^T)^{1/2}||_* = \mathrm{Tr}((\mathcal{H}_q\mathcal{H}_q^T)^{1/2}) = \mathrm{Tr}((\mathcal{H}_q^T\mathcal{H}_q)^{1/2}) = \sum_k \varsigma_k^q. \tag{4.36}$$

A different concept of energy of simplicial complexes related to their Euler characteristics was defined by O. Knill [161].

Directed Simplicial Katz Centralities

Let $\mathcal{H}_q \in \mathbb{R}^{n_q \times n_q}$. We define the *simplicial in-q-Katz centrality* of $\sigma \in \mathcal{V}_q$ as the simplicial analogue of the directed version of the Katz centrality which considers the incoming arcs (2.53), i.e.

$$K_q^-(\sigma) = \left[\langle 1|(I_{n_q} - \alpha\mathcal{H}_q)^{-1}\right]_\sigma, \tag{4.37}$$

and we define the *simplicial out-q-Katz centrality* of $\sigma \in \mathcal{V}_q$ as the simplicial analogue of the directed version of the Katz centrality which considers the outgoing arcs (2.54), i.e.

$$K_q^+(\sigma) = \left[(I_{n_q} - \alpha\mathcal{H}_q)^{-1}|1\rangle\right]_\sigma, \tag{4.38}$$

where $|1\rangle = (1, ..., 1)^T$ and I_{n_q} is the $n_q \times n_q$ identity matrix. The subscript σ in the brackets represents the position corresponding to σ in the vector inside the brackets.

The attenuation factor must be $\alpha \neq 1/\lambda_1^q$, where λ_1^q is the largest eigenvalue of $\mathcal{H}_q$. Analogously to the graph Katz centrality, typically the attenuation factor is chosen to be $\alpha < 1/\lambda_1^q$.

Simplicial Eigenvector Centralities

The *right simplicial q-eigenvector centrality* of $\sigma \in \mathcal{V}_q$ is defined as the simplicial analogue of the eigenvector centrality (2.55), i.e. as the entry corresponding to σ of the right eigenvector associated with the largest eigenvalue (λ_1^q) of $\mathcal{H}_q$:

$$C_{e,r}^q(\sigma) = \left(v_1^q\right)_\sigma = \left(\frac{1}{\lambda_1^q}\mathcal{H}_q v_1^q\right)_\sigma. \tag{4.39}$$

The *left simplicial q-eigenvector centrality* of σ is defined in an analogous way but considering the left eigenvector, i.e.

$$C_{e,l}^q(\sigma) = \left(v_1^q\right)_\sigma = \left(\frac{1}{\lambda_1^q}\mathcal{H}_q^T v_1^q\right)_\sigma. \tag{4.40}$$

Similarly as observed for the graph eigenvector centrality in Subsection 2.3.5, the Perron-Frobenius theorem (Theorem 2.2.1) guarantees that the right and left eigenvectors associated with λ_1^q are non-negative, thus the right and left simplicial q-eigenvalue centralities of every σ are non-negative.

Simplicial Spectral Entropy

Let $\mathcal{X}$ be a path complex or a directed flag complex associated with a digraph without double-edges. Let $\{\mu_i^n\}_i$ be the spectrum of the Hodge n-Laplacian matrix of $\mathcal{X}$, $[\mathcal{L}_n]$, with multiplicities, and let

$$p(\mu_i^n) = \frac{\mu_i^n}{\sum_i \mu_i^n} \tag{4.41}$$

be the *n-eigenvalue probabilities*, i.e. the contribution of μ_i^n in the Hodge n-Laplacian spectrum (Proposition 3.1.3 guarantees that $\mu_i^n \geq 0$, $\forall i$). Assuming the conventions $0/0 = 0$ and $0 \log_2 0 = 0$, we define the *simplicial n-spectral entropy* of $\mathcal{X}$ as the Shannon entropy of the n-eigenvalue probabilities, i.e.

$$S_n(\mathcal{X}) = -\sum_i p(\mu_i^n) \log_2 p(\mu_i^n). \tag{4.42}$$

Alternative definitions of spectral entropies associated with simplicial complexes were proposed by Maletić and Rajković [177] and Baccini et al. [23].

4.3 Simplicial Similarity Comparison Methods

In this section, we introduce several similarity comparison methods, such as *topological structure distances* and *simplicial kernels*, for digraph-based complexes based on their

algebraic/topological/structural properties, such as the number of weakly/strongly q-connected components, the number of directed simplices of certain dimensions, Betti numbers, etc. Also, we propose a *simplicial spectral distance* to compare complexes via their Hodge Laplacian spectra.

Throughout this part, all directed flag complexes are considered to be associated with simple digraphs *without double-edges*.

4.3.1 Topological Structure Vectors and Structure Distances

In Subsection 3.3.2 we introduced structure vectors associated with directed flag complexes, each capturing a specific structural property of the complex. Here we introduce additional structure vectors that take into account other topological features of these complexes, for instance, the Betti numbers $\beta_n = \dim H_n$ (which are associated with the topology of the network) and the lengths of the bars in the persistence barcodes, and we present specific structure vectors for path complexes; subsequently, we propose a general formula for similarity comparison between two digraph-based complexes based on these novel structure vectors.

Definition 4.3.1. Given a directed flag complex $\mathcal{X}$, with $\dim \mathcal{X} = N$, we define five *topological structure vectors* associated with it, namely:

1. The $1st$ *topological structure vector* is defined as $\mathrm{Str}_1(\mathcal{X}) = (s_0^1, ..., s_N^1)$, where s_i^1 is the number of directed i-simplices contained in $\mathcal{X}$, $i = 0, ..., N$, i.e. $\mathrm{Str}_1(\mathcal{X})$ is equal to the second structure vector as introduced in Definition 3.3.23.

2. The $2nd$ *topological structure vector* is defined as $\mathrm{Str}_2(\mathcal{X}) = (s_0^2, ..., s_N^2)$, where s_q^2 is the number of weakly q-connected components of $\mathcal{X}$, $q = 0, ..., N$, i.e. $\mathrm{Str}_2(\mathcal{X})$ is equal to the first weak structure vector as introduced in Definition 3.3.23.

3. The $3rd$ *topological structure vector* is defined as $\mathrm{Str}_3(\mathcal{X}) = (s_0^3, ..., s_N^3)$, where s_q^3 is the number of strongly q-connected components of $\mathcal{X}$, $q = 0, ..., N$, i.e. $\mathrm{Str}_3(\mathcal{X})$ is equal to the first strong structure vector as introduced in Definition 3.3.23.

4. The $4th$ *topological structure vector* is defined as $\mathrm{Str}_4(\mathcal{X}) = (s_0^4, ..., s_{max}^4)$, where $s_i^4 = \beta_i$ is the i-th Betti number of $\mathcal{X}$, $i = 0, ..., max$.

5. The $5th$ *topological structure vector* is defined as $\mathrm{Str}_5(\mathcal{X}) = (s_0^5, ..., s_{max}^5)$, where s_i^5 is the average bar length in the i-th persistence barcode of $\mathcal{X}$, $i = 0, ..., max$.

We could consider more structure vectors (such as the third weak/strong structure vector introduced in Definition 3.3.23), but here we will only deal with these five vectors.

In the case where $\mathcal{X}$ is path complex, since we did not define any kind of "higher-order connectivity" between its elementary p-paths, we do not have an equivalent for path complexes of the 2nd and 3rd topological structure vectors as defined above. Accordingly, we define different topological structure vectors for path complexes as follows.

Definition 4.3.2. Let $\mathcal{P}$ be a path complex and let's denote by N the length of its largest elementary p-path. We define five *topological structure vectors* associated with $\mathcal{P}$, namely:

1. The $1st$ *topological structure vector* is defined as $\mathrm{Str}_1(\mathcal{P}) = (s_0^1, ..., s_N^1)$, where s_i^1 is the number of elementary i-paths in $\mathcal{P}$, $i = 0, ..., N$.

2. The $2nd$ *topological structure vector* is defined as $\mathrm{Str}_2(\mathcal{P}) = (s_0^2, ..., s_N^2)$, where s_i^2 is the number of $(i, 1/i)$-DQCs associated with the elementary i-paths of $\mathcal{P}$, $i = 0, ..., N$.

3. The $3rd$ *topological structure vector* is defined as $\mathrm{Str}_3(\mathcal{P}) = (s_0^3, ..., s_N^3)$, where $s_i^3 = \dim \Omega_i$, $i = 0, ..., N$.

4. The $4th$ *topological structure vector* is defined as $\mathrm{Str}_4(\mathcal{P}) = (s_0^4, ..., s_{max}^4)$, where $s_i^4 = \beta_i$ is the i-th Betti number, i.e. the dimension of the i-th path homology of $\mathcal{P}$, $i = 0, ..., max$.

5. The $5th$ *topological structure vector* is defined as $\mathrm{Str}_5(\mathcal{P}) = (s_0^5, ..., s_{max}^5)$, where s_i^5 is the average bar length in the i-th persistence barcode of $\mathcal{P}$, $i = 0, ..., max$.

Due to the computational cost of finding ∂-invariant p-paths, in the 2nd topological structure vector only the elementary p-paths of $\mathcal{P}$ were considered.

In addition, note that if G is a digraph without double-edges and $\mathcal{X}$ and $\mathcal{P}$ are its directed flag complex and its path complex, respectively, the k-th entry of the vector $(\mathrm{Str}_2(\mathcal{P}) - \mathrm{Str}_1(\mathcal{X}))$ is equal to the number of ∂-invariant elementary k-paths that do not belong to any directed k-clique.

Remark 4.3.1. For a weighted directed flag complex, if one wishes to take the weights into account, it is enough to compute the premetric Dowker complex (see Definition 3.1.14).

We now propose a general formula for similarity comparison between two directed flag complexes, or between two path complexes, based on the topological structure vectors defined above.

Definition 4.3.3. Given two path complexes or two directed flag complexes, $\mathcal{X}_1$ and $\mathcal{X}_2$, let $\mathrm{Str}_n(\mathcal{X}_1) = (s_0^{1,n}, ..., s_{N_1^n}^{1,n})$ and $\mathrm{Str}_n(\mathcal{X}_2) = (s_0^{2,n}, ..., s_{N_2^n}^{2,n})$ be their respective n-th topological structure vectors. Without loss of generality, suppose $N_1^n \geq N_2^n$. Then consider the new vector $\mathrm{Str}_n^*(\mathcal{X}_2) = (s_0^{2,n}, ..., s_{N_1^n}^{2,n})$, such that $s_k^{2,n} = 0$ for all $N_2^n < k \leq N_1^n$. Let $||\cdot||_2$ denote the Euclidean norm (or 2-norm) in $\mathbb{R}^{N_1^n}$. We define the (normalized) *n-th topological structure distance* between $\mathcal{X}_1$ and $\mathcal{X}_2$ by

$$\widehat{T}_{tsd}^n(\mathcal{X}_1, \mathcal{X}_2) = \begin{cases} T_{tsd}^n(\mathcal{X}_1, \mathcal{X}_2), \text{ if } \mathcal{X}_1 \neq \emptyset \text{ or } \mathcal{X}_2 \neq \emptyset, \\ 0, \text{ if } \mathcal{X}_1 = \mathcal{X}_2 = \emptyset, \end{cases} \tag{4.43}$$

where

$$T_{tsd}^n(\mathcal{X}_1, \mathcal{X}_2) = \frac{||\mathrm{Str}_n(\mathcal{X}_1) - \mathrm{Str}_n^*(\mathcal{X}_2)||_2}{||\mathrm{Str}_n(\mathcal{X}_1)||_2 + ||\mathrm{Str}_n^*(\mathcal{X}_2)||_2}. \tag{4.44}$$

The vectors Str_n correspond to those defined in Definition 4.3.1 when $\mathcal{X}_1$ and $\mathcal{X}_2$ are directed flag complexes, and correspond to those defined in Definition 4.3.2 when $\mathcal{X}_1$ and $\mathcal{X}_2$ are path complexes.

It can be easily verified that $0 \leq \widehat{T}_{tsd}^n \leq 1$. Indeed, since all entries of the vectors Str_n are non-negative, by the triangular inequality of the Euclidean norm, we have $||\mathrm{Str}_n(\mathcal{X}_1) - \mathrm{Str}_n^*(\mathcal{X}_2)||_2 \leq ||\mathrm{Str}_n(\mathcal{X}_1)||_2 + ||\mathrm{Str}_n^*(\mathcal{X}_2)||_2$.

4.3.2 Simplicial Kernels

As we already discussed in Section 2.4, graph kernels are distance-based algorithms that produce a similarity score as the output of the comparison between two graphs. Martino et al. [180] proposed four kernels for simplicial complexes, namely: *histogram cosine kernel*, *weighted Jaccard kernel*, *edit kernel*, and *stratified edit kernel*. The first two kernels are based on the count of simplices belonging simultaneously to both simplicial complexes being compared, and the last two kernels are based on the counting of edit operations. In what follows, we propose adaptations of these four simplicial kernels to digraph-based complexes.

Definition 4.3.4. Given two path complexes or two directed flag complexes, $\mathcal{X}_1$ and $\mathcal{X}_2$, let $\mathrm{Str}_1(\mathcal{X}_1) = (s_0^{1,1}, ..., s_{N_1^1}^{1,1})$ and $\mathrm{Str}_1(\mathcal{X}_2) = (s_0^{2,1}, ..., s_{N_2^1}^{2,1})$ be their respective 1st topological structure vectors. Without loss of generality, suppose $N_1^1 \geq N_2^1$. Then consider the new vector $\mathrm{Str}_1^*(\mathcal{X}_2) = (s_0^{2,1}, ..., s_{N_1^1}^{2,1})$, such that $s_k^{2,1} = 0$ for all $N_2^1 < k \leq N_1^1$. Let $\langle \cdot, \cdot \rangle$ denote the Euclidean inner product in $\mathbb{R}^{N_1^1}$. The (normalized) *histogram cosine kernel* (HCK) is defined by

$$K_{HC}(\mathcal{X}_1, \mathcal{X}_2) = \frac{\langle \mathrm{Str}_1(\mathcal{X}_1), \mathrm{Str}_1^*(\mathcal{X}_2) \rangle}{\sqrt{\langle \mathrm{Str}_1(\mathcal{X}_1), \mathrm{Str}_1(\mathcal{X}_1) \rangle} \sqrt{\langle \mathrm{Str}_1^*(\mathcal{X}_2), \mathrm{Str}_1^*(\mathcal{X}_2) \rangle}}. \tag{4.45}$$

The 1st topological structure vectors correspond to those defined in Definition 4.3.1 when $\mathcal{X}_1$ and $\mathcal{X}_2$ are directed flag complexes, and correspond to those defined in Definition 4.3.2 when $\mathcal{X}_1$ and $\mathcal{X}_2$ are path complexes.

Definition 4.3.5. Given two path complexes or two directed flag complexes, $\mathcal{X}_1$ and $\mathcal{X}_2$, the *Jaccard kernel* is defined as

$$K_J(\mathcal{X}_1, \mathcal{X}_2) = \begin{cases} 1 - \frac{|\mathcal{X}_1 \cap \mathcal{X}_2|}{|\mathcal{X}_1 \cup \mathcal{X}_2|}, \text{ if } \mathcal{X}_1 \neq \emptyset \text{ or } \mathcal{X}_2 \neq \emptyset, \\ 0, \text{ if } \mathcal{X}_1 = \mathcal{X}_2 = \emptyset, \end{cases} \tag{4.46}$$

where $|\mathcal{X}_1 \cap \mathcal{X}_2|$ represents the cardinality of the intersection and $|\mathcal{X}_1 \cup \mathcal{X}_2|$ represents the cardinality of the union.

The Jaccard kernel is a normalized similarity distance. In fact, since $|\mathcal{X}_1 \cap \mathcal{X}_2| \leq |\mathcal{X}_1 \cup \mathcal{X}_2|$, we have $0 \leq K_J \leq 1$, and $K_J = 0$ when $\mathcal{X}_1 = \mathcal{X}_2$ and $K_J = 1$ when $\mathcal{X}_1 \cap \mathcal{X}_2 = \emptyset$.

Definition 4.3.6. Given two path complexes or two directed flag complexes, $\mathcal{X}_1$ and $\mathcal{X}_2$, let $e(\mathcal{X}_1, \mathcal{X}_2)$ be an edit distance between them (i.e. a distance based on the number of changes necessary to convert one into the other - see Section 2.4). We define the (normalized) *edit kernel* as

$$K_E(\mathcal{X}_1, \mathcal{X}_2) = \begin{cases} 1 - \bar{e}(\mathcal{X}_1, \mathcal{X}_2), \text{ if } \mathcal{X}_1 \neq \emptyset \text{ or } \mathcal{X}_2 \neq \emptyset, \\ 0, \text{ if } \mathcal{X}_1 = \mathcal{X}_2 = \emptyset, \end{cases} \tag{4.47}$$

where

$$\bar{e}(\mathcal{X}_1, \mathcal{X}_2) = \frac{2e(\mathcal{X}_1, \mathcal{X}_2)}{|\mathcal{X}_1| + |\mathcal{X}_2| + e(\mathcal{X}_1, \mathcal{X}_2)}. \tag{4.48}$$

Definition 4.3.7. Given two directed flag complexes (or two path complexes), $\mathcal{X}_1$ and $\mathcal{X}_2$, let $\mathcal{D}$ denote the set of all different dimensions (or lengths) of the simplices (or elementary paths) present in these two complexes. The *stratified edit kernel* (SEK) is defined as

$$K_{SE}(\mathcal{X}_1, \mathcal{X}_2) = \frac{1}{|\mathcal{D}|} \sum_{k \in \mathcal{D}} K_E(\mathcal{X}_1^k, \mathcal{X}_2^k), \tag{4.49}$$

where $\mathcal{X}_1^k \subseteq \mathcal{X}_1$ and $\mathcal{X}_2^k \subseteq \mathcal{X}_2$ are the subsets of all directed k-simplices (or elementary k-paths) in the respective complexes, and K_E is the edit kernel (4.47).

4.3.3 Simplicial Spectral Distance

The distances between two digraph-based complexes defined in the previous two sections are mainly based on the structural properties of these complexes. Now we present a novel distance based on their Hodge Laplacian spectra, i.e. a *higher-order spectral distance*, which is also connected to their topological structures via their spectra (see Subsection 3.1.6).

Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two directed flag complexes (or two path complexes) with Hodge n-Laplacian matrices $[\mathcal{L}_n^1]$ and $[\mathcal{L}_n^2]$, respectively. Let $\mu_0^n(\mathcal{X}_1) \leq ... \leq \mu_{N_1}^n(\mathcal{X}_1)$ and $\mu_0^n(\mathcal{X}_2) \leq ... \leq \mu_{N_2}^n(\mathcal{X}_2)$ be the spectra of $[\mathcal{L}_n^1]$ and $[\mathcal{L}_n^2]$, respectively. Without loss of generality, suppose $N_1 \leq N_2$. The *simplicial n-spectral distance* between $\mathcal{X}_1$ and $\mathcal{X}_1$ is defined as

$$\widehat{D}_{\mathcal{L}_n}(\mathcal{X}_1, \mathcal{X}_2) = \begin{cases} D_{\mathcal{L}_n}(\mathcal{X}_1, \mathcal{X}_2), \text{ if } [\mathcal{L}_n^1] \neq 0 \text{ or } [\mathcal{L}_n^2] \neq 0, \\ 0, \text{ if } [\mathcal{L}_n^1] = [\mathcal{L}_n^2] = 0, \end{cases} \tag{4.50}$$

where

$$D_{\mathcal{L}_n}(\mathcal{X}_1, \mathcal{X}_2) = \frac{1}{S_n(\mathcal{X}_1, \mathcal{X}_2)} \left(\sum_{k=0}^{N_1} |\mu_k^n(\mathcal{X}_1) - \mu_k^n(\mathcal{X}_2)| + \sum_{k=N_1+1}^{N_2} |\mu_k^n(\mathcal{X}_2)| \right), \tag{4.51}$$

with the normalization term

$$S_n(\mathcal{X}_1, \mathcal{X}_2) = \sum_{k=0}^{N_1} |\mu_k^n(\mathcal{X}_1)| + \sum_{k=0}^{N_2} |\mu_k^n(\mathcal{X}_2)|. \tag{4.52}$$

It can be easily verified that $0 \leq \widehat{D}_n \leq 1$. Indeed, by Proposition 3.1.3, for all n, all eigenvalues of the Hodge n-Laplacian matrices are non-negative, thus $|\mu_k^n(\mathcal{X}_1) - \mu_k^n(\mathcal{X}_2)| \leq |\mu_k^n(\mathcal{X}_1)| + |\mu_k^n(\mathcal{X}_2)|$, for all k.

Moreover, notice that the distance (4.50) compare solely the spectra of the Hodge n-Laplacians of $\mathcal{X}_1$ and $\mathcal{X}_2$ for a same order n.

4.4 Examples with Random Digraphs

In this last part, we present some examples of applications of the simplicial characterization measures and simplicial similarity comparison distances to random digraphs built through the four random digraph models introduced in Section 2.5.

Example 4.4.1. Figure 4.3 presents four sparse (density ≤ 0.5) random digraphs of same order $n = 20$ obtained through four different random models, namely: k-regular (KR) model with parameter $k = 6$; Erdős-Rényi $G(n, p)$ (ER) model with parameter $p =$

0.2; Watts-Strogatz (WS) model with parameters $k = 10$ and $p = 0.2$; and Barabási-Albert (BA) model with parameters $\alpha = 0.41, \beta = 0.54, \gamma = 0.05, \delta_{in} = 0.2$, and $\delta_{out} = 0.0$. Also, Figure 4.3 depicts the respective q-digraphs associated with these four digraphs, for $q = 0, 1, 2$. As a convention, let's refer to the original digraphs as (-1)-digraphs, i.e. we will use $q = -1$ to designate the original networks.

We applied eighteen simplicial characterization measures (described in Table B.1)[1] to these four random digraphs and their respective q-digraphs, $q = 0, 1, 2$, whose results are presented in Table 4.2, and we performed a pairwise comparison between the original digraphs using nine different simplicial similarity comparison distances (described in Table B.2)[2], whose results are presented in Table 4.3. To compute the measures, since some of them depend on the order (N) of the network (i.e., are *N-dependent* [266]), we used a fixed N equal to the order of the largest network among all levels q.

The idea of this example is to present a simple numerical analysis of the results just to get a sense of how the simplicial measures behave at each level q for digraphs with different topologies, and not to perform a rigorous statistical analysis. With that said, inspecting the results, we obtained some interesting insights:

- Distance-based simplicial measures: The BA network presents the lowest values for the global q-efficiency (measure 2) and for the average shortest q-walk length (measure 1) in relation to other networks at all levels q, which might suggest that information is propagated less efficiently in this network than in the other networks, including at higher-order levels of network topological organization.

- Simplicial centrality measures: The in- and out-q-degree centralities (measures 4 and 5, respectively), as expected, are greater at the level $q = 0$ for all networks and become smaller as we increase the level q since the q-digraphs become more disconnected (sparse). However, the global q-reaching centrality (measure 8) increases for some networks, which might suggest that higher-order structures have a greater influence on the information flow.

- Simplicial segregation measures: The ER and BA networks have high in-q-rich-club coefficients (measure 10) in the original networks which might suggest the presence of densely connected nodes with high in-degrees, however, this property is not preserved at other levels q. Also, the out-q-rich-club coefficients (measure 11) are equal to zero for all the original networks and at level $q = 2$, but they are greater than zero at levels $q = 0$ and $q = 1$.

- Simplicial similarity comparison distances: For most of the distances, the pairs KR-BA, ER-BA, and WS-BA produced values closest to 1, which might suggest

[1] We used the parameter $k = 6$ for the simplicial in- and out-q-degree rich-club coefficients.

[2] For the bottleneck, Wasserstein, Betti, 4th, and 5th topological structure distances we only considered the 0-th Betti numbers. For the simplicial n-spectral distance, we considered $n = 1$.

that these pairs have the most different topologies compared to each other. However, most distances produced values close to 0 for the pairs KR-WS and ER-WS, which might suggest that their topologies are similar.

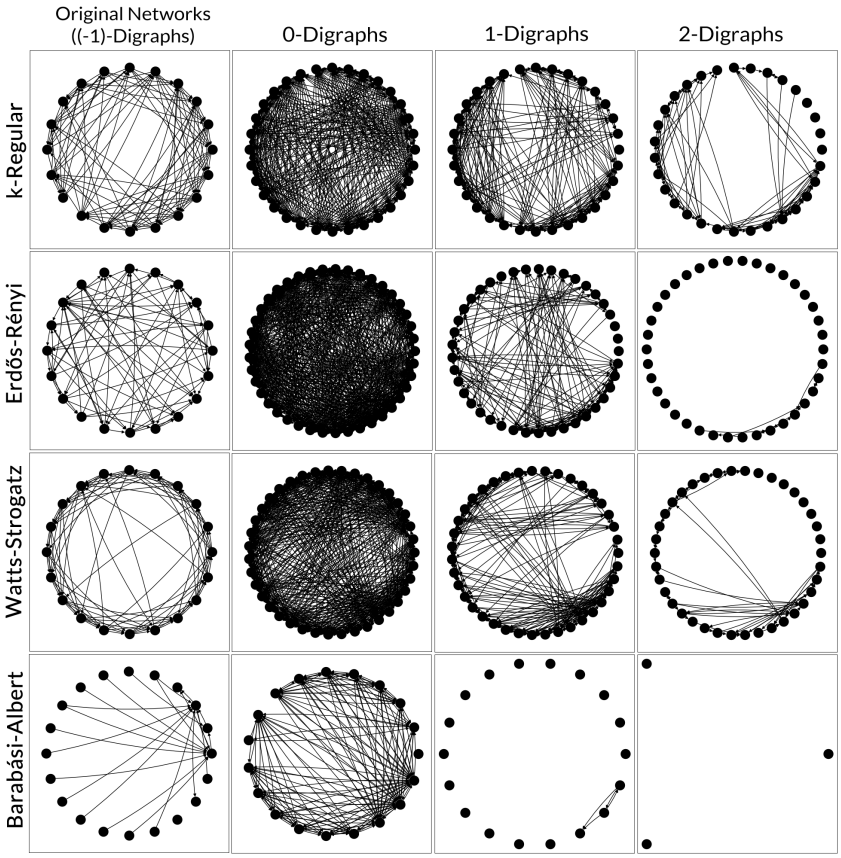


Figure 4.3: Random digraphs and their respective q-digraphs ($q = 0, 1, 2$).

Table 4.2: Results of the simplicial characterization measures for the random digraph models at each level $q = -1, 0, 1, 2$. See Table B.1 for the measure ids.

	$q=-1$				$q=0$				$q=1$				$q=2$			
Measure	**KR**	**ER**	**WS**	**BA**	**KR**	**ER**	**WS**	**BA**	**KR**	**ER**	**WS**	**BA**	**KR**	**ER**	**WS**	**BA**
1	2.03	1.66	2.22	0.49	1.57	1.55	1.58	1.54	2.35	3.12	2.6	0.67	1.94	1.0	6.06	0.0
2	0.09	0.07	0.08	0.02	0.29	0.49	0.43	0.09	0.21	0.23	0.27	0.0	0.08	0.01	0.08	0.0
3	0.0	0.0	0.01	0.01	0.2	0.26	0.18	0.24	0.16	0.31	0.13	0.21	0.1	0.19	0.12	0.0
4*	0.12	0.16	0.14	0.18	0.37	0.57	0.45	0.33	0.24	0.2	0.22	0.04	0.14	0.06	0.08	0.0
5*	0.12	0.18	0.1	0.06	0.37	0.55	0.43	0.33	0.24	0.16	0.18	0.04	0.1	0.04	0.06	0.0
6*	12.3	12.5	12.6	12.3	24.5	34.0	30.0	16.5	19.1	21.1	22.5	2.0	9.83	4.0	9.02	0.0
7*	53.3	66.9	45.7	36.0	50.6	54.6	44.7	48.9	251	301	157	0.0	169	6.0	348	0.0
8	0.24	0.27	0.24	0.07	0.23	0.15	0.17	0.23	0.24	0.2	0.18	0.04	0.36	0.07	0.27	0.0
9	0.12	0.09	0.1	0.03	0.36	0.5	0.4	0.29	0.34	0.23	0.25	0.04	0.17	0.0	0.06	0.0
10	0.0	12.83	0.0	11.0	0.44	0.48	0.43	0.58	0.67	2.84	1.06	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.44	0.45	0.43	0.75	0.55	13.25	1.68	0.0	0.0	0.0	0.0	0.0
12	4.28	4.3	4.45	5.11	4.95	5.2	5.26	4.07	4.75	4.9	5.11	5.49	4.52	5.46	5.11	0.0
13	1.51	2.01	1.8	0.66	2.9	3.77	3.17	1.82	2.69	2.94	3.2	0.38	2.03	0.78	1.96	0.0
14	1.21	1.73	1.46	1.26	2.78	3.97	3.09	1.54	2.76	2.76	2.81	0.24	2.15	0.82	1.87	0.0
15	-360	-408	-258	-381	-126	-782	-759	-60	-103	-28	-74	0.0	-69	0.0	-15	0.0
16	-360	-785	-298	-471	-139	-790	-750	-63	-134	-25	-87	0.0	-132	0.0	-22	0.0
17	33.4	30.2	33.0	9.23	78.7	106	104	33.3	61.6	63.1	71.3	2.73	37.2	8.49	34.6	0.0
18*	2.41	2.25	2.18	1.98	1.65	1.67	1.66	29.3	8.76	3.65	4.59	1.22	2.15	1.33	1.51	0.0

* The results correspond to the maximum values obtained among all nodes.

Table 4.3: Results of pairwise comparisons of the random digraphs via the simplicial distances. See Table B.2 for the distance ids.

Distance	KR-ER	KR-WS	KR-BA	ER-WS	ER-BA	WS-BA
1	0.0	0.0	0.5	0.0	0.5	0.5
2	0.0	0.0	0.707	0.0	0.707	0.707
3	0.0	0.0	1.0	0.0	1.0	1.0
4	0.559	0.362	0.895	0.254	0.677	0.794
5	0.0	0.0	1.0	0.0	1.0	1.0
6	0.0	0.0	0.0	0.0	0.0	0.0
7	0.762	0.898	0.393	0.953	0.741	0.582
8	0.972	0.981	0.988	0.939	0.982	0.986
9	0.432	0.299	0.874	0.408	0.740	0.845

Example 4.4.2. In this example, we performed a statistical analysis on four global simplicial characterization measures (namely, average shortest directed simplicial q-walk length (measure 1)), directed simplicial global simplicial q-efficiency (measure 2), simplicial global q-reaching centrality (measure 8), and average directed simplicial q-clustering coefficient (measure 9) (measure ids are in accordance with Table B.1)) computed on q-digraphs, for $q = -1, 0, 1, 2$, obtained from samples produced by Monte Carlo simulations in the four random digraph models described in Section 2.5 to investigate whether the properties associated with the network topology of each model persist at each level q. Due to the limitations of computation time and processing power, we limited ourselves to small and sparse networks. The number of nodes was set to $n = 20$ and, to avoid dense digraphs, we limited the parameters in such a way that the digraph densities were always ≤ 0.5. We ran a Monte Carlo simulation on the parameters of each digraph model using uniform distributions:

- k-Regular (KR): $k \sim U(1, 10)$ (integer part);
- Erdős-Rényi $G(n, N)$ (ER): $N \sim U(0, 190)$ (integer part);
- Watts-Strogatz (WS): $k = 10$, $p \sim U(0, 1)$;
- Barabási-Albert (BA): $\delta_{in} \sim U(0, 1)$, $\delta_{out} = 0; \alpha = 0.41, \beta = 0.54, \gamma = 0.05$.

The mean of each measure was obtained from 50 simulations carried out for each model, repeated 20 times. Since some of the measures are N-dependent, we used a fixed N equal to the order of the largest network among all q. Figure 4.4 presents the means and standard deviations obtained for each measure. Subsequently, we compared the means obtained for each random model, for each level $q = -1, 0, 1, 2$, and for each measure, using ANOVA. We found statistically significant differences ($p < 0.05$) in all four measures at all levels q, except for measure 1 at level $q = 0$ ($F = 0.1513, p = 0.135$). Also, as a post hoc test, we performed a Bonferroni's multiple comparison test and we found statistically significant differences ($p < 0.05$) for all pairs of models, for all

measures, at all levels q, except for the pair ER-KR in measure 1 ($q = 1$: $p = 0.069$; $q = 2$: $p = 0.778$), measure 8 ($q = -1$: $p = 0.596$, $q = 2$: $p = 0.151$), and measure 9 ($q = -1$: $p = 0.416$; $q = 1$: $p = 0.052$; $q = 2$: $p = 0.324$), and for the pair KR-WS in measure 1 ($q = -1$: $p = 0.361$) and measure 2 ($q = 0$: $p = 0.542$).

Moreover, we notice that the ER networks presented a global q-reaching centrality significantly larger than the other network models at levels $q = 0, 1, 2$, which might suggest that the higher-order structures of ER networks propagated information more efficiently than the higher-order structures of the other networks. Also, for measures 8 and 9, it was not possible to find any significant difference between the KR and ER networks in the original networks ($q = -1$), as well as between the KR and WS networks in measures 1 and 2, but it was possible to find significant differences between these networks at other higher-order levels q.

We may conclude that the differences between the global topological properties for each random model are, in general, preserved in higher-order structural and connectivity levels, but may differ in different ways at each level. Furthermore, we saw that for some measures it is not possible to identify topological/structural, and/or functional differences between one network model and another, but these differences may be evident at higher-order levels of topological organization (clique organization), and this is one of the advantages of taking the q-digraphs associated with the networks into account in the analysis.

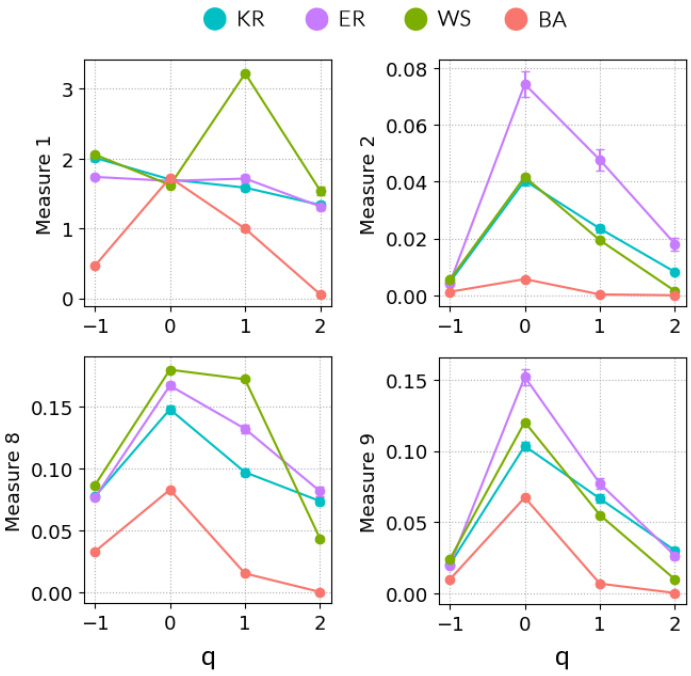


Figure 4.4: Means and standard deviations of the four simplicial measures computed for the q-digraphs, $q = -1, 0, 1, 2$, corresponding to each random digraph model. See Table B.1 for the measure ids.

Part II

Brain Connectivity Networks and a Quantitative Graph/Simplicial Analysis of Epileptic Brain Networks

Chapter 5

Brain Connectivity Networks

> *Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary following William of Occam he should seek an economical description of natural phenomena.*
>
> — George E. P. Box [42]

It is a well-known fact that brain activities depend on several different brain areas rather than being isolated occurrences in one or more distinct brain regions. Neuroanatomical structures located in different regions interact to process information, thus creating structural and functional brain networks. The signals from these neurophysiological activities can be captured by techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), which can then be subjected to mathematical and statistical methods, such as partial directed coherence (PDC), which can estimate the causal relationship between two regions by characterizing the influence that one exerts on the other through the concept of Granger causality; these connections can be represented by directed graphs, and thus can be analyzed by tools from graph theory and computational algebraic topology.

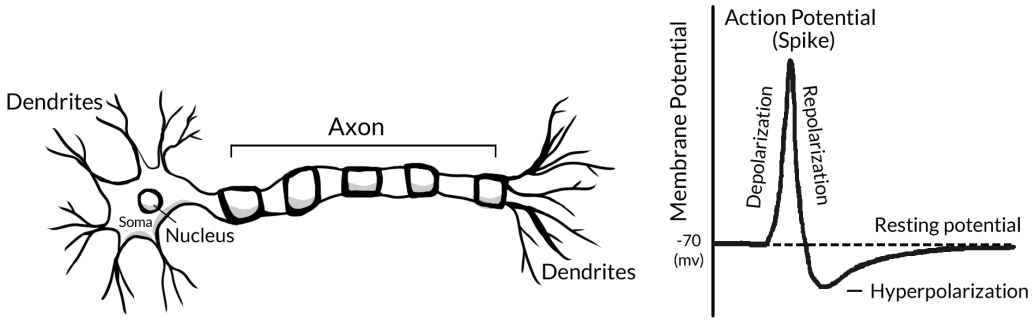
In this chapter, we discuss the general aspects of brain connectivity networks, starting with the biophysical principles of brain signals, going through the common methods for acquiring these signals, especially EEG, and then discussing brain connectivity estimators, with special attention to PDC and its variants. Afterwards, we discuss about the different types of brain connectivity and briefly about the applications of graph theoretical analysis (GTA) and topological data analysis (TDA) in modern network neuroscience research.

5.1 Biophysical Principles of Brain Signals

The human brain is part of a larger system called the *central nervous system* (CNS). The *nervous system* consists of the *peripheral nervous system* (PNS) and the CNS, and its basic units are the *neurons* [154]. Neurons are specialized cells in the nervous system that are responsible for transmitting information. Each neuron is composed of a cell body (soma) and a nucleus, dendrites, and an axon, as is schematically represented in Figure 5.1a. To be more precise, Figure 5.1a presents just a general schematic representation of a neuron, since neurons may vary substantially in size and shape, and they may be classified according to their specific morphology. For instance, some neurons have axons that are only a fraction of millimeters long and others can have axons that are many centimeters long; some neurons known as *pyramidal neurons* (discovered by Santiago Ramón y Cajal (1852 - 1934)) have a pyramid-shaped cell body, and they form the most common class of neurons in the neocortex, accounting for approximately 75-90% of all neurons [262].

The transmission of information between neurons occurs through connections called *synapses*. The neurons that receive inputs from other neurons are called *postsynaptic* neurons and the neurons that give inputs to other neurons are called *presynaptic* neurons, i.e. postsynaptic neurons receive inputs through synapses from presynaptic neurons. It is clear from the previous passage that synaptic connections are not symmetrical. Moreover, synapses can be electrical or chemical. Electrical synapses allow the direct flow of ionic currents from one neuron to another. Chemical synapses, however, involve the release of *neurotransmitters* from the synaptic vesicles of the presynaptic neuron into the *synaptic cleft* (the small gap between the neurons). These neurotransmitters then bind to receptors on the postsynaptic neuron, producing changes in the membrane potential of this neuron (known as *postsynaptic potentials* (PSPs)), leading to the initiation (in which case the PSPs are known as *excitatory* PSPs) or inhibition (in which case the PSPs are known as *inhibitory* PSPs) of a new action potential [70, 154].

An *action potential* (or *spike*) is an electrical impulse that travels down the axon of a neuron. When a neuron emits an action potential, it is said to be "firing." The *resting potential* is the potential inside the neuron when it is not firing, and it is approximately -70mV. An action potential is initiated when the membrane potential of the neuron reaches a certain threshold, causing an influx of sodium ions (Na^+) into the cell. This influx of positively charged sodium ions results in a rapid increase of the membrane potential (depolarization), generating an action potential, followed by a rapid decrease (repolarization) of this potential, reaching values lower than the resting potential for a brief period of time (hyperpolarization), and then returning to the resting potential (see Figure 5.1b) [262]. Once the action potential reaches the end of the axon, it triggers the release of neurotransmitters into the synaptic cleft, which bind to receptors on the postsynaptic neuron and lead to the transmission of information [70].



(a) Neuron (schematic representation). (b) Action potential.

Figure 5.1: Graphical representation of a neuron and an action potential.

In the previous paragraphs, we discussed very briefly the biophysical bases of brain activity at the cellular level. However, the brain has a multiscale organization (multiple levels of organization), from molecules to synapses, cells, neuronal networks, and brain areas [59, 112]. This is an important feature of the brain because, although there are (invasive) techniques to detect signals from just a few neurons, we are often interested in the functioning of the brain at a global level. We can study brain functions by utilizing the signals produced by the activity of populations of neurons, which can be recorded by non-invasive devices. The acquisition of brain signals is a fundamental aspect of neuroscience research and clinical diagnosis.

Numerous methods have been developed to measure the neurophysiological activity of the brain, all of which rely on specific signals resulting from the dynamics of brain processes, for example: *electroencephalography* (EEG) involves placing electrodes on the scalp to detect electrical activity [231]; *magnetoencephalography* (MEG) measures the magnetic fields produced by brain's electrical activity, and it requires highly sensitive devices called SQUIDs (superconducting quantum interference devices) [199]; *functional magnetic resonance imaging* (fMRI) detects changes in blood flow and oxygenation levels in the brain [145]; *positron emission tomography* (PET) is a nuclear imaging technique that requires the injection of a radioactive tracer into the bloodstream for subsequent measurement of the tracer's distribution and concentration in the brain [141].

5.2 Electroencephalography

Electroencephalography (EEG) is a non-invasive and economically feasible technique that allows the recording of the brain's electrical activity through electrodes attached to the scalp. In a 1929 report, titled *On the Electroencephalogram of Man*, the German neuropsychiatrist Hans Berger (1873–1941) described for the first time a successful observation of a human EEG, and was later considered the father of EEG [170, 231].

The electrical signals captured by the electrodes are generated by large populations of neurons, since only these large populations are able to generate electric potentials strong enough to be captured by scalp EEG because of the distance between the scalp electrodes and the neurons as well as the high electrical resistance of the medium between them (e.g., skin, skull, dura mater, and cerebral spinal fluid). It is thought that the extracellular potentials captured by EEG are produced by postsynaptic potentials (PSPs) of pyramidal neurons [231].

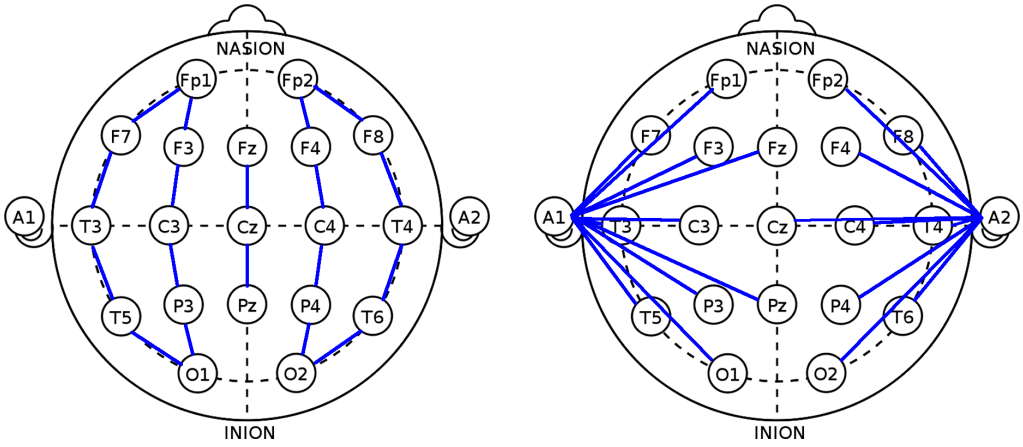
The effects caused by the recording of electrical potentials at a distance from their sources, as occurs in EEG recordings, are known as *volume conduction* [224]. Generally speaking, a medium, for example, skin, skull, dura mater, and cerebrospinal fluid, will occupy the space between an electrode placed on the scalp and a cerebral source of electrical potential. Electrical signals conducted across this medium result in volume conduction, since the electrical impulses dissipate and refract as a result of this medium conducting them. Volume conduction effects are present in all EEG recordings [264], and they might affect the resulting EEG signals.

The standard system for the arrangement of electrodes on the head is known as *international 10-20 system*[1]. It is implemented by positioning electrodes on the scalp using standard coordinates that are determined relatively to three reference points on the skull, namely: the *nasion* (the deepest point between the nose and the forehead), the *inion* (the lowest posterior point of the skull above the neck), and the *pre-auricular points* (commonly denoted by A1 and A2). These electrodes are spaced in ratios of 10% or 20% of the length of the distances between these three reference points. The electrode locations are indicated by a combination of a *letter* and a *number*. The letter indicates the corresponding cortical area (Fp: pre-frontal, F: frontal, T: temporal, P: parietal, O: occipital, C: central; the letter "z" is used to indicate the electrodes in the center of the head (Fz, Cz, and Pz)). The number indicates the relative position; odd and even numbers correspond to the left and right hemispheres, respectively [112, 170]. Figure 5.2 presents the configuration of 21 electrodes corresponding to the international 10-20 system.

EEG recordings are differential recordings, in which the difference between the voltages in two electrodes is measured by a differential amplifier. In other words, the differential amplifier receives the voltages from two electrodes and returns the difference between them, highly amplified. The patterns of chosen electrode pairs (channels) can vary, and these variations are called *montages*. Each signal is the amplified difference between the voltages of two electrodes, for example, in a longitudinal bipolar montage, we have the pairs T5-O1, Fz-Cz, etc. (see Figure 5.2a), and in a referential (or monopolar) montage, e.g., in an ipsilateral ear montage (see Figure 5.2b), we have the pairs T5-Ref, Fz-Ref, etc., where "Ref" is the electrode of reference (e.g., A1 or

[1]There are alternative systems to the 10-20 system, such as the 10-10 system, that can accommodate extra electrodes for a more detailed EEG.

A2).



(a) Longitudinal bipolar montage. (b) Ipsilateral ear reference montage.

Figure 5.2: International 10-20 electrode system, with 21 electrodes.

The frequencies of EEG signals can be divided into five frequency bands: delta [0.1 Hz, 4 Hz), theta [4 Hz, 8 Hz), alpha [8 Hz, 14 Hz), beta [14 Hz, 30 Hz), and gamma ($\geq$ 30 Hz), where [,) denotes semi-closed intervals, i.e. we do not include the highest frequency of each interval. In the literature, we can find other slightly different ranges for these frequency bands [52, 60, 214, 231], but here we will consider this convention. Typically, the clinically significant EEG frequency components have a range of 0.1 to 100 Hz [231].

In the previous paragraphs, we described the non-invasive EEG modality known as *scalp EEG*. However, two invasive EEG modalities that require surgically implanted intracranial electrodes are the *electrocorticography* (ECoG) and the *stereoelectroencephalography* (sEEG) [202]. Moreover, scalp EEG has high temporal resolution, which allows the measurement of electrical activities that occur in the order of milliseconds; however, it has low spatial resolution, while ECoG and sEEG have high temporal and spatial resolutions.

In clinical practice, it is common to use Video-EEG, which is the combination of EEG with video recording to allow simultaneous monitoring of brain electrical activity and the patient's behavior.

5.2.1 Stationarity of Signals

When we are analyzing real-world signals, such as electrophysiological signals, a concept that must be taken into consideration is the concept of *stationarity*.

Stationarity [179] refers to a property of a signal that remains constant over time. A stationary (wide-sense stationary) signal has statistical properties that do not change with time, such as its *mean* and *autocorrelation*. This means that the behavior of the signal is predictable and can be analyzed using statistical tools. For instance, if a signal

is stationary, its power spectral density can be estimated using the autocorrelation function, which provides a measure of the signal's frequency content.

In practice, however, most of the signals are non-stationary. In particular, EEG signals are essentially non-stationary due to the time-varying nature of the underlying brain processes [160, 269]. One way to tackle the problem of non-stationarity is to consider short segments of the signal as locally stationary (short-window methods), thus the methods that require stationarity can be applied from segment to segment.

5.2.2 EEG Artifacts

An *artifact* in an EEG is a component of the signal that is not a brain signal, i.e., it is a component of the signal that does not correspond to a true cerebral activity [170]. EEG artifacts may be caused by electrical signals originating from the patient's extracerebral physiological activities (*physiological artifacts*) or by electrical signals originating from external sources, such as equipment and the environment (*non-physiological artifacts*); common examples of these two types of artifacts are:

- **Physiological artifacts:** Eye blinks, eye movements, tongue movement (glossokinetic artifact), cardiac activity, and muscle contractions.
- **Non-physiological artifacts:** Electromagnetic interference from AC power lines (line frequencies typically are 50 Hz or 60 Hz), electrode detachment, EEG cable movement, and interference from other external electronic equipment.

The presence of artifacts in the EEG data may reduce the statistical power of the analysis; therefore, it is essential in the EEG preprocessing stage, before the analysis *per se*, to identify and remove or attenuate artifacts. There are several techniques used for artifact handling, for instance, low-frequency filters (to remove very low frequencies, e.g., frequencies < 1 Hz), high-frequency filters (to remove high frequencies, e.g., frequencies > 70 Hz), notch filters (to eliminate signals at a specific frequency, e.g., 50 Hz or 60 Hz frequencies), regression methods, wavelet transform, principal component analysis (PCA), and independent component analysis (ICA) (to remove eye blinks, eye movements, and muscle artifacts) [149].

5.2.3 Clinical Uses and Limitations of EEG and Video-EEG

As already discussed previously, EEG is a tool to assess the brain's electrical activity, and Video-EEG combines EEG with video recording to monitor both the brain's electrical activity and the patient's behavior. Despite the diverse clinical uses, in what follows, we focus on describing some important uses and limitations of EEG and Video-EEG in the diagnosis and treatment of epilepsy [189]:

- **Diagnosing epilepsy:** EEG is commonly used to diagnose epilepsy and differentiate it from other neuronal disorders. Video-EEG can provide additional information, such as the timing and duration of seizures, which can help in the diagnosis and treatment of epilepsy.
- **Monitoring seizures:** EEG and Video-EEG can be used to monitor seizures and track their frequency, duration, and severity. This information can be used to adjust medication dosages and evaluate the effectiveness of treatment, or as presurgical evaluation for patients with drug-resistant epilepsy (DRE).

Nevertheless, there are some limitations in the use of EEG and Video-EEG, such as:

- **Limited spatial resolution and sensitivity:** EEG provides a measure of overall brain activity but has limited spatial resolution. It cannot identify the precise location of epileptiform activity occurring far from the cranial vault, i.e. epileptiform activity originating from deep brain structures.
- **False-positives:** EEG can produce false-positive results, indicating the presence of epileptiform activity when none is present.
- **Inconvenience:** EEG and Video-EEG require the placement of electrodes on the scalp, which can be uncomfortable and time-consuming for patients.

EEG and Video-EEG are useful tools in clinical practice, especially in the diagnosis and treatment of epilepsy, but they have some limitations that must be considered when interpreting results.

5.3 From Brain Signals to Brain Connectivity

Brain signals captured by brain mapping modalities, such as fMRI, and electrophysiological methods, such as EEG, provide information about the complex and dynamic activities of the brain. These activities involve the interaction of several brain areas, connected by both anatomical and functional associations [240]. Brain connectivity comprises the description of how different brain structures interact and influence or are influenced by each other over time [108, 109], encompassing *structural connectivity* (anatomical connections), *functional connectivity* (statistical dependencies), and *effective connectivity* (causal influences) [239].

The connections and interactions between different brain areas can be measured and interpreted through connectivity estimators. These estimators are mathematical and statistical methods that can be computed from data obtained from fMRI, EEG, MEG,

etc. Different approaches can be used, consequently, we can have directed and non-directed, bivariate and multivariate, linear and non-linear, time-domain and frequency-domain connectivity estimators [53].

The most common connectivity estimators are correlation and coherence. They are examples of linear, bivariate estimators. Examples of non-linear, bivariate approaches include mutual information, transfer entropy, and phase synchronization [206]. All these measures can be used to infer functional connectivity. Several estimators such as transfer entropy, Granger causality index (GCI) [43, 116], directed coherence (DC) [225], partial directed coherence (PDC) [17], and directed transfer function (DTF) [153], are based on the concept of Granger causality, which is a cause-effect relation concept (the past values of one time series can predict current values of another), suggesting a directional influence or causality between brain regions, and thus they can be applied to infer effective connectivity. GCI, PDC, and DTF are directed, model-based, multivariate estimators, and whereas GCI is defined in the time-domain, DTF and PDC are defined in the frequency-domain. Moreover, information-theoretic forms of PDC and DTF, the information PDC (iPDC) and the information DTF (iDTF), which reinterpreted these estimators in terms of the mutual information, were introduced in [259, 261]. Another very popular model-based estimator used to infer causal interactions (but not based on Granger causality), and therefore used to infer effective connectivity, is the dynamic causal modeling (DCM) [110], which is a Bayesian model-based approach formulated in terms of stochastic differential equations or ordinary differential equations. As is reasonable to expect, all of the previous estimators have limitations and are influenced by the quality of the data, data variability, volume conduction effects, and data preprocessing steps [53].

As already discussed in Chapter 1, in this thesis we are interested in effective connectivity networks obtained through the iPDC estimator, thus in the next sections we will introduce all the mathematical formalism behind the PDC and its variants, beginning with the theory of vector autoregressive models and the concept of Granger causality.

5.4 A Brief Introduction to Multivariate Autoregressive Models

This section is based on the textbook by H. Lütkepohl [176]; however, here we adopt the notation and some observations presented in [226].

Multivariate autoregressive models (MVAR models), also called *vector autoregressive models* (VAR models), are stochastic process models used to capture the relationship between various quantities as they change over time. Formally, they are defined as follows.

Definition 5.4.1. A *VAR model of order* $p \in \mathbb{N}$, denoted by VAR(p), is given by

$$X(n) = \nu + A_1 X(n-1) + A_2 X(n-2) + \cdots + A_p X(n-p) + W(n), \tag{5.1}$$

where p represents the number of lags (past values) in the model, $X(n) = (x_1(n), ..., x_N(n))^T$ is an $N \times 1$ random vector, $n \in \mathbb{Z}$, A_i are $N \times N$ fixed coefficient matrices, $\nu = (\nu_1, ..., \nu_N)^T$ is an $N \times 1$ fixed intercept vector, and $W(n) = (w_1(n), ..., w_N(n))^T$ is an N-dimensional innovation process (or white noise), i.e. $E(W(n)) = 0$, $E(W(n)W(n)^T) = \Sigma_W$ and $E(W(n)W(m)^T) = 0$, for $n \neq m$, such that $E(\cdot)$ is the mean and Σ_W is the covariance matrix of $W(n)$. Here, Σ_W is assumed to be nonsingular.

Explicitly, Equation (5.1) is written as

$$\begin{bmatrix} x_1(n) \\ \vdots \\ x_N(n) \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_N \end{bmatrix} + \sum_{r=1}^{p} A_r \begin{bmatrix} x_1(n-r) \\ \vdots \\ x_N(n-r) \end{bmatrix} + \begin{bmatrix} w_1(n) \\ \vdots \\ w_N(n) \end{bmatrix}, \tag{5.2}$$

where

$$A_r = \begin{bmatrix} a_{11}(r) & \cdots & a_{1N}(r) \\ \vdots & \ddots & \vdots \\ a_{N1}(r) & \cdots & a_{NN}(r) \end{bmatrix}. \tag{5.3}$$

The coefficients $a_{ij}(r)$ of the matrices A_r represent the linear interaction effect of $x_j(n-r)$ on $x_i(n)$. The stochastic innovation processes $w_i(n)$ represent the part of the dynamic behavior that cannot be predicted from the past observations of the processes. Therefore, $w_i(n)$ are not time correlated but can exhibit instantaneous correlations among each other, which are described by their covariance matrix [226]:

$$\Sigma_W = \begin{bmatrix} \sigma_{11}(r) & \cdots & \sigma_{1N}(r) \\ \vdots & \ddots & \vdots \\ \sigma_{N1}(r) & \cdots & \sigma_{NN}(r) \end{bmatrix}, \tag{5.4}$$

where $\sigma_{ij} = \mathrm{Cov}(w_i(n), w_j(n))$.

A multivariate time series can be regarded as a (finite) realization of a vector stochastic process, thus it can be modeled by a VAR(p) model. There are numerous algorithms to estimate the VAR(p) model parameters from the input multivariate time series, such as the Arfit, multichannel Levinson, Viera-Morf, and Nuttall–Strand algorithms [179, 230]. The quality of the fitting, however, depends on the model order selection. This involves determining an optimal number for p, i.e. an optimal number of lags to include in the model. Choosing an appropriate p is essential, since a very small p can lead to loss of information about the series, and a large p can cause overfitting. There

are several criteria that can be used for this purpose, including: Akaike's information criterion (AIC), Hannan–Quinn's criterion, and Bayesian–Schwarz's criterion [176].

Now, let us formally define the concept of *stationarity* for $X(n)$ as a multivariate time series.

Definition 5.4.2. We say that a multivariate time series $X(n) = X_n = (x_1(n), ..., x_N(n))^T$ is *stationary* if

$$E(X_n) = \mu = (\mu_1, ..., \mu_N)^T, \quad \forall n, \quad \text{and} \tag{5.5}$$

$$E[(X_n - \mu)(X_{n-h} - \mu)^T] = E[(X_{n+h} - \mu)(X_n - \mu)^T], \quad \forall n \quad \text{and} \quad h = 0, 1, 2, ... \tag{5.6}$$

That is, if its first (mean) and second moments are *time invariant*.

5.5 Granger Causality

In a 1969 paper [124], the econometrician Clive Granger introduced the concept of causality for stationary stochastic processes that can easily be extended to time series [176]. The concept of *Granger causality* (or *G-causality*) lies in the idea that the cause cannot come after the effect. For time series, if we say that the series $x(n)$ "Granger-causes" (or "G-causes") the series $y(n)$, then the past values of $x(n)$ must contain information that helps to predict $y(n)$, in addition to the information contained in the past values of $y(n)$:

$$y(n) = \alpha + \beta_1 y(n-1) + \beta_2 y(n-2) + \cdots + \delta_1 x(n-1) + \delta_2 x(n-2) + \cdots + w(n). \tag{5.7}$$

An interesting property of Granger causality is that *it is not reciprocal*, i.e. $x(n)$ Granger-cause $y(n)$ does not imply that $y(n)$ Granger-cause $x(n)$. This makes it a suitable approach to directional causal processes, such as brain connectivity.

Formally, for time series, Granger causality is defined as follows. Let $x_n = x(n)$ and $y_n = y(n)$ be two time series and let $\{\Omega_n, n = 0, 1, 2, ...\}$ be a set of relevant information accumulated up to n (including n), containing at least x_n and y_n. Set $\bar{\Omega}_n = \{\Omega_s : s < n\}$, $\bar{\bar{\Omega}}_n = \{\Omega_s : s \leq n\}$, and let $\bar{x}_n$, $\bar{y}_n$, and $\bar{\bar{x}}_n$ be similar definitions. Given the information set B, let $P_n(y|B)$ be the predictor of y_n with the minimum mean square error (MSE); let $\sigma^2(y|B)$ be the corresponding MSE of the predictor.

Definition 5.5.1. We say that x_n *Granger-causes* y_n if

$$\sigma^2(y_n|\bar{\Omega}_n) < \sigma^2(y_n|\bar{\Omega}_n - \bar{x}_n). \tag{5.8}$$

In other words, y_n can be better predicted if we use all available information about the past of both x_n and y_n.

Definition 5.5.2. We say that x_n *instantaneously causes* y_n in the Granger sense if

$$\sigma^2(y_n|\bar{\Omega}_n, \bar{\bar{x}}_n) < \sigma^2(y_n|\bar{\Omega}_n). \tag{5.9}$$

That is, the present value of y_n is better predicted if the present value of x_n is taken into account.

Observation 5.5.1. Granger causality can be tested through linear prediction models [176]. For instance, for a VAR(p) model (5.1), testing for the existence of Granger causality from $x_i(n)$ to $x_j(n)$ is equivalent to verifying the hypothesis

$$H : a_{ij}(r) = 0, \quad \forall r = 1, ..., p, \tag{5.10}$$

i.e. $x_i(n)$ Granger-causes $x_j(n)$ if there is at least some coefficient $a_{ij}(r)$ different from zero. Moreover, note that $a_{ij}(r) = 0$ does not imply that $a_{ji}(r) = 0$, due to the non-reciprocity of Granger causality.

5.6 Partial Directed Coherence and its Variants

In this section, we introduce the mathematical formalism behind the partial directed coherence and its variants, specifically the generalized partial directed coherence and the information partial directed coherence, together with their asymptotic properties. The main reference for this part is [226].

5.6.1 Partial Directed Coherence

The concept of *partial directed coherence* (PDC) was first proposed by Baccalá and Sameshima [16, 17, 18] as a multivariate generalization of the *directed coherence* (DC) proposed by Saito and Harashima [225]. PDC is a quantifier based on the Granger causality concept, and can be considered a representation of Granger causality in the frequency-domain.

Formally, PDC is defined as follows. Consider a VAR(p) model given by

$$X(n) = \sum_{r=1}^{p} A_r X(n-r) + W(n), \tag{5.11}$$

where $X(n) = (x_1(n), ..., x_N(n))^T$ is a stationary multivariate time series (e.g., $X(n)$ may represents N channels of EEG signals in time n) and $W(n) = (w_1(n), ..., w_N(n))^T$ is an N-dimensional stationary Gaussian innovation process (with covariance matrix Σ_W). The order p may be estimated by one of the order selection criteria mentioned in Section 5.4.

By properly obtaining the coefficients $a_{ij}(r)$ of the matrices A_r (see Section 5.4), we can define a frequency-domain representation of (5.11) by defining the matrix $A(f)$ as follows:

$$A(f) = \sum_{r=1}^{p} A(r)e^{-i2\pi fr}. \tag{5.12}$$

Defining $\bar{A}(f) = I - A(f) = [\bar{a_1}(f)\bar{a_2}(f)...\bar{a_m}(f)]$, where $\bar{a}_j(f)$ represents the j-th column of the matrix $\bar{A}(f)$, the entries of $\bar{A}(f)$ are given by

$$\bar{A}_{ij}(f) = \begin{cases} 1 - \sum_{r=1}^{p} a_{ij}(r)e^{-i2\pi fr}, & \text{if i = j,} \\ -\sum_{r=1}^{p} a_{ij}(r)e^{-i2\pi fr}, & \text{otherwise.} \end{cases} \tag{5.13}$$

Definition 5.6.1. The *partial directed coherence* (PDC) from j to i at frequency f is defined by

$$\pi_{ij}(f) := \frac{\bar{A}_{ij}(f)}{\sqrt{\bar{a}_j^H(f)\bar{a}_j(f)}}, \tag{5.14}$$

where the superscript H denotes the Hermetian transpose.

Note that because of the dependence on $a_{ij}(r)$ in Equation (5.14), the nullity of $\pi_{ij}(f)$ at a given frequency implies the lack of G-causality from j to i. Also, the expression $\pi_{ij}(f)$ denotes the *direction* and *intensity* of the information flow from j to i at frequency f, and it satisfies the following normalization properties:

$$0 \leq |\pi_{ij}(f)|^2 \leq 1, \tag{5.15}$$

$$\sum_{i=1}^{N} |\pi_{ij}(f)|^2 = 1, \quad \forall j = 1, ..., N. \tag{5.16}$$

Observation 5.6.1. In the previous paragraphs we introduced the PDC as a quantifier based on the coefficients of a VAR model, however, the PDC does not depend on this specific model, but can be formulated through other multivariate models, such as the *vector moving average* (VMA) model and the *vector autoregressive moving average* (VARMA) model [21].

5.6.2 Generalized PDC

A scaling-invariant version of PDC, called *generalized partial directed coherence* (gPDC), was introduced by Baccalá et al. [22]. As stated in their article, the central problem of connectivity analysis is to analyze the hypothesis

$$H_0 : \pi_{ij}(f) = 0, \tag{5.17}$$

whose rejection implies the existence of a directed connection from $x_j(n)$ to $x_i(n)$, which cannot be explained by other series observed simultaneously. Considering a scenario where a series $y(n)$ Granger-causes $x(n)$, if $y(n)$ is amplified by a constant α, and taking $u(n) = \alpha y(n)$, we would eventually have $|\pi_{xu}(f)|^2 \to 0$, as α grows. To solve this problem, a generalization of the PDC, which makes it invariant to eventual gains affecting a time series, was defined as follows.

Definition 5.6.2. The *generalized partial directed coherence* (gPDC) from j to i at frequency f is defined by

$$\pi_{ij}^{(w)}(f) := \frac{\frac{1}{\sigma_i^2}\bar{A}_{ij}(f)}{\sqrt{\bar{a}_j^H(f)\,\mathrm{diag}(\sigma_i^2)^{-1}\bar{a}_j(f)}}, \tag{5.18}$$

where $\mathrm{diag}(\sigma_i^2)$ is the diagonal matrix of the variances σ_i^2 of the innovation processes $w_i(n)$.

Equation (5.18) preserves the normalization properties (5.15) and (5.16).

5.6.3 Information PDC

The *information partial directed coherence* (iPDC), introduced by Takahashi et al. [259, 261], is a modification of the PDC expression that formalizes the relationship between PDC and information flow. In the following, we expose the formal definition of the iPDC according to the above-mentioned articles.

Let $x = \{x(n)\}_{n\in\mathbb{Z}}$ and $y = \{y(n)\}_{n\in\mathbb{Z}}$ be two discrete-time stochastic processes. We can evaluate the relationship between x and y through the *mutual information rate* (MIR) (see [66] for more details on information theory), which compares the joint probability density with the product of the marginal probability densities of x and y:

$$\mathrm{MIR}(x,y) = \lim_{m\to\infty} \frac{1}{m+1} E\left[\log \frac{dP(x(1),...,x(m),y(1),...,y(m))}{dP(x(1),...,x(m))dP(y(1),...,y(m))}\right], \tag{5.19}$$

where dP denotes the probability density function. Note that if $\mathrm{MIR}(x,y) = 0$, then x and y are independent.

Let $\omega = 2\pi f$ (angular frequency) and let $S_{xx}(\omega)$ and $S_{yy}(\omega)$ be the auto-spectrum of x and y, respectively, and $S_{xy}(\omega)$ the cross-spectrum. The *coherence* between x and y is given by

$$C_{xy}(\omega) = \frac{S_{xy}(\omega)}{\sqrt{S_{xx}(\omega)S_{yy}(\omega)}}. \tag{5.20}$$

For stationary Gaussian processes, Equation (5.19) is related to the coherence C_{xy}

through the expression

$$\mathrm{MIR}(x,y) = -\frac{1}{4\pi}\int_{-\pi}^{\pi} \log(1-|C_{xy}(\omega)|^2)\,d\omega. \tag{5.21}$$

Now, consider a VAR model given by (5.11), with $p = +\infty$, and such that $\Sigma_W = E(w(n)w(n)^T)$ is the positive-definite covariance matrix of the N-dimensional stationary Gaussian process $W(n)$. A sufficient condition for the existence of this VAR model, with $p = +\infty$, is that the spectral density matrix associated with x is invertible at all frequencies and uniformly bounded above and below. Hence, the entries of the matrix $\bar{A}(\omega) = I - A(\omega) = [\bar{a}_1(\omega)\bar{a}_2(\omega)...\bar{a}_m(\omega)]$ are given by

$$\bar{A}_{ij}(\omega) = \begin{cases} 1 - \sum_{r=1}^{+\infty} a_{ij}(r)e^{-i\omega r}, & \text{if i = j,} \\ -\sum_{r=1}^{+\infty} a_{ij}(r)e^{-i\omega r}, & \text{otherwise.} \end{cases} \tag{5.22}$$

Definition 5.6.3. The *informational partial directed coherence* (iPDC) from j to i is defined by

$$\iota\pi_{ij}(\omega) := \frac{\sigma_i^{-1/2}\bar{A}_{ij}(\omega)}{\sqrt{\bar{a}_j^H(\omega)\Sigma_W^{-1}\bar{a}_j(\omega)}}, \tag{5.23}$$

where $\omega \in [-\pi,\pi)$ and $\sigma_i = E(w_i^2(n))$.

Theorem 5.6.1. *Let $X(n)$ be the stationary multivariate time series satisfying the VAR model (5.11). Then we have*

$$\iota\pi_{ij}(\omega) = C_{w_i\eta_j}(\omega), \tag{5.24}$$

where $\eta_j(n) = x_j(n) - E[x_j(n)|\{x_r(m), r \neq j, m \in \mathbb{Z}\}]$ is the partial process associated with $x_j(n)$, given the remaining series $\{x_r(m)\}_{r\neq j, m\in\mathbb{Z}}$.

The previous theorem shows that the iPDC from j to i measures the amount of common information between the partial process η_{ij} and the innovation w_i. In fact, replacing (5.24) in Equation (5.21) we get

$$\mathrm{MIR}(w_i,\eta_j) = -\frac{1}{4\pi}\int_{-\pi}^{\pi} \log(1-|\iota\pi_{ij}(\omega)|^2)\,d\omega. \tag{5.25}$$

5.6.4 General Expression for all PDC Variants

As noticed by Baccalá et al. [15], the PDC from j to i (5.14), and its two variants, gPDC (5.18) and iPDC (5.23), can all be obtained from the same general formula for

a given frequency f:

$$\pi_{ij}(f) = \frac{1}{s} \frac{\bar{A}_{ij}(f)}{\sqrt{\bar{a}_j^H(f) S \bar{a}_j(f)}}, \tag{5.26}$$

where the variables s and S are given according to the Table 5.1.

Table 5.1: Variables s and S according to the PDC type.

Variable	PDC	gPDC	iPDC
s	1	$\sigma_i^{-1/2}$	$\sigma_i^{-1/2}$
S	I	$\text{diag}(\sigma_i^2)^{-1}$	Σ_W^{-1}

Sometimes the notation $\text{diag}(\sigma_i^2) = (I \odot \Sigma_W)$ is used, where "$\odot$" denotes the Hadamard product of matrices (element-wise product).

Other forms of PDC have been introduced over time, such as the *time-varying generalized orthogonalized PDC* (tv-gOPDC) [197], which tries to reduce the effect of volume conduction, and the *total PDC* (tPDC) [20], which takes into account the instantaneous Granger causality.

5.6.5 Asymptotic Properties of the PDC and its Variants

As stated in [226], the problem of statistical connectivity inference, whether performed in the time or frequency-domain, actually involves two distinct problems:

- *The connectivity detection problem*: To detect the presence of significant connectivity at a given frequency.
- *The connectivity quantification problem*: To determine the confidence interval of the estimated value when it is significant at a given frequency.

In [260], it was shown that both problems can be rigorously examined from the perspective of asymptotic statistics.

The availability of confidence intervals based on a single trial makes it possible to consistently compare connection strengths under various experimental situations without the need to do repeated experiments based on ANOVA for inference [226]. In view of this, Baccalá et al. [15] worked to determine the asymptotic behavior of the three previously introduced forms of PDC, showing that significant values of PDC (gPDC, iPDC) are asymptotically Gaussian, and this normality is not verified when there is no connectivity.

In what follows, we summarize the results obtained in [15], omitting the proofs and some details. Consider the VAR model (5.11). Let n_s be the number of observed data points. Let $\theta^T = \alpha^T \sigma^T$ be the vector of parameters, where $\sigma = \text{vec}(\Sigma_W)$ and

$\alpha = \text{vec}(A_1 ... A_p)$, with $\text{vec}(\cdot)$ denoting the *vectorization operator* (i.e., the operator that converts a matrix into a column vector, by stacking all columns of the matrix). Note that the vector θ incorporates the dependence of a_{ij} and σ_{ij} according to the chosen PDC type.

The *confidence intervals* and the *limit of the null hypothesis* for the general form of the PDC (5.26) can be calculated by dividing its parameter dependence on the parameter vector θ, considering its decomposition into numerator and denominator:

$$|\pi_{ij}(f)|^2 = \pi(\theta) = \frac{\pi_n(\theta)}{\pi_d(\theta)}, \tag{5.27}$$

where the subscripts n and d indicate "numerator" and "denominator", respectively. Thus, the following results are valid:

- *Confidence Intervals*: For a large n_s, Equation (5.27) is asymptotically normal, i.e.

$$\sqrt{n_s}(|\hat{\pi}_{ij}(f)|^2 - |\pi_{ij}(f)|^2) \to \mathcal{N}(0, \gamma^2(f)), \tag{5.28}$$

  where $\gamma^2(f)$ is a frequency-dependent variance which depends on the PDC type.

- *Null hypothesis threshold*: The variance $\gamma^2(f)$ is identically zero under the null hypothesis

$$H_0 : |\pi_{ij}(f)|^2 = 0, \tag{5.29}$$

  therefore, Equation (5.30) is no longer valid, i.e. asymptotic normality is no longer satisfied. This requires the consideration of the next term of the Taylor expansion of the asymptotic expression of (5.27), which has a dependency $\mathcal{O}(n_s^{-1})$. The resulting distribution corresponds to a linear combination of at least two χ_1^2, with appropriate and frequency-dependent multiplication coefficients:

$$n_s \bar{a}_j^H(f) S \bar{a}_j(f)(|\hat{\pi}_{ij}(f)|^2 - |\pi_{ij}(f)|^2) \xrightarrow{d} \sum_{k=1}^{q} l_k(f)\chi_1^2, \tag{5.30}$$

  where the coefficients $l_k(f)$ depend only on the numerator $\pi_n(\theta)$, and "$\xrightarrow{d}$" designates the convergence in distribution.

That is, when the null hypothesis (5.29) is not rejected (lack of connectivity), the PDC tends to a distribution χ_1^2, and when (5.29) is rejected, the PDC tends to a normal distribution.

5.6.6 Examples with Simulations

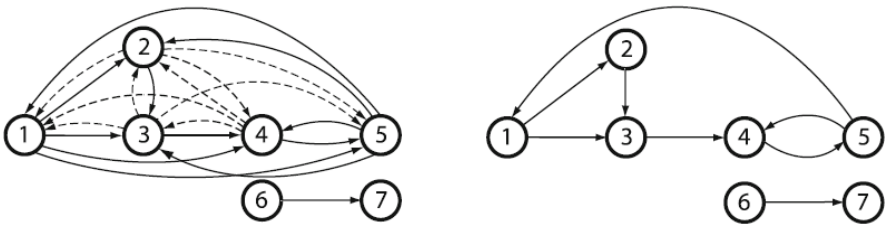
Given a multivariate time series $X(n) = (x_1(n), ..., x_N(n))$, by calculating the PDC between $x_i(n)$ and $x_j(n)$, for a given frequency f, for all i and j, a weighted connectivity

digraph can be constructed by drawing an arc from x_j to x_i (which correspond to the nodes) if and only if $|\pi_{ij}(f)|^2 \neq 0$ (where $\pi_{ij}(f)$ represents the general formula (5.26)), where the arc weights are equal to the values $|\pi_{ij}(f)|^2$. The same applies to DTF.

In the following example, which is based on the article [19], we applied the iPDC and iDTF estimators in a toy model consisting of seven ($N = 7$) linear difference equations as exposed in [16]. As commented previously, like PDC, DTF is a multivariate estimator based on the concept of Granger causality, which can describe the causal influence of one time series on another at a given frequency, and an information-theoretic form of DTF, the iDTF, was introduced in [259, 261]. The aim of this example is twofold: 1) to present how iPDC can be used to create a connectivity digraph; 2) to compare the differences in the connections generated by these estimators, which are direct and indirect in nature for iDFT, and only direct for iPDC. We emphasize that in this text we do not discuss in details the DTF/iDTF estimators, and it's up to the reader to consult [153, 261] for more details.

Example 5.6.1. Figures 5.3a and 5.3b present the directed graphs produced by the iDTF and iPDC estimator (only statistically significant ($p < 0.01$) connections were considered), respectively, when applied to the following VAR model [16]:

$$\begin{cases} x_1(n) = 0.95\sqrt{2}x_1(n-1) - 0.9025x_1(n-2) + 0.5x_5(n-2) + w_1(n) \\ x_2(n) = -0.5x_1(n-1) + w_2(n) \\ x_3(n) = 0.4x_1(n-4) - 0.4x_2(n-2) + w_3(n) \\ x_4(n) = -0.5x_3(n-1) + 0.25\sqrt{2}x_4(n-1) + 0.25\sqrt{2}x_5(n-1) + w_4(n) \\ x_5(n) = -0.25\sqrt{2}x_4(n-1) - 0.25\sqrt{2}x_5(n-1) + w_5(n) \\ x_6(n) = 0.95\sqrt{2}x_6(n-1) - 0.9025\sqrt{2}x_6(n-2) + w_6(n) \\ x_7(n) = -0.1x_6(n-2) + +w_7(n) \end{cases} \tag{5.31}$$



(a) iDTF network. (b) iPDC network.

Figure 5.3: Examples of iDTF and iPDC networks. Dashed lines represent weak connections (adapted from [19]).

The iDTF is an influenceability (reachability) estimator, i.e. it is unable to differentiate between direct and indirect interactions, whereas the iPDC is a directed connectivity estimator, i.e. it can detect direct interactions.

5.7 Brain Connectivity Networks

In this last section, we describe the different types of brain connectivity networks, presenting some examples; afterwards, we present some concerns related to brain connectivity networks estimated from neurophysiological data, and we conclude with a literature review and some considerations about the state of the art of graph theoretical analysis and topological data analysis in modern network neuroscience research.

5.7.1 The Different Types of Brain Connectivity Networks

Brain connectivity refers to the ways in which different brain structures interact and influence or are influenced by each other during sensory, motor, or cognitive tasks. As mentioned previously, brain connectivity encompasses three main types of connectivity, namely: *structural connectivity*, *functional connectivity*, and *effective connectivity* [108, 109, 143, 164, 239, 240].

Structural (anatomical) connectivity refers to a set of anatomical connections that link neural elements. These connections vary in scale, from large-scale networks of inter-regional routes to small circuits of individual neurons [240]. They form the *connectome* through which synapses between neighboring neurons and nerve fibers connect spatially distinct brain regions [241, 242, 246]. At shorter time scales (seconds to minutes), anatomical connections can be considered as static, but at longer time scales (hours to days), they can be thought of as plastic or dynamic [240]. Structural networks of the brain can be estimated through techniques such as diffusion tensor imaging (DTI) and fiber tractography [191].

Functional connectivity refers to the temporal interdependence of activation patterns of anatomically separate brain areas [240]. It displays the statistical relationships between different and remote-located populations of neurons. It is dependent on statistical metrics, such as correlation, covariance, or spectral coherence [164]. Since statistical relationships are highly reliant on time, the basis of functional connectivity analysis is time series data of neurophysiological activity, which can be extracted from EEG, MEG, fMRI, or other techniques. Unlike structural connectivity, functional connectivity does not necessarily depend on the anatomical connections of neuronal units. As mentioned in Section 5.3, examples of functional connectivity network estimators include correlation, coherence, mutual information, transfer entropy, and phase synchronization.

Effective connectivity reflects the causal relationships between activated brain regions by characterizing the influence that one brain structure has on another [240]. It may depict the directed effects inside a neuronal network by combining structural and effective connections. Effective connectivity is also time-dependent. Moreover, methods based on Granger causality may be used to infer causality from time series data of neu-

rophysiological signals. As discussed in Section 5.3, examples of effective connectivity network estimators include transfer entropy, GCI, DC, PDC, DTF, and DCM.

Although the previous classification between functional and effective connectivity categories is widely accepted in the literature, it may not be accurate enough to describe multivariate models. Accordingly, Baccalá and Sameshima [19] proposed an alternative classification based on two novel connectivity categories: *G-connectivity* and *G-influentiability*. G-connectivity (G stands for Granger) describes the direct, immediate, and active coupling between brain structures, but excludes active interactions that occur through intermediate (indirect) structures, and may be estimated by PDC (see Figure 5.5a). G-influentiability, on the other hand, describes both direct and indirect active connections, and may be estimated by DTF (reachability) (see Figure 5.5b). This description allows us to classify connectivity networks according to the nature of their links, as shown in Table 5.2.

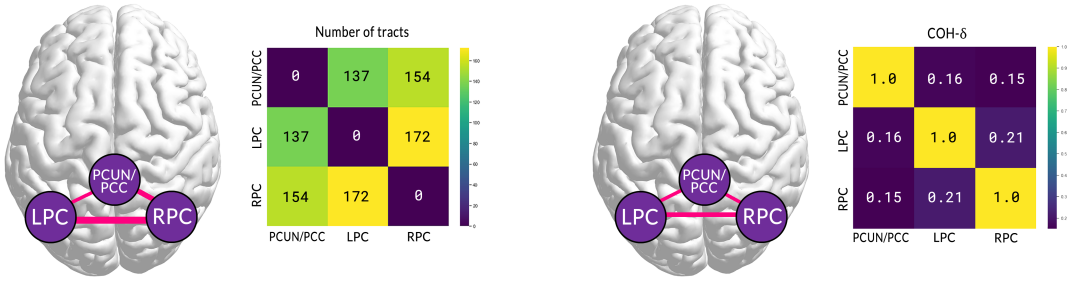
Table 5.2: G-connectivity/G-influentiability classification.

	Direct	**Indirect**
Active	$PDC \neq 0$	$PDC = 0$ and $DTF \neq 0$
Inactive	$PDC = 0$	$DTF = 0$

The dynamics of functional and effective connectivity networks may be analyzed through sliding window techniques. These techniques include selecting a temporal window of a specified length and using the data within it to estimate the connectivity network through a chosen measure. A (discrete) temporal network is created by shifting the window in time by a certain number of data points and repeating the procedure. This may be thought of as a quantification of the dynamics of the measure's behavior [147].

Example 5.7.1. In this example, we constructed structural and functional connectivity networks based on results extracted from the article [60]. The authors used EEG, fMRI, and DTI data from three patients diagnosed with autism spectrum disorder (ASD) to investigate how structural brain networks correlate with functional brain networks. Here we used three regions of interest (ROIs) determined in their study, namely: the Precuneus/Posterior Cingulate Cortex (PCUN/PCC), the Left Parietal Cortex (LPC), and the Right Parietal Cortex (RPC). The functional connectivities between these ROIs (nodes) were estimated through several functional connectivity measures from EEG signals, for five frequency bands (δ(1-4 Hz), θ(4-8 Hz), α(8-12 Hz), β(12-30 Hz), γ(30-45 Hz)), and we chose the results corresponding to the *coherence* in the delta band (COH-δ). The structural connectivities were obtained through DTI analysis, from which the number of white matter fibre tracts connecting the ROIs was estimated. Figure 5.4a

presents a structural connectivity network, along with its weighted adjacency matrix, in which the edge weights correspond to the number of tracts between the ROIs for to the results obtained for patient 1 (subject 1). Figure 5.4b presents a functional connectivity network, along with its weighted adjacency matrix, in which the edge weights correspond to the coherence (COH-δ) between the ROIs for the results obtained for the same patient.

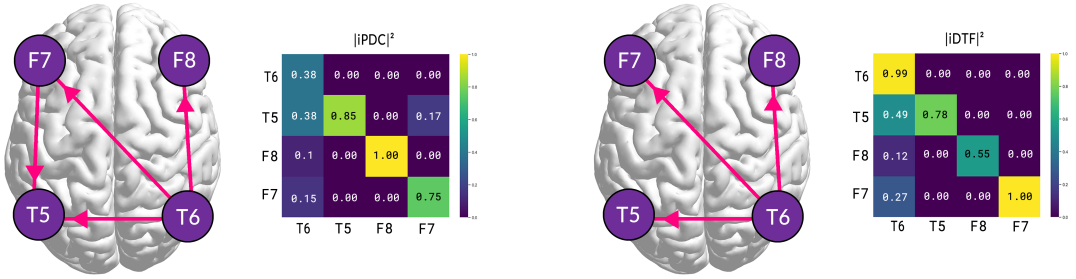


(a) Structural connectivity.

(b) Functional connectivity.

Figure 5.4: Structural and functional connectivity networks corresponding to patient 1 (subject 1). Structural connectivities correspond to the number of tracts between the ROIs, and the functional connectivities correspond to the coherence in the delta band (COH-δ) between the ROIs (edge thickness is proportional to the magnitude of the connectivity measure) (3D model generated with the BrainNet Viewer software [282]).

Example 5.7.2. As commented in Example 5.6.1, PDC and DTF are *directed* connectivity estimators, and thus it is possible to construct a connectivity digraph when applied to a multivariate time series. In this example, we constructed G-connectivity and G-influentiability networks based on data used in the article [14] (the data were provided by the authors). The authors used EEG data from nine patients diagnosed with left/right mesial temporal lobe epilepsy (MTLE) to assess the seizure lateralization through graph theoretical analysis (GTA) of connectivity networks estimated via PDC. The records were performed using a total of 29 electrodes placed according to the international 10–20 system in referential montage, with a sampling rate of 200 Hz. We chose four channels (F7-Ref, F8-Ref, T5-Ref, T6-Ref) and computed the iPDC and iDTF estimators on a 10s epoch (patient 1), starting at seizure onset, in the delta frequency band (1-4 Hz). The iPDC and iDTF networks were estimated via the MATLAB package asympPDC (see Appendix A), using the Nuttall-Strand algorithm to estimate the parameters of the VAR model, with a fixed order set at $p = 2$, and the statistically significant connections were estimated asymptotically (see Subsection 5.6.5) with a significance level of 0.1%. Figures 5.5a and 5.5b present the G-connectivity network obtained via iPDC and the G-influentiability network obtained via iDTF, respectively, along with their respective weighted adjacency matrices.



(a) G-connectivity network (iPDC). (b) G-influentiability network (iDTF).

Figure 5.5: G-connectivity (iPDC) and G-influentiablility (iDTF) digraphs, along with their weighted adjacency matrices (3D model generated with the BrainNet Viewer software [282]).

5.7.2 Concerns about Brain Connectivity Networks

There are some concerns about the estimation and validity of brain connectivity networks that should be considered in their analysis. Below, we list some of the main points that must be taken into consideration when obtaining and interpreting brain connectivity [53, 164, 226].

- **Variability of data extraction:** As briefly discussed in Section 5.1, there are several techniques to detect and measure the neurophysiological activity of the brain, such as fMRI, EEG, and MEG. These techniques vary widely in terms of spatial and temporal resolution, and the specific aspects of brain activity that they measure. Therefore, the data acquisition method directly impacts the interpretation of the connectivity networks subsequently estimated.

- **Data preprocessing:** Data preprocessing procedures, such as motion correction, high pass filtering, and RETROICOR correction (RTC) for fMRI data, and downsampling and choice of reference for EEG data, may influence connectivity estimates. For instance, Výtvarová et al. [270] found that the strength of the connectivities in functional connectivity networks obtained from fMRI data was statistically significantly affected by high-pass filtering, and that the topology of these networks was affected by RTC.

- **Selection of connectivity estimator:** As already discussed in Section 5.3, there are numerous measures for estimating brain connectivity. Measures such as correlation, coherence, phase synchronization, GCI, DC, PDC, DTF, and DCM may yield different or even conflicting results under similar conditions. Therefore, the choice of the connectivity estimator can have a decisive impact on the interpretation of the results.

- **Volume conduction:** In Section 5.2 we already discussed the phenomenon of

volume conduction and how it affects EEG signals. Nevertheless, its effects may also impact the results of other techniques, such as MEG. Volume conduction effects can lead to false assumptions about the interactions between different brain regions. For instance, considering a functional connectivity network estimated via correlation from EEG signals, two regions might appear to be connected because their signal patterns are correlated. However, this correlation might be due to the signals from one region spreading to the electrodes near another region, rather than any true functional interaction between them. The effects of volume conduction are inevitable, but there are methods capable of reducing them [139]. In particular, both DTF and PDC are affected by volume conduction [45].

- **Generalizability:** Brain connectivity networks are often estimated from data obtained from relatively small populations and in specific contexts (whether from healthy individuals or individuals with some neurological disorder), therefore, it is essential to take this limitation into account in the applicability of results to general populations and contexts.

All of the points mentioned above impact the analysis of brain connectivity networks; therefore, they should be taken into consideration when interpreting the results.

5.7.3 Applications of Brain Connectivity Networks Analysis

As we saw in the previous sections, brain connectivity networks can be represented as graphs (sometimes referred to as *brain graphs* [47]), where the nodes represent different brain structures and the edges represent specific associations between them (e.g., anatomical connections, statistical dependencies, or causal influences). Thus, they can be analyzed through algorithms and quantitative methods from graph theory and network science [46, 94, 223, 240, 252]. In recent years, graph theoretical analysis (GTA) of brain networks has proven to be a powerful analytic tool, and has been widely used in the area of network neuroscience [96, 103].

In Chapter 1, we have already briefly discussed the applications of graph theory in the analysis of brain networks. Many of the GTA have shown that brain structural, functional, and effective networks may present, depending on the context, several non-trivial topological and organizational properties [289], such as hierarchical organization [28], clustering, modularity [223, 240], presence of hubs (which may be interpreted as important brain regions) [95, 244], presence of structural and functional network motifs [245, 240], and small-world organization [3, 29, 201, 247].

Moreover, graph theoretical analysis of brain networks has played an important role in the search for *biomarkers* for different aspects of brain functioning. According to the National Institutes of Health (NIH) [134], a biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biological

processes, pathogenic processes or pharmacologic responses to a therapeutic intervention." Alterations in structural or functional connectivity of the brain may be used as biomarkers to detect structural and functional abnormalities in network connectivity patterns, which might be indicators of neurological disorders or specific cognitive processes [274]. As mentioned in the systematic review carried out by Farahani et al. [96] of GTA studies of brain networks using fMRI data, graph theory applications in human cognition include to find biomarkers (in this case, fMRI-based biomarkers) for the human intelligence (e.g., shorter characteristic path lengths in functional networks were associated with better intellectual performances [265]), working memory (e.g., better performance of working memory was associated with lower local efficiency [253]), aging brain (e.g., global efficiency was essentially unchanged over the lifespan, whereas local efficiency and the rich-club coefficient increased until adulthood in healthy individuals and decreased with age [96]), and behavioral performance in natural environments (e.g., small-world organizations were maintained in individuals in both normal and hyperthermia conditions, whereas decreased clustering coefficients, local efficiency, and small-worldness indices were observed in heat-exposed individuals [213]), and applications in neurological disorders include to find biomarkers for disorders such as epilepsy (e.g., disruption of global integration and local segregation was observed in patients with chronic epilepsy [268]), Alzheimer's disease (AD) (e.g., patients with AD showed a significantly decreased clustering coefficient and characteristic path length compared to healthy individuals [71]), multiple sclerosis (MS) (e.g., patients diagnosed with MS showed decreased global efficiency relative to healthy individuals [173]), autism spectrum disorders (ASD) (e.g., in individuals diagnosed with ASD, modularity, clustering coefficient, and local efficiency are relatively reduced compared with healthy individuals [157]), and attention-deficit/hyperactivity disorder (ADHD) (e.g., patients with ADHD showed increased local efficiency and decreased global efficiency compared to healthy individuals [272]). Be aware that the previous network properties were obtained for (*undirected*) functional networks estimated through different methods.

Besides the fMRI-based biomarkers, particularly in the study of neuropathologies, the discovery of EEG-based biomarkers (electrophysiological biomarkers) is of special interest due to the portability and low cost of EEG compared to fMRI. Among the neuropathologies studied and the topological and functional characteristics found in connectivity networks obtained from EEG data, we can mention, for example [172, 251]: patients with Parkinson's disease (PD) showed increased connectivity strength in the theta band compared with healthy controls [263]; patients with epilepsy showed abnormally regular functional networks relative to healthy controls [142]; patients with schizophrenia (SCZ) showed a decreased clustering coefficient, decreased global and local efficiency, and increased average characteristic path length compared to healthy controls [286]. Again, the previous network properties were obtained for (*undirected*) functional networks estimated through different methods.

Recently explored approaches to the study of brain networks have involved concepts from computational (algebraic) topology and computational geometry, such as simplicial complexes, homology, homotopy, Betti numbers, and persistent homology (PH). The set of all these tools in data analysis, as we already mentioned in Chapter 1, is identified by the umbrella term topological data analysis (TDA) [51]. Applications in network neuroscience include, for example, assessing neural functions and structures through clique topology and PH of clique complexes built out of functional or structural brain networks [119, 120, 207, 218, 237], to characterize functional brain networks of patients with ADHD and ASD through PH [167, 168], to characterize [182] and detect [99, 208, 256, 273] epileptic seizures via PH, and to assess the dynamics of functional brain connectivity through persistence vineyards [287]. Also, concepts from Q-Analysis have been used to study functional connectivity networks [258] and human connectomes [8].

In summary, understanding the dynamic nature of brain networks is essential for understanding brain function and dysfunction in healthy individuals and individuals diagnosed with neurological disorders. As we have seen, there are a myriad of techniques, including EEG, fMRI, MEG, and mathematical, statistical, and computational methods, to estimate and study the organization, topology, and dynamics of brain networks, with the aim of identifying patterns of neural activity that are associated with specific brain processes.

In the next chapter, we will discuss in more depth the role of brain connectivity in individuals with epilepsy and how GTA and TDA can be useful in the study of epileptic brain networks.

Chapter 6

Epilepsy as a Disorder of Brain Connectivity

> *Hippocrates left behind him only a single discussion of the function of the brain and the nature of consciousness. It was included in a lecture delivered to an audience of medical men on epilepsia, the affliction that we still call epilepsy. Here is an excerpt from this lecture, this amazing flash of understanding: "Some people say that the heart is the organ with which we think and that it feels pain and anxiety. But it is not so. Men ought to know that from the brain and from the brain only arise our pleasures, joys, laughter and tears. Through it, in particular, we think, see, hear and distinguish the ugly from the beautiful, the bad from the good, the pleasant from the unpleasant.... To consciousness the brain is messenger." And again, he said: "The brain is the interpreter of consciousness." In another part of his discussion he remarked, simply and accurately, that epilepsy comes from the brain "when it is not normal."*
>
> — Wilder Penfield [205]

Disorders of brain connectivity encompass a wide range of conditions affecting neural connectivity, communication, and information processing in the brain. As we discussed in the previous chapter, several abnormalities in brain connectivity networks, whether structural, functional, or effective, have been observed in brain diseases such as Parkinson's disease, Alzheimer's disease, multiple sclerosis, schizophrenia, and epilepsy. The discovery of links between the topology of brain networks and neurological disorders has increasingly encouraged researchers to use mathematical methods from graph theory, network science, and computational topology to understand the changes in the dynamics of brain connectivity that underlie the symptoms and progression of specific neuropathologies, which could potentially lead to new diagnostic methods and therapeutic approaches.

In this chapter, we discuss in more detail the neuropathology of epilepsy, which is considered one of the most prominent brain network disorders. We discuss its general characteristics, the different types of epilepsy, the classification of epileptic seizures, the etiology, epidemiology, diagnosis and treatment of epilepsy, and, finally, how it is characterized as a network disorder by exploring the topological/structural and functional properties of epileptic networks through methods from graph theory and computational topology.

6.1 An Introduction to Epilepsy

In this section, we present the main characteristics related to epilepsy, the different types of epilepsy, the classification of epileptic seizures, the etiology, epidemiology, diagnosis, and treatment of epilepsy.

6.1.1 General Aspects of Epilepsy

Epilepsy is one of the most common neurological disorders, present in populations all over the globe. It is a disorder of the central nervous system (CNS) typified by recurrent and non-induced seizures (defined as a transient period of excessively synchronous (hypersynchronous) of abnormal neuronal activity manifested in the brain in a localized or generalized way) that occur over an interval of time, and which may occur spontaneously [6, 112, 275]. Epilepsy is defined by the number and frequency of non-induced seizures and can be classified as *focal* (seizure onset involves a specific area of the brain), *generalized* (seizure onset involves both cerebral hemispheres), *combined generalized and focal*, or *unknown* (see next subsection), depending on the area of its origin in the brain [101].

The first phase of a seizure is known as *pre-ictal phase* (or *aura*) and it occurs immediately before the *ictal phase*, which corresponds to the seizure itself. The *post-ictal phase* is the period right after the ictal phase. The *interictal period* corresponds to the period between seizures. However, determining the precise length of the pre-ictal, ictal, or post-ictal phase is not very clear and may vary depending on each case [24]. Although some crises may last for a short time, most of them last from seconds to minutes.

The *epileptogenic zone* (EZ) refers to the cortical regions involved in the genesis and propagation of the epileptiform activity. The EZ is most likely equivalent to the *seizure focus*, which is defined as the location in the brain from whence the seizure began [193]. For focal epilepsies, the side (left or right cerebral hemisphere) of seizure onset (or the side of seizure focus) is known as the *lateralization* of the seizure. The most common form of focal epilepsy is the *temporal lobe epilepsy* (TLE).

The identification of graphoelements in EEG recordings (patterns in the EEG signal) is a commonly used method in the study of epilepsy. For instance, the majority of individuals with epilepsy exhibit typical *interictal epileptiform discharges* (IEDs), often known as spike ($< 70\mu$s duration), spike-and-wave, or sharp-wave (70–200μs duration) discharges [250]; in TLE, an EEG pattern that has been consistently observed is the *temporal intermittent rhythmic delta activity* (TIRDA), which is characterized by an intermittent ($\geq$ 3s), rhythmic, 1-4Hz activity in the anterior temporal region [104].

The diagnosis of epilepsy is complicated by the fact that many of the signs and symptoms of epilepsy occur over brief, irregular periods of time (e.g., seizures, IDEs). This means that in clinical EEG examinations, brain activity may appear normal; even so, EEG is the most used method to confirm the diagnosis of epilepsy. Moreover, the definitive diagnosis and choice of therapy is based on the analysis of the EEG by a specialist, which can take hours; for this and other reasons, the availability of automated systems capable of detecting epilepsy accurately (and eventually in real-time), would be of great value in clinical practice [67, 275].

For patients diagnosed with epilepsy, antiepileptic drugs (AEDs) are the main form of treatment. However, some patients present drug-resistant epilepsy (DRE) [216], which is a pharmacoresistant form of the disease that represents one-third of epilepsies [98] and, in these cases, a surgical intervention may be necessary to resect or disconnect the supposed EZ [221].

6.1.2 A Closer Look into the Neuropathology of Epilepsy

In the previous subsection, we briefly discussed the general aspects related to the neuropathology of epilepsy. Now, we present in more detail the classification of the different types of seizures and epilepsies, the etiology, epidemiology, diagnosis, and treatment of epilepsy.

Classification of Seizure Types and Epilepsies

Epileptic seizures may be classified according to the type of onset, level of awareness, and responsiveness [100]. According to the International League Against Epilepsy (ILAE), epileptic seizures can be classified into *focal onset*, *generalized onset*, and *unknown onset* [101]:

- **Focal onset (aware/impaired awareness):** These seizures originate from a specific area of the brain and may or may not involve loss of consciousness (impaired awareness). Seizures of this type can additionally be classified into *motor onset* (automatisms, atonic, clonic, epileptic spasms, hyperkinetic, myoclonic, tonic); *non-motor onset* (autonomic, behavior arrest, cognitive, emotional, sensory); and *focal to bilateral tonic-clonic*.

- **Generalized onset:** These seizures involve both hemispheres of the brain from the onset and typically result in loss of consciousness (impaired awareness). Seizures of this type can additionally be classified into: *motor onset* (tonic-clonic, clonic, myoclonic, myoclonic-tonic-clonic, myoclonic-atonic, epileptic spasms); and *non-Motor (absence)* (typical, atypical, myoclonic, eyelid myoclonia).
- **Unknown onset:** Seizures where their specific origin or onset cannot be determined with certainty. Seizures of this type can additionally be classified into: *motor* (tonic-clonic, epileptic spasms); *non-motor* (behavior arrest); and *unclassified*.

In addition, the epilepsy type is classified into four categories, namely: *generalized epilepsy*, *focal epilepsy*, *combined generalized and focal epilepsy*, and *unknown epilepsy*.

Etiology

The causes of epilepsy can vary widely. Since there are many distinct processes governing the electrical activity of neurons, there exists a wide variety of possibilities that can disturb these processes, which can lead to a variety of reasons for epilepsy and seizures. Accordingly, six etiologic categories for epilepsy have been established by the ILAE Task Force [229], namely: *structural, genetic, infectious, metabolic, immunological*, and *unknown*.

- **Structural etiology:** Structural etiology corresponds to the case in which abnormalities are found on structural neuroimaging that, together with results obtained from electrophysiological tests, suggest that the abnormalities are the cause of the patient's seizures.
- **Genetic etiology:** Genetic etiology refers to the case in which the disorder is thought to be directly caused by a known or suspected genetic abnormality.
- **Infectious etiology:** Infectious etiology refers to the case in which epilepsy results as a result of a known infection. This is the most common etiology.
- **Metabolic etiology:** Metabolic etiology refers to the case in which epilepsy results from a suspected metabolic disorder.
- **Immunological etiology:** Immunological etiology refers to the case in which epilepsy results from an immunological disorder in the patient's organism.
- **Unknown etiology:** Unknown etiology refers to the case in which the cause of the epilepsy is not known.

Epidemiology

The incidence of epilepsy worldwide is contained in the range between 0.5 and 1.5% and is higher in developing countries, mainly due to poor prevention conditions and lack of access to suitable treatments. The prevalence follows this 1% trend and is also higher in developing countries.

Although the occurrence of a spontaneous seizure (i.e., without an identified cause) does not guarantee the diagnosis of epilepsy, the risk of a second seizure is around 40%. After a second seizure, the risk increases to almost 100%. In addition, up to 30% of the patients diagnosed with epilepsy may present a pharmacoresistant form of the disorder (i.e., DRE), and 10% may need a surgical resection or disconnection of the EZ [112].

Diagnosis

The first step towards diagnosing epilepsy is the occurrence of at least one non-inducted (unprovoked) seizure. Afterwards, the next steps of the diagnosis include medical history (e.g., family history of epilepsy, history of brain infection, traumatic brain injury, febrile seizures, etc.), physical examination (blood pressure test, skin exam, checking for signs of infection, cancer, etc.), EEG, and neuroimaging (e.g., MRI) [184]. Based on the results of these tests and evaluations, a neurologist or epileptologist can make a diagnosis of epilepsy and recommend an appropriate treatment.

Treatment

Epilepsy, to this day, is a neuropathology that cannot be cured. Nevertheless, there are several treatments available to control seizures in patients diagnosed with the disease, allowing them to live an unrestricted life. The three most common treatments for epilepsy are: administration of antiepileptic drugs (pharmacological treatment); surgical intervention to resect or disconnect the EZ (neurosurgical treatment); and neurostimulation [112].

6.2 Grapho-Topological Characteristics of Epileptic Brain Connectivity Networks

Epilepsy is notoriously a brain connectivity network disorder, characterized by a direct relation between abnormal network dynamics and clinical manifestations [112]. A fact that has been constantly observed in the literature is that patients diagnosed with epilepsy present alterations in brain connectivity networks, whether structural, functional, or effective when compared to healthy individuals [96, 172, 251]. Many of these findings come from results obtained through graph theoretical analysis (GTA) and topological data analysis (TDA) of these connectivity networks, and that is the main reason

why these methods have been widely used to try to answer questions such as: How do epileptic brain networks differ from healthy brain networks? How does brain network topology change in the ictal phase? How can epileptic networks be characterized? How to detect and predict epileptic seizures?

In what follows, we summarize some results obtained from studies that indicate the association of epilepsy with changes in brain networks, particularly associated with (undirected) structural and functional networks estimated from MRI/fMRI, EEG, sEEG, or DTI data.

- Bernhardt et al. [32] estimated structural connectivity networks through MR-based cortical thickness correlations from MRI data obtained from 122 patients with drug-resistant temporal lobe epilepsy (TLE) (52 males and 70 females; age range: 17-62) and 47 healthy controls (23 males and 24 females; age range: 18-66). They found that patients with TLE showed changes in the distribution of hubs, increased path lengths, increased clustering coefficients, and increased vulnerability to targeted attacks compared with healthy controls.

- Liao et al. [169] estimated functional connectivity networks through Pearson's correlation from resting-state fMRI data obtained from 18 patients with mesial temporal lobe epilepsy (MTLE) and 27 healthy controls. They found a pattern of significantly increased local connectivity and decreased global connectivity (typical characteristics of regular networks) in patients with MTLE compared with healthy controls.

- Ponten et al. [212], estimated functional connectivity networks through synchronization likelihood from sEEG data obtained from 7 patients diagnosed with MTLE. They found that, during the ictal phase, a change in the topology of the networks occurred, with an increase in the clustering coefficient and the characteristic path length, especially in the delta, theta, and alpha bands, which might suggest a change from a small-world organization (which seems to be characteristic of healthy individuals) to a more regular organization.

- Bonilha et al. [40] estimated structural connectivity networks from DTI data of 12 patients with MTLE (5 with left MTLE and 7 with right MTLE) and 26 healthy controls. They found that, in the thalamus, ipsilateral insula, and superior temporal region, patients with MTLE showed increased degree, local efficiency, clustering coefficient, and limbic network clustering compared with healthy controls.

- Vlooswijk et al. [268] estimated functional connectivity networks through Pearson's correlation from fMRI data obtained from 41 patients diagnosed with chronic epilepsy (20 males and 21 females; age range: 22-63) and 23 healthy controls (9

males and 14 females; age range: 18-58). They found a disruption of global integration and a disruption of local segregation in patients with epilepsy, and efficient small-world properties (high clustering coefficients and short characteristic path lengths) in healthy controls.

- Horstmann et al. [142] estimated functional connectivity networks through cross-correlation and mean phase coherence from EEG and MEG data obtained from 21 patients diagnosed with drug-resistant epilepsy (9 males and 12 females) and 23 healthy controls (12 males and 11 females). They found that epileptic brain networks showed abnormally regular functional networks relative to healthy controls.

- Merelli et al. [182] estimated functional connectivity networks through Pearson's correlation from EEG data obtained from 10 patients diagnosed with focal epilepsy (5 males and 5 females; age range: 0.5-19). From these networks, they constructed clique complexes and computed weighted persistent entropy for the pre-ictal and ictal phases. They found that the analysis of the persistent entropy can detect the transition between the pre-ictal and ictal phases.

Next, we present some results obtained for (directed) effective connectivity networks estimated from EEG data.

- Hu et al. [144] estimated G-connectivity networks through PDC from EEG data obtained from 10 patients diagnosed with focal epilepsy (5 males and 5 females; age range: 0.5-19). They found that, in the delta band, the total degree (the sum of in-degree and out-degree) at the center lobe during the ictal phase was significantly lower compared with the interictal period, and the clustering coefficients were significantly increased in the frontal, parietal, and temporal lobes during the ictal phase compared compared with the interictal period.

- Baccalá et al. [14] estimated G-connectivity networks through PDC from EEG data obtained from 9 patients with left / right MTLE. They found that channels in the hemisphere corresponding to the seizure focus belong to strongly connected subdigraphs, suggesting a possible graph-based biomarker to identify laterality.

- Coito et al. [62] estimated G-connectivity networks through PDC from EEG obtained from 16 patients with TLE (eight with left TLE (LTLE) and eight right TLE (RTLE); 12 males and 4 females; age range: 15-56). They found significantly different patterns between the networks of the LTLE and RTLE groups: ipsilateral predominance in LTLE and bilateral predominance in RTLE.

As we have seen, different data acquisition methods, connectivity estimators, and network quantifiers may provide different interpretations of the same phenomenon.

However, it is notorious across these different analysis modalities that there are discrepancies between connectivity networks of healthy individuals and individuals diagnosed with some type of epilepsy, as well as alterations in the brain networks during the ictal phase compared with other periods. Therefore, it is worth searching for graph-based biomarkers for the characterization of epileptic brain networks.

Chapter 7

Quantitative Graph/Simplicial Analysis of Epileptic Brain Networks

> *Le hasard n'est que la mesure de notre ignorance. Les phénomènes fortuits sont, par définition, ceux dont nous ignorons les lois.*
> *(Chance is only the measure of our ignorance. Fortuitous phenomena are, by definition, those whose laws we are ignorant of.)*
>
> — Henri Poincaré [211]

In this chapter, we apply some of the simplicial characterization measures and simplicial similarity comparison distances introduced in Chapter 4 to directed flag complexes built out of G-connectivity networks estimaed through iPDC from EEG signals of patients diagnosed with left temporal lobe epilepsy.

More specifically, we applied the iPDC estimator to epileptic EEG signals and constructed G-connectivity networks in three different frequency bands (delta, theta, and alpha) for the left and right brain hemispheres of each patient, and for three seizure phases (pre-ictal, ictal, and post-ictal phase). Subsequently, we constructed directed flag complexes from these networks, computed some chosen simplicial characterization measures and simplicial similarity comparison distances for each hemisphere, frequency, and seizure phase, in five levels of organization q ($q = 0, 1, 2, 3, 4$) of the complexes, and, finally, we performed a statistical analysis to evaluate the seizure phases and lateralization in each frequency band, in each hemisphere, and at each level q. Furthermore, we computed the standard graph measures introduced in Chapter 2 and compare their statistical results with the results obtained for their simplicial analogues.

The aim of this analysis is twofold: 1) to understand how epileptic networks and their higher-order counterparts change throughout different seizure phases, in different

frequency bands, and for each cerebral hemisphere; 2) to identify novel biomarkers for epileptic brain networks associated with their higher-order structures and higher-order connectivities.

7.1 EEG Data Acquisition and Preprocessing

The EEG data used in this study was obtained from the Siena Scalp EEG Database (SSED)[1] [78], which consists of EEG recordings acquired from 14 patients (9 males and 5 females), diagnosed with epilepsy, at the Unit of Neurology and Neurophysiology at the University of Siena, Italy. The EEG signals were recorded using a Video-EEG with sampling rate of 512 Hz, and were made available in the European Data Format (EDF). Three types of seizures were identified and classified according to the criteria of the ILAE, namely: focal onset impaired awareness, focal onset without impaired awareness, and focal to bilateral tonic–clonic. Moreover, the documentation includes annotations on the start and end times of the ictal phase for each recording.

From the SSED, we selected eight patients with left temporal lobe epilepsy (TLE) based on the quality of the signal. For these patients, the records were performed using a total of 29 electrodes (Fp1, F3, C3, P3, O1, F7, T3, T5, Fc1, Fc5, Cp1, Cp5, F9, Fz, Cz, Pz, Fp2, F4, C4, P4, O2, F8, T4, T6, Fc2, Fc6, Cp2, Cp6, F10), placed according to the international 10–20 system in a referential montage. Table 7.1 presents detailed information about the selected patients: columns 1 (Pat. id) presents the patient identification; columns 2 (Age) reports their ages; column 3 (Gender) reports their gender; column 4 (Localization) reports the location of the seizure focus; column 5 (Lateralization) reports the lateralization (left or right) of the seizure focus; column 6 (Time) presents the total registration time in minutes.

Table 7.1: Patient information.

Pat. id	Age	Gender	Localization	Lateralization	Time (min)
PN01	46	Male	Temporal Lobe	Left	809
PN06	36	Male	Temporal Lobe	Left	722
PN07	20	Female	Temporal Lobe	Left	523
PN09	27	Female	Temporal Lobe	Left	410
PN12	71	Male	Temporal Lobe	Left	246
PN13	34	Female	Temporal Lobe	Left	519
PN14	49	Male	Temporal Lobe	Left	1408
PN16	41	Female	Temporal Lobe	Left	303

Figure 7.1 presents the whole preprocessing work flow performed on the EEG data. Specifically, the preprocessing was performed in EEGLAB and included the following

[1]Public available at https://physionet.org/content/siena-scalp-eeg/1.0.0/

steps: (1) The raw EEG signals were downsampled from 512 Hz to 256 Hz. (2) A 1 Hz high-pass filter was applied to each channel. (3) A notch filter at 50Hz (European standard power line frequency) was applied to each channel using the CleanLine[2] EEGLAB plugin. (4) The midline electrodes (Fz, Cz, and Pz) were removed as they may be contaminated with the electrical activities of both hemispheres, thus a final configuration with 26 channels was obtained, as is schematically represented in Figure 7.2. (5) Finally, an independent components analysis (ICA) (employing the InfoMax algorithm) was performed to remove artifactual components. Figure 7.3 presents a cut of 162 seconds of preprocessed signal (with the midline electrodes) as an example.

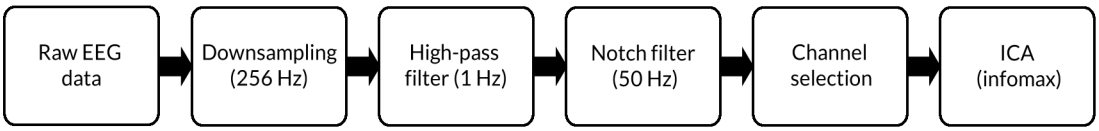


Figure 7.1: EEG data preprocessing workflow.

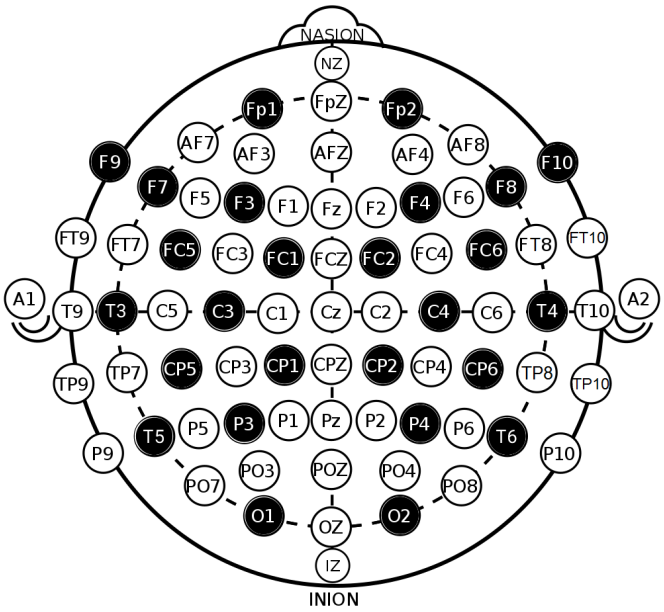


Figure 7.2: Configuration of the 26 selected electrodes (represented in black), according to the international 10–20 system.

[2] https://www.nitrc.org/projects/cleanline

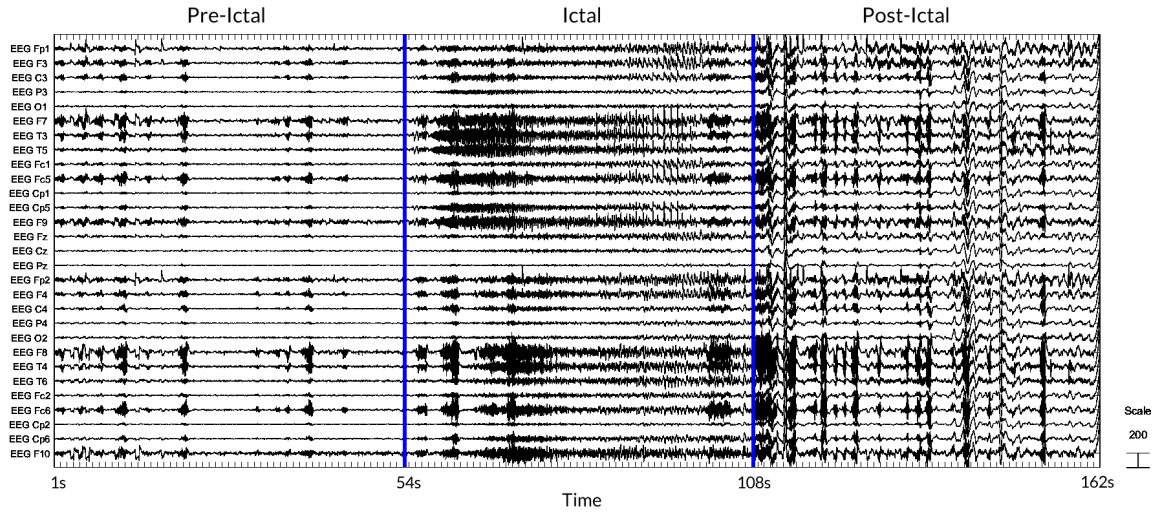


Figure 7.3: Example of preprocessed signal corresponding to patient PN01 (crisis 1). This signal represents a 162s cut from the original signal, in which the first 54s correspond to the pre-ictal phase, followed by 54s of the ictal phase, and the final 54s correspond to the post-ictal phase (separated by blue lines). The vertical scale is set to 200 μV.

7.2 Analysis of Seizure Phases and Lateralization in Left Temporal Lobe Epilepsy

In this section, we present the methodology and the results of the analysis performed in the preprocessed EEG data described in the previous section.

7.2.1 Methodology

Brain G-Connectivity Networks

We started our analysis by identifying the pre-ictal, ictal, and post-ictal phases in the preprocessed EEG signals based on the annotations provided together with the data, and then dividing the signals from each phase into 30s epochs: 30s immediately before the seizures, 30s starting on the seizures onset, and 30s immediately after the seizures.

Afterwards, we applied the iPDC estimator on each 30s epoch using a sliding window technique with fixed-size windows of 10s and 80% overlap, for three different frequency bands, namely: delta band [1 Hz, 4 Hz) (here, we are considering only frequencies ≥ 1 Hz), theta band [4 Hz, 8 Hz), and alpha band [8 Hz, 14 Hz). This procedure resulted in 11 weighted digraphs, or G-connectivity networks (corresponding to a discrete-time dynamic digraph), for each seizure phase in each of the frequency bands, where the weights correspond to the strengths of the connections.

The iPDC networks were computed via the MATLAB package `asympPDC` version 3.0 (see Appendix A). The VAR autoregression coefficients of these networks were estimated

through the Nuttall-Strand's method. The order for the VAR models was estimated by three different information criteria, namely: Akaike's information criterion, Hannan-Quinn's criterion, and Bayesian–Schwartz's criterion, which yielded an order equal to 2; however, the connectivity patterns differed just slightly from models with fixed order set to 12, therefore we chose to use a fixed order equal to 12. Only the statistically significant connections were considered, and they were estimated asymptotically (see Subsection 5.6.5) with a significance level of 0.1%. Figures C.22, C.23, and C.24 present examples of the dynamic iPDC networks (in discrete time) obtained for patient PN01 for the pre-ictal, ictal, and post-ictal phases in the delta band, respectively.

After obtaining the iPDC networks, as we are not taking the weights into account, we converted all of them into binary networks. Also, we removed all double-edges in such a way that the clique structures of the networks were preserved. Subsequently, we separated each of these networks into two networks as follows: we separated the nodes corresponding to the left hemisphere from the nodes corresponding to the right hemisphere, preserving the intrahemispheric connections in both hemispheres, disregarding the interhemispheric connections, thus obtaining two networks, one for each hemisphere. We performed this procedure for each of the three seizure phases (pre-ictal, ictal, post-ictal) in each frequency band (delta, theta, alpha).

Simplicial Characterization Measures

In order to analyze the directed higher-order connectivity and higher-order topology of the networks through simplicial characterization measures, we computed the q-digraphs, for $q = 0, 1, 2, 3, 4$, for the networks from the right and left hemispheres, for each seizure phase in each frequency band. Figures 7.4 and 7.5 present examples of q-digraphs, for $q = 0, 1$, constructed from networks of the right and left hemispheres of patients PN01 and PN12, respectively.

Due to the limited nature of this study, we chose nine among all simplicial measures introduced in Chapter 4. We chose six global measures and three local measures, and at least one belonging to each of the five categories of simplicial measures (excluding the discrete curvatures):

- Distance-based simplicial measures: Global q-efficiency (global);
- Simplicial centralities: In- and out-q-degree centralities (local), q-harmonic centrality (local), and global q-reaching centrality (global);
- Simplicial segregation measures: Average q-clustering coefficient (global);
- Simplicial entropies: In- and out-q-degree distribution entropies (global);
- Spectrum-related simplicial measures: q-energy (global).

Bearing in mind that the order (N) of the network can impact the simplicial measures that depend on N (N-dependent), we used a fixed N (N_{max}) equal to the largest N among all q-digraphs, for $q = 0, 1, 2, 3, 4$, for all seizure phases in all frequency bands, for both hemispheres of all patients. For each simplicial measure and for each patient, we computed the mean and the standard deviation between the q-digraphs constructed from each of the 11 digraphs, for each seizure phase in each frequency band and for both hemispheres. For the local simplicial measures, we computed the mean of the maximum values obtained among all nodes.

Furthermore, we applied the usual graph measures presented in Chapter 2 corresponding to their simplicial analogues exposed above in the iPDC networks of the right and left hemispheres, following the same procedure described above for the simplicial measures.

Both the construction of the q-digraphs and the computation of the simplicial and graph measures were carried out through the Python package `DigplexQ` (see Appendix A).

Simplicial Similarity Comparison Distances

For the simplicial distances, we only considered the iPDC networks of the right and left hemispheres, i.e. we did not take into account their q-digraphs. As in the case of the simplicial characterization measures, due to the limited nature of this study, we chose eight simplicial distances, namely: bottleneck distance, Wasserstein distance, Betti distance, 1st, 4th, and 5th topological structure distances, histogram cosine kernel, and Jaccard kernel. Also, for the bottleneck, Wasserstein, Betti, 4th, and 5th topological structure distances we only considered the 0-th Betti numbers, and for the histogram cosine kernel and Jaccard kernel we considered the directed flag complexes associated with the networks.

Let d be any of the eight chosen simplicial distances. Given two digraphs, G_1 and G_2, let's denote by $d(G_1, G_2)$ the value produced by the distance d in the corresponding structures obtained from these two digraphs, for example, if d is the bottleneck distance, then $d(G_1, G_2) = d_{W_\infty}(P_1, P_2)$, where P_1 and P_2 are the persistent diagrams associated with G_1 and G_2, respectively, and if d is the Jaccard kernel, then $d(G_1, G_2) = K_J(\mathcal{X}_1, \mathcal{X}_2)$, where $\mathcal{X}_1$ and $\mathcal{X}_2$ are the directed flag complexes of G_1 and G_2, respectively. For each of the simplicial distances, we produced different distributions, for each patient, according to the cerebral hemisphere and seizure phase, for every frequency band:

- Distributions $d(G_{ic}^R, G_{ic}^R)$ and $d(G_{ic}^L, G_{ic}^R)$, where G_{ic}^R are randomly chosen digraphs from the right hemisphere and G_{ic}^L are randomly chosen digraphs from the left hemisphere corresponding to the ictal phase (without repetition).

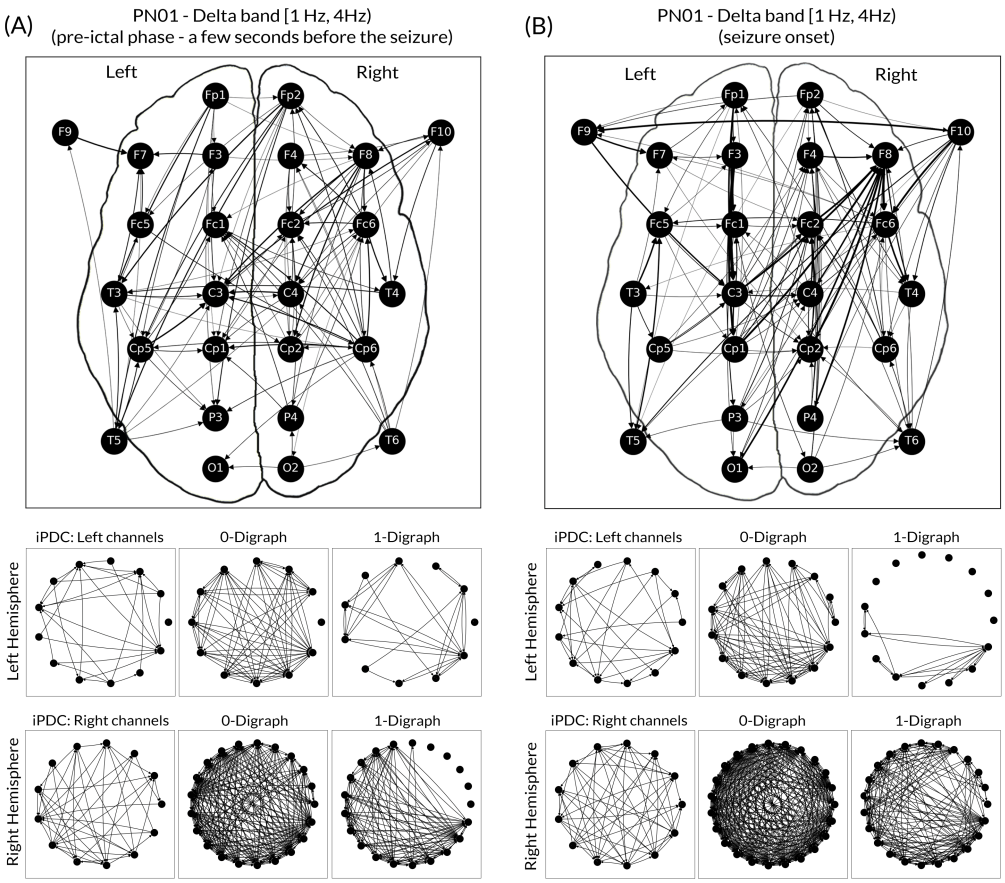


Figure 7.4: iPDC networks (arcs thicknesses are proportional to the weights) of patient PN01 computed in a 10s interval a few seconds before the seizure (A) and in a 10s interval starting on seizure onset (B) in the delta band, together with the (weightless) q-digraphs corresponding to the right and left hemispheres networks (considering only intrahemispheric connections) for $q = 0, 1$.

- Distributions $d(G^L_{pre}, G^L_{pre})$ and $d(G^L_{pre}, G^L_{ic})$, where G^L_{pre} are randomly chosen digraphs from the left hemisphere corresponding to the pre-ictal phase, and G^L_{ic} are randomly chosen digraphs from the left hemisphere corresponding to the ictal phase (without repetition).

- Distributions $d(G^L_{pos}, G^L_{pos})$ and $d(G^L_{ic}, G^L_{pos})$, where G^L_{pos} are randomly chosen digraphs from the left hemisphere corresponding to the post-ictal phase, and G^L_{ic} are randomly chosen digraphs from the left hemisphere corresponding to the ictal phase (without repetition).

- Distributions $d(G^R_{pre}, G^R_{pre})$ and $d(G^R_{pre}, G^R_{ic})$, where G^R_{pre} are randomly chosen digraphs from the right hemisphere corresponding to the pre-ictal phase, and G^R_{ic} are randomly chosen digraphs from the right hemisphere corresponding to the ictal phase (without repetition).

- Distributions $d(G^R_{pos}, G^R_{pos})$ and $d(G^R_{ic}, G^R_{pos})$, where G^R_{pos} are randomly chosen di-

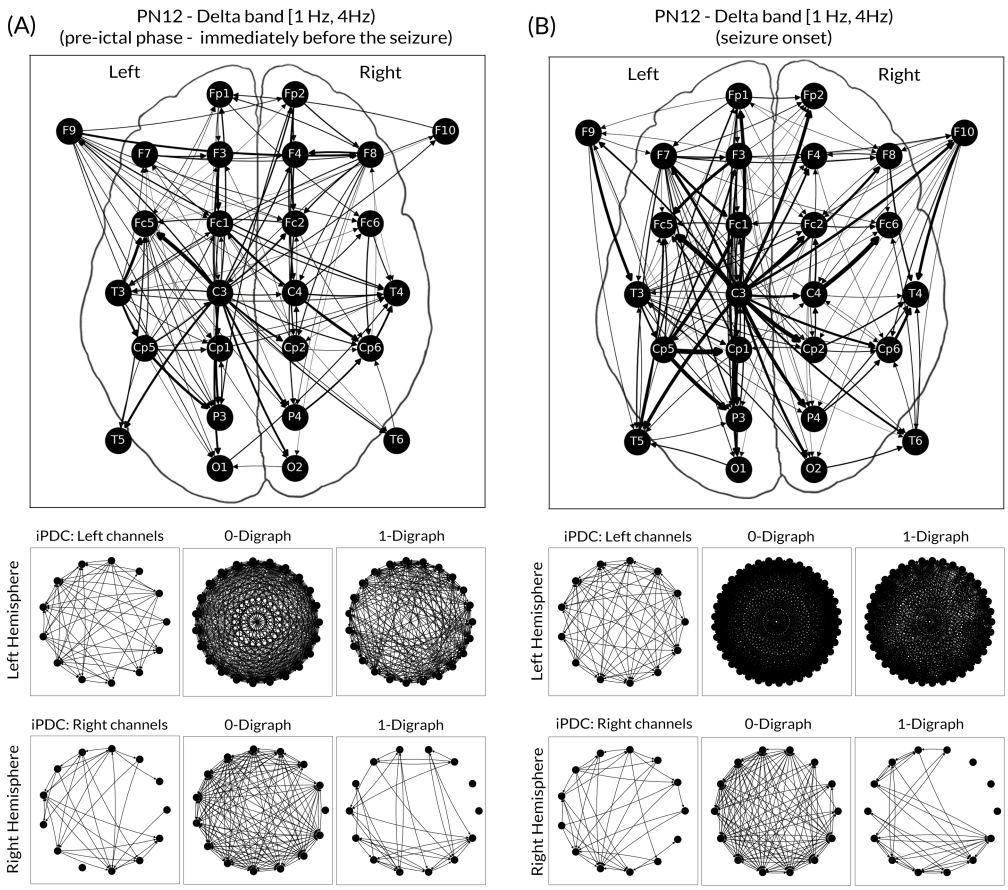


Figure 7.5: iPDC networks (arcs thicknesses are proportional to the weights) of patient PN12 computed in a 10s interval immediately before the seizure (A) and in a 10s interval starting on seizure onset (B) in the delta band, together with the (weightless) q-digraphs corresponding to the right and left hemispheres networks (considering only intrahemispheric connections) for $q = 0, 1$.

graphs from the right hemisphere corresponding to the post-ictal phase, and G_{ic}^{R} are randomly chosen digraphs from the right hemisphere corresponding to the ictal phase (without repetition).

Afterwards, we computed the means and standard deviations for each distribution, in each frequency band, for each patient.

The computation of the simplicial distances were carried out through the Python package `DigplexQ` (see Appendix A).

Statistical Analysis

Finally, we performed the following statistical analysis:

- For each simplicial measure, Wilcoxon paired tests, at a significance level $\alpha = 0.05$, were performed to verify the statistical differences between the simplicial measure

of the pre-ictal phase and ictal phase as well as of the ictal phase and post-ictal phase, in each frequency band (delta, theta, alpha) and at each level $q = -1, 0, 1, 2, 3, 4$ (where $q = -1$ represents the original networks), for the right and left hemispheres separately. In addition, we computed the Pearson's correlation between the simplicial measures across the levels $q = -1, 0, 1, 2, 3, 4$ for each seizure phase (pre-ictal, ictal, and post-ictal), and also across the seizure phases at each level $q = -1, 0, 1, 2, 3, 4$, both for each frequency band and for each cerebral hemisphere.

- For each simplicial distance, Wilcoxon paired tests, at a significance level $\alpha = 0.05$, were performed to verify the statistical differences between the means of the following distributions: $d(G^R_{ic}, G^R_{ic})$ and $d(G^L_{ic}, G^R_{ic})$ (to verify the differences between the left and the right hemispheres in the ictal phase); $d(G^L_{pre}, G^L_{pre})$ and $d(G^L_{pre}, G^L_{ic})$ (to verify the differences between the pre-ictal phase and the ictal phase in the left hemisphere); $d(G^L_{pos}, G^L_{pos})$ and $d(G^L_{ic}, G^L_{pos})$ (to verify the differences between the ictal phase and the post-ictal phase in the left hemisphere); $d(G^R_{pre}, G^R_{pre})$ and $d(G^R_{pre}, G^R_{ic})$ (to verify the differences between the pre-ictal phase and the ictal phase in the right hemisphere); and $d(G^R_{pos}, G^R_{pos})$ and $d(G^R_{ic}, G^R_{pos})$ (to verify the differences between the ictal phase and the post-ictal phase in the right hemisphere), for each frequency band.

The statistical analysis was carried out through the Python statistical package `Pingouin` (see Appendix A).

Figure 7.6 summarizes the analysis work flow, including the preprocessing step.

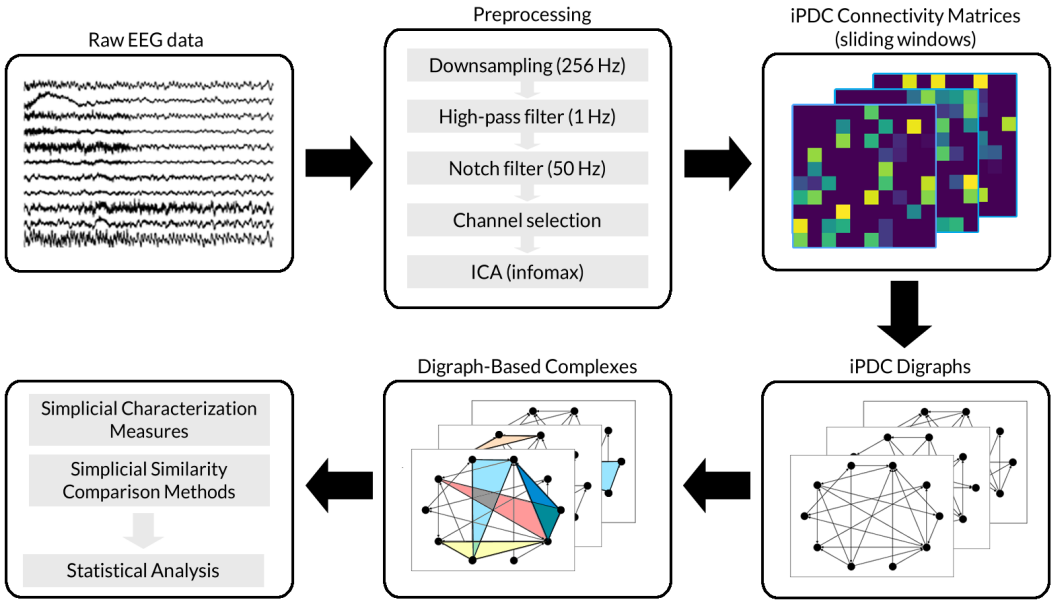


Figure 7.6: Summary of the analysis workflow.

7.2.2 Results and Discussion

Simplicial Characterization Measures

All results discussed in this part can be found in Appendix C. Figure C.1 through C.12 present the grand means (mean of all of the means) and standard deviations of the simplicial measures; Table C.1 through C.5 present the W-statistics and the p-values of the Wilcoxon paired tests; and Figure C.13 through C.21 present the Pearson's correlation coefficients between the simplicial measures. These tables and figures present the results for the original iPDC networks (henceforth referred to as (-1)-digraphs or level $q = -1$) and for the q-digraphs, $q = 0, 1, 2, 3, 4$, for the right and left hemispheres, for each seizure phase (pre-ictal, ictal, post-ictal), and for each frequency band (delta, theta, alpha).

In the following, we present and discuss only statistically significant results ($p < 0.05$) and, for the Pearson's correlation coefficient r, we consider two measures to be strongly positively correlated if $r \geq 0.7$ and strongly negatively correlated if $r \leq -0.7$ ($p < 0.05$). Also, for the sake of simplicity, we refer to directed cliques simply as cliques, we omit the term "directed simplicial" from the nomenclature of the measures, we use the notations δ, θ, and α to indicate the delta, theta, and alpha bands, respectively, and we adopt LH and RH to indicate the left hemisphere and the right hemisphere, respectively.

Global q-efficiency: The global q-efficiency increased significantly in the ictal phase compared with the pre-ictal phase in the δ band at levels $q = -1, 0, 1, 2$ in both hemispheres and at level $q = 3$ in the LH, and in θ band at levels $q = 2, 3$ in the LH and at level $q = -1$ in the RH. Also, this measure is strongly positively correlated with most of the measures across all seizure phases at all levels q, and across all levels q in all seizure phases, for all frequencies, and in both hemispheres. These results suggest that the efficiency of transmitting information at a global level in each hemisphere increases in the ictal phase, not only at the level of nodes but also at the level of cliques, specifically in the δ band.

In-q-degree centrality: The in-q-degree centrality increased significantly in the ictal phase compared with the pre-ictal phase in the δ band at levels $q = 0, 1, 2, 3$ in the LH and at levels $q = 1, 2$ in the RH, and in θ band at levels $q = 2, 3$ in the LH and at levels $q = 1, 2, 3$ in the RH. Also, this measure is strongly negatively correlated with most of the measures in relation to all seizure phases at level $q = -1$ in the LH and it is strongly positively correlated with most of the measures across all seizure phases at all other levels q, and across all levels q in all seizure phases, for all frequencies and for both hemispheres. These results suggest that the importance of nodes in network communication, in relation to the inner flow, did not change significantly in the ictal phase compared to the pre-ictal phase, however there was a significant increase at higher

levels, that is, the importance of cliques in higher-order communication increased in the ictal phase, in both hemispheres, in both δ and θ bands.

Out-q-degree centrality: The out-q-degree centrality increased significantly in the ictal phase compared with the pre-ictal phase in the δ band at levels $q = 0, 1, 2, 3$ in the LH and at levels $q = 1, 2$ in the RH, and in θ band at levels $q = 2, 3$ in the LH and at levels $q = 1, 2, 3$ in the RH. Also, this measure is strongly negatively correlated with most of the measures in relation to all seizure phases at level $q = -1$ in the LH and it is strongly positively correlated with most of the measures across to all seizure phases at all other levels q, and across to all levels q in all seizure phases, for all frequencies and for both hemispheres. These results suggest that the importance of nodes in network communication, in relation to the outer flow, did not change significantly in the ictal phase compared to the pre-ictal phase, however, there was a significant increase at higher levels, that is, the importance of cliques in higher-order communication increased in the ictal phase, in both hemispheres, in both δ and θ bands.

q-Harmonic centrality: The q-harmonic centrality increased significantly in the ictal phase compared with the pre-ictal phase in the δ band at levels $q = 0, 1, 2, 3$ in the LH and at levels $q = -1, 0, 2$ in the RH, and in θ band at levels $q = 2, 3$ in the LH and at levels $q = -1, 2$ in the RH. Also, this measure is strongly positively correlated with most of the measures across all seizure phases at all levels q, for all frequencies, and for both hemispheres. These results suggest that, in the ictal phase, higher-order cliques transmit information more efficiently, in both hemispheres, especially in the δ band, on the other hand, nodes transmit information more efficiently in the RH, in both δ and θ bands.

Global q-reaching centrality: The global q-reaching centrality increased significantly in the ictal phase compared with the pre-ictal phase in the δ band at levels $q = 0, 1, 2, 3$ in the LH and at levels $q = 0, 2, 3$ in the RH, and in θ band at level $q = 2$ in the LH and at levels $q = 0, 1, 2$ in the RH. Also, this measure is strongly negatively correlated with most of the measures in relation to all seizure phases at level $q = -1$ in both hemispheres and it is strongly positively correlated with most of the measures across all seizure phases at all other levels q, and across all levels q in all seizure phases, for all frequencies and for both hemispheres. These results suggest that, in the ictal phase, higher-order cliques have more influence in the information flow, in both hemispheres, especially in the δ band.

Average q-clustering coefficient: The average q-clustering coefficient increased significantly in the ictal phase compared with the pre-ictal phase in the δ band at levels $q = -1, 0, 1, 2$ in the LH and at levels $q = 1, 2$ in the RH, and in θ band at levels $q = 1, 2, 3$ in the LH and at levels $q = -1, 1, 2, 3$ in the RH. Also, this measure is

strongly positively correlated with most of the measures across all seizure phases at all levels q, and across all levels q in all seizure phases, for all frequencies and for both hemispheres. These results suggest that, in the ictal phase, cliques tend to form higher-order clusters, in both hemispheres, in both δ and θ bands.

In-q-degree distribution entropy: The in-q-degree distribution entropy increased significantly in the ictal phase compared with the pre-ictal phase in the δ band at levels $q = 1, 2, 3$ in the LH and at levels $q = -1, 0, 2, 3$ in the RH, and in θ band at level $q = 2$ in the LH and at level $q = -1, 1, 2, 3$ in the RH. Also, this measure is strongly positively correlated with most of the measures in relation to all seizure phases at all levels q and across all levels q in all seizure phases, for all frequencies and for both hemispheres. These results suggest that, in the ictal phase, the higher-order networks, in relation to their inner higher-order flow, become "more random" (or "less regular"), i.e. they get closer to a random model, in both hemispheres, in both δ and θ bands.

Out-q-degree distribution entropy: The out-q-degree distribution entropy increased significantly in the ictal phase compared with the pre-ictal phase in the δ band at levels $q = 1, 2, 3$ in the LH and at levels $q = -1, 0, 2, 3$ in the RH, and in θ band at level $q = 2$ in the LH and at level $q = -1, 1, 2, 3$ in the RH. Also, this measure is strongly positively correlated with most of the measures in relation to all seizure phases at all levels q and across all levels q in all seizure phases, for all frequencies and for both hemispheres. These results suggest that, in the ictal phase, the higher-order networks, in relation to their outer higher-order flow, become "more random" (or "less regular"), i.e. they get closer to a random model, in both hemispheres, in both δ and θ bands.

q-Energy: The q-energy increased significantly in the ictal phase compared with the pre-ictal phase in the δ band at levels $q = 0, 1, 2, 3$ in the LH and at levels $q = -1, 0, 2, 3$ in the RH, and in θ band at levels $q = 2, 3$ in the LH and at levels $q = -1, 3$ in the RH. Also, this measure is strongly positively correlated with most of the measures across all seizure phases at all levels q, for all frequency bands and for both hemispheres. These results suggest that, in the ictal phase, the higher-order networks become more connected, in both hemispheres, in both δ and θ bands.

In short, all simplicial characterization measures showed an increase from the pre-ictal phase to the ical phase, for various levels of topological organization q, in both hemispheres, suggesting an increase in clustering, efficiency, and connectivity, and a shift towards a "more random" organization (both in relation to inner and outer flow) in higher-order networks, especially in the δ and θ bands. However, here a caveat is necessary, because while the directed clustering coefficient takes into account the total degree of the nodes, the in and out entropies only take into account the in and out degrees, respectively, therefore not necessarily an increase in the directed clustering coefficient will lead to a decrease in some of these entropies.

Simplicial Similarity Comparison Distances

All results discussed in this part can be found in Appendix C. Table C.6 through C.9 present the W-statistics and the p-values of the Wilcoxon paired tests. These tables present the results for each frequency band (delta, theta, alpha).

In the following, we present and discuss only statistically significant results (p-values < 0.05). Also, just as before, we refer to directed cliques simply as cliques, and we use the notations δ, θ, and α to indicate the delta, theta, and alpha bands, respectively, and LH and RH to indicate the left hemisphere and the right hemisphere, respectively.

Bottleneck distance: The bottleneck distance showed statistically significant difference between the pre-ictal phase and the ictal phase in the RH, in the θ band.

Wasserstein distance: The Wasserstein distance showed statistically significant difference between the pre-ictal phase and the ictal phase in both hemispheres, in the θ band, and between the ictal phase and the post-ictal phase in both hemispheres, but in the θ band in the LH and in the α band in the RH.

Betti distance: The Betti distance showed a statistically significant difference between the pre-ictal phase and the ictal phase in both hemispheres, in the θ band.

First topological distance: The first topological distance showed a statistically significant difference between the pre-ictal phase and the ictal phase in the LH, in the θ band, and between the ictal phase and the post-ictal phase in the RH, in the δ, θ, and α bands.

Fifth topological distance: The fifth topological distance showed a statistically significant difference between the ictal phase and the post-ictal phase in LH, in the θ band, and in the RH, in the δ band.

Histogram cosine kernel: The histogram cosine kernel showed a statistically significant difference between the ictal phase and the post-ictal phase in the RH, in the δ, θ, and α bands.

Jaccard kernel: The Jaccard kernel showed statistically significant difference between the pre-ictal phase and the ictal phase and between the ictal phase and the post-ictal phase in both hemispheres, in the δ, θ, and α bands.

All these results suggest that there is a change in the clique topology of the networks of both hemispheres, both from the pre-ictal phase to the ictal phase and from the ictal phase to the post-ictal phase, especially in the θ band. Moreover, we highlight that no statistically significant difference was observed between the left hemispheres and the right hemisphere in the ictal phase for any distance, at any frequency.

7.3 Conclusions

Despite all the efforts of the scientific community, our understanding of the dynamic processes underlying epilepsy still has many gaps. In this study we constructed G-connectivity networks from EEG data recorded before, during and after epileptic seizures in patients diagnosed with left temporal lobe epilepsy, obtaining therefore an estimate of the directionality and dynamics of the information flow between different brain regions in various seizure phases; subsequently, we applied novel simplicial characterization measures on the q-digraphs constructed from these networks, in addition to their usual graph counterparts in the original networks, to investigate: 1) How do the topological and functional properties of G-connectivity networks and their respective q-digraphs change during the seizure in each hemisphere and in each frequency band, both at the node level and at the various clique topology levels? 2) Does the analysis of higher-order structures of brain connectivity networks provide novel and better biomarkers for seizure dynamics and also for the laterality of the seizure focus than the usual theoretical graph analyses?

We observed that all simpicial characterization measures showed statistically significant increases in their magnitudes from the pre-ictal phase to the ictal phase, for various levels q, for both hemispheres, especially in the δ and θ bands but no statistically significant changes were observed from the ictal phase to the post-ictal phase, which may suggest that several topological and functional aspects of the brain networks change from the pre-ictal phase to the ictal phase, at various higher-order levels of topological organization (clique organization), especially in the δ and θ bands. However, most of the usual graph measures did not detect significant differences in the left hemisphere, which may suggest that changes in the network topology at the node level ($q = -1$) do not undergo as many changes as in the higher-order network topology. Furthermore, most of the simplicial measures are strongly positively correlated across the seizure phases (at all levels q) and across the levels q (in all seizure phases), for all frequency bands and for both hemispheres. Also, most of the simplicial distances revealed changes in the clique topology of the networks of both hemispheres, from the pre-ictal phase to the ictal phase, especially in the θ band, which reinforces the findings obtained by the simplicial characterization measures. Regarding the laterality of the seizure focus, the analysis through simplicial distances did not find any statistically significant difference between the left and right hemispheres clique topology in the ictal phase.

In conclusion, despite our several limitations, such as the limited number of patients (eight patients) and all the limitations associated with each method used in the study, we emphasize that we found evidence that the analysis of higher-order structures represented by q-digraphs obtained from G-connectivity networks may be a reliable way to find biomarkers associated with epileptic networks, but its establishment as a viable rigorous method will depend on future work, as well as its applicability to other

disorders of brain connectivity networks.

Chapter 8

Final Considerations

Wir müssen wissen. Wir werden wissen. (We must know. We will know.)

— Epitaph on the gravestone of David Hilbert

The initial motivation for writing this thesis was the development of new methods to analyze the topology of directed graphs obtained from brain connectivity estimators based on the concept of Granger causality, especially the iPDC, to contribute to the general investigation of the dynamics of brain activity. As research progressed, we noticed the increasing application and usefulness of methods derived from computational topology in network analysis, mainly based on clique complexes. Thus, we decided to transpose several of these methods and concepts to directed graphs, by considering their directed clique complexes and path complexes, and two main objectives were outlined:

1. To develop rigorously a new quantitative theory for digraph-based complexes (or, as we can consider it, a step towards the formalization of a "quantitative simplicial theory"), with special emphasis on directed higher-order connectivity between directed cliques;
2. To apply the methods of the new theory to epileptic brain networks obtained through iPDC to quantitatively investigate their higher-order topologies and search for new biomarkers based on their directed higher-order connectivities, thus pointing out potential applications of the theory in network neuroscience.

To accomplish these objectives, we divided the thesis into two parts, Part I, to carry out objective 1, and Part II, to carry out objective 2, whose developments are summarized in the following paragraphs.

In Part I, having considered the multidisciplinary nature of this thesis and to make it self-contained, we first presented, in Chapter 2, all the fundamental and necessary concepts of graph theory, starting with a discussion of binary and equivalence relations, followed by the introduction of concepts associated with graphs and digraphs, passing

through algebraic and spectral graph theory, graph measures, graph similarities, and at last a brief discussion on random graphs. Subsequently, in Chapter 3, we presented the basic theory of simplicial complexes and directed clique complexes associated with digraphs, including the case of weighted digraphs, passing through simplicial homology, persistent homology, and combinatorial Laplacians associated with these complexes, followed by the presentation of paths complexes and their homologies, and, at last, we introduced a novel theory associated with directed higher-order connectivity between directed cliques, which provided the conception of new concepts such as directed higher-order adjacencies (upper and lower adjacencies) and maximal q-digraphs, and new concepts for directed Q-Analysis. Finally, in Chapter 4, we introduced new characterization measures for maximal q-digraphs, adapted from graph measures, such as distance-based measures, segregation measures, centrality measures, entropy measures, and spectrum-related measures, and similarity comparison methods for directed clique complexes and path complexes, based on graph similarity comparison methods, such as structure distances and graph kernels, and, at last, we presented some examples with random digraph models.

In Part II, we started by presenting, in Chapter 5, the theory of brain connectivity networks, discussing briefly the biophysical principles of brain signals and the techniques for acquiring such signals, with special attention to EEG, followed by a discussion of connectivity estimators, with an emphasis on PDC and its variants, especially gPDC and iPDC, and, finally, we discussed the different types of brain connectivity networks, especially structural, functional, and effective networks, and also about the applications of graph theory and computational (algebraic) topology in the analysis of these networks to investigate the dynamics of brain activity in different contexts. In Chapter 6, we studied the neuropathology of epilepsy in more detail, discussing its main characteristics, etiologies, epidemics, diagnoses, and treatments, and also discussed several studies pointing out alterations in the network properties of epileptic brain networks when compared to brain networks obtained from healthy individuals, as well as changes during the ictal period compared to other periods. Finally, in Chapter 7, we performed an analysis of epileptic brain networks, estimated through iPDC from EEG data from patients diagnosed with left temporal lobe epilepsy, using the novel quantitative methods for directed clique complexes developed in previous chapters, to investigate how certain properties of these networks, and their corresponding higher-order structures (q-digraphs), alter according to the phases of seizures, and also according to the cerebral hemispheres, in different frequency bands.

Furthermore, we developed the Python package `DigplexQ` (see Appendix A), which contains the implementation of algorithms for calculating directed clique complexes and path complexes of some given digraph from its adjacency matrix, and also contains the implementation of various simplicial characterization measures and simplicial similarity comparison methods introduced in Chapter 4.

Due to the limited nature of this thesis, many of the simplicial characterization measures and simplicial similarity comparison methods have not been applied to real-world data, and a comparison between the weighted and unweighted versions of these measures and methods has also been lacking. Moreover, the applicability of concepts developed to path complexes has not been explored. In addition to these gaps, there are several other opportunities to be explored in future work regarding the theory involving directed higher-order connectivity and directed Q-Analysis, not only in relation to its mathematical foundations but also in relation to its applicability in several areas, besides network neuroscience, such as biology, social sciences, transportation planning, etc.

Bibliography

[1] F. Abdelnour, M. Dayan, O. Devinsky, T. Thesen, and A. Raj. Estimating brain's functional graph from the structural graph's laplacian. *Proceedings of SPIE*, 9597:176–180, 2015.

[2] F. Abdelnour, M. Dayan, O. Devinsky, T. Thesen, and A. Raj. Functional brain connectivity is predictable from anatomic network's laplacian eigen-structure. *NeuroImage*, 172:728–739, 2018.

[3] S. Achard and E. Bullmore. Efficiency and cost of economical brain functional networks. *PLoS Comput Biol*, 3(2):e17, 2007.

[4] R. Aharoni, E. Berger, and R. Meshulam. Eigenvalues and homology of flag complexes and vector representations of graphs. *Geom. Funct. Anal.*, 15:555–566, 2005.

[5] M. Ahmadlou, H. Adeli, and A. Adeli. Graph theoretical analysis of organization of functional brain networks in adhd. *Clinical EEG and neuroscience*, 43(1):5–13, 2012.

[6] G. Alarcón and A. Valentín. *Introduction to Epilepsy.* Cambridge University Press., 2012.

[7] M. Andjelković, N. Gupte, and B. Tadić. Hidden geometry of traffic jamming. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 91(5):052817, 2015.

[8] M. Andjelković, B. Tadić, and R. Melnik. The topology of higher-order complexes associated with brain-function hubs in human connectomes. *arXiv: Neurons and Cognition*, 2020.

[9] G. Arizmendi and O. Arizmendi. Energy and randić index of directed graphs. *Linear and Multilinear Algebra*, 71:2696–2707, 2021.

[10] R. H. Atkin. An algebra for patterns on a complex. i. *Internat. J. Man-Machine Stud.*, 6:285–307, 1974.

[11] R. H. Atkin. *Mathematical structure in human affairs*. Heinemann, London, 1974.

[12] R. H. Atkin. An algebra for patterns on a complex. ii. *Internat. J. Man-Machine Stud.*, 8:483–448, 1976.

[13] R. H. Atkin. *Combinatorial Connectivities in social systems: An Application of Simplicial Complex Structures to the Study of Large Organizations*. Birkhäuser Basel, 1977.

[14] L. A. Baccalá, M. Y. Alvarenga, K. Sameshima, C. L. Jorge, and L. H. Castro. Graph theoretical characterization and tracking of the effective neural connectivity during episodes of mesial temporal epileptic seizure. *Journal of Integrative Neuroscience*, 3(4):379–395, 2004.

[15] L. A. Baccalá, C. S. de Brito, D. Y. Takahashi, and K. Sameshima. Unified asymptotic theory for all partial directed coherence forms. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 371(1997):20120158, 2013.

[16] L. A. Baccalá and K. Sameshima. Overcoming the limitations of correlation analysis for many simultaneously processed neural structures. 130:33–47, 2001.

[17] L. A. Baccalá and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics*, 84(6):463–474, 2001.

[18] L. A. Baccalá and K. Sameshima. Partial directed coherence: Some estimation issues. *In World Congress on Neuroinformatics, Edited by: Rattay, F., Vienna: ARGESIM/ASIM Verlag*, page 546–553, 2001.

[19] L. A. Baccalá and K. Sameshima. Causality and influentiability: The need for distinct neural connectivity concepts. *In: Slezak D., Tan AH., Peters J.F., Schwabe L. (eds) Brain Informatics and Health. BIH 2014. Lecture Notes in Computer Science*, 8609:424–435, 2014.

[20] L. A. Baccalá and K. Sameshima. Frequency domain repercussions of instantaneous granger causality. *Entropy (Basel, Switzerland)*, 23(8):1037, 2021.

[21] L. A. Baccalá and K. Sameshima. Partial directed coherence and the vector autoregressive modelling myth and a caveat. *Front. Netw. Physiol.*, 2:845327, 2022.

[22] L. A. Baccalá, K. Sameshima, and D. Y. Takahashi. Generalized partial directed coherence. *15th International Conference on Digital Signal Processing*, pages 163–166, 2007.

[23] F. Baccini, F. Geraci, and G. Bianconi. Weighted simplicial complexes and their representation power of higher-order network data and topology. *Phys. Rev. E*, 106:034319, 2022.

[24] M. Bandarabadi, J. Rasekhi, C. A. Teixeira, M. R. Karami, and A. Dourado. On the proper selection of preictal period for seizure prediction. *Epilepsy & behavior : E& B*, 46:158–166, 2015.

[25] J. Bang-Jensen and G. Z. Gutin. *Digraphs: Theory, Algorithms and Applications*. Springer-Verlag, London, 2nd edition, 2009.

[26] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[27] H. Barcelo, X. Kramer, R. Laubenbacher, and C. Weaver. Foundations of a connectivity theory for simplicial complexes. *Adv. Appl. Math.*, 26:97–128, 2001.

[28] D. S. Bassett, E. Bullmore, B. A. Verchinski, V. S. Mattay, D. R. Weinberger, and A. Meyer-Lindenberg. Hierarchical organization of human cortical networks in health and schizophrenia. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 28(37):9239–9248, 2008.

[29] D. S. Bassett and E. T. Bullmore. Small-world brain networks revisited. *The Neuroscientist: a review journal bringing neurobiology, neurology and psychiatry*, 23(5):499–516, 2017.

[30] A. Bavelas. Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*, 20:725–730, 1950.

[31] L. W. Beineke and R. J. (Eds.) Wilson. *Topics in Algebraic Graph Theory*. Cambridge University Press, Cambridge, 2004.

[32] B. C. Bernhardt, Z. Chen, Y. He, A. C. Evans, and N. Bernasconi. Graph-theoretical analysis reveals disrupted small-world organization of cortical thickness correlation networks in temporal lobe epilepsy. *Cerebral cortex (New York, N.Y. : 1991)*, 21(9):2147–2157, 2011.

[33] N. Biggs. *Algebraic Graph Theory*. Cambridge University Press, Cambridge, 1993.

[34] W. D. Blizard. The development of multiset theory. *Modern Logic*, 1:319–352, 1991.

[35] P. Boldi and S. Vigna. Axioms for centrality. *Internet Mathematics*, 10:222–262, 2014.

[36] B. Bollobás. *Modern Graph Theory*. Graduate Texts in Mathematics, Springer-Verlag, New York, 1998.

[37] B. Bollobás. *Random Graphs*. Cambridge University Press, 2nd edition, 2001.

[38] B. Bollobás, C. Borgs, J.T. Chayes, and O. Riordan. Directed scale-free graphs. *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, page 132–139, 2003.

[39] P. F. Bonacich. Power and centrality: A family of measures. *Am. J. Sociol.*, 92:1170–1182, 1987.

[40] L. Bonilha, T. Nesland, G. U. Martz, J. E. Joseph, M. V. Spampinato, J. C. Edwards, and et al. Medial temporal lobe epilepsy is associated with neuronal fibre loss and paradoxical increase in structural connectivity of limbic structures. *J. Neurol. Neurosurg. Psychiatr.*, 83:903–909, 2012.

[41] M. Borgwardt. *Graph kernels*. PhD thesis, Ludwig-Maximilians-Universität München, Fakultät für Mathematik, Informatik und Statistik, 2007.

[42] G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

[43] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler. Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by granger causality. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9849–9854, 2004.

[44] M. Brunato, H. H. Hoos, and R. Battiti. On effectively finding maximal quasi-cliques in graphs. *In: Maniezzo V., Battiti R., Watson JP. (eds) Learning and Intelligent Optimization. LION 2007. Lecture Notes in Computer Science*, 5313:41–55, 2008.

[45] C. Brunner, M. Billinger, M. Seeber, T. R. Mullen, and S. Makeig. Volume conduction influences scalp-based connectivity estimates. *Frontiers in computational neuroscience*, 10:121, 2016.

[46] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3):186–198, 2009.

[47] E. T. Bullmore and D. S. Bassett. Brain graphs: Graphical models of the human brain connectome. *Annual Review of Clinical Psychology*, 7:113 – 140, 2011.

[48] S. Cao, M. Dehmer, and Y. Shi. Extremality of degree-based graph entropies. *Information Sciences*, 278:22–33, 2014.

[49] L. Caputi and H. Riihimäki. Hochschild homology, and a persistent approach via connectivity digraphs. *arXiv:2204.00462*, 2022.

[50] G. E. Carlsson and V. D. Silva. Zigzag persistence. *Foundations of Computational Mathematics*, 10:367–405, 2008.

[51] F. Chazal and B. Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Front. Artif. Intell.*, 4:667963, 2021.

[52] S. Chiang and Z. Haneef. Graph theory findings in the pathophysiology of temporal lobe epilepsy. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 125(7):1295–1305, 2014.

[53] G. Chiarion, L. Sparacino, Y. Antonacci, L. Faes, and L. Mesin. Connectivity analysis in eeg data: A tutorial review of the state of the art and emerging trends. *Bioengineering*, 10(3):372, 2023.

[54] S. Chowdhury and F. Mémoli. Persistent homology of directed networks. *2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 77–81, 2016.

[55] S. Chowdhury and F. Mémoli. A functorial dowker theorem and persistent homology of asymmetric networks. *Journal of Applied and Computational Topology*, 2:115–175, 2018.

[56] S. Chowdhury and F. Mémoli. Persistent path homology of directed networks. *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1152–1169, 2018.

[57] F. R. K. Chung. *Complex Graphs and Networks*. Vol. 107 of CBMS Regional Conference Series in Mathematics, AMS Bookstore, 1992.

[58] F. R. K. Chung. *Spectral graph theory*. Vol. 92 of CBMS Regional Conference Series in Mathematics, AMS Bookstore, 1992.

[59] P. S. Churchland and T. J. Sejnowski. *The computational brain*. MIT Press, 1992.

[60] B. A. Cociu, S. Das, L. Billeci, W. Jamal, K. Maharatna, S. Calderoni, A. Narzisi, and F. Muratori. Multimodal functional and structural brain connectivity analysis in autism: A preliminary integrated approach with eeg, fmri, and dti. *IEEE Transactions on Cognitive and Developmental Systems*, 10(2):213–226, 2018.

[61] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37:103–120, 2007.

[62] A. Coito, G. Plomp, M. Genetti, E. Abela, R. Wiest, M. Seeck, C. M. Michel, and S. Vulliemoz. Dynamic directed interictal connectivity in left and right temporal lobe epilepsy. *Epilepsia*, 56(2):207–217, 2015.

[63] M. Coombs, A. Jarrah, and R. Laubenbacher. Foundations of combinatorial time series analysis. *preprint*, 2000.

[64] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press and McGraw-Hill, 2nd edition, 2001.

[65] L. D. Costa, F. A. Rodrigues, G. Travieso, and P. R. Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167 – 242, 2005.

[66] T. M. Covers and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

[67] I. Covert, B. Krishnan, I. Najm, J. Zhan, M.W. Shore, J. Hixson, and M. Po. Temporal graph convolutional networks for automatic seizure detection. *Proceedings of the 4th Machine Learning for Healthcare Conference*, PMLR 106:160–180, 2019.

[68] S. Dantchev and I. Ivrissimtzis. Simplicial complex entropy. *arXiv:1603.07135*, 2016.

[69] R. J. M. Dawson. Homology of weighted simplicial complexes. *Cahiers de Topologie et Géométrie Différentielle Catégoriques*, 31(3):229–243, 1990.

[70] P. Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 2005.

[71] W. de Haan, Y. A. Pijnenburg, R. L. Strijers, Y. van der Made, W. M. van der Flier, P. Scheltens, and C. J. Stam. Functional neural network analysis in frontotemporal dementia and alzheimer's disease using eeg and graph theory. *BMC neuroscience*, 10:101, 2009.

[72] O. De la Cruz Cabrera, M. Matar, and L. Reichel. Centrality measures for node-weighted networks via line graphs and the matrix exponential. *Numer. Algor.*, 88:583–614, 2021.

[73] S. C. De Lange, M. A. de Reus, and M. P. van den Heuvel. The laplacian spectrum of neural networks. *Frontiers in Computational Neuroscience*, 7(189), 2014.

[74] M. Dehmer, F. Emmert-Streib, and Y. Shi. Quantitative graph theory: A new branch of graph theory and network science. *ArXiv:1710.05660*, 2017.

[75] M. Dehmer and A. Mowshowitz. A history of graph entropy measures. *Inf. Sci.,*, 181:57–78, 2011.

[76] M. Dehmer and F. E. Streib (eds.). *Quantitative Graph Theory: Mathematical Foundations and Applications*. CRC Press, 2014.

[77] M. Dehmer (ed.). *Structural Analysis of Complex Networks*. Birkhäuser/Springer, 2011.

[78] P. Detti, G. Vatti, and G. Zabalo Manrique de Lara. Eeg synchronization analysis for seizure prediction: A study on data of noninvasive recordings. *Processes*, 8(7):846, 2020.

[79] R. Diestel. *Graph Theory*. Graduate Texts in Mathematics, Springer-Verlag, 2017.

[80] J. Dieudonné. *A History of Algebraic and Differential Topology, 1900 - 1960*. Modern Birkhäuser Classics, Birkhäuser Boston, 1989.

[81] M. P. do Carmo. *Riemmanian Geometry*. Birkhäuser Boston, MA, 1992.

[82] C. F. Earl and J. H. Johnson. Graph theory and q-analysis. *Environment and Planning B: Planning and Design*, 8(4):367–391, 1981.

[83] H. Edelsbrunner and J. L. Harer. *Computational topology: An introduction*. American Mathematical Society, Providence, RI, 2010.

[84] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.

[85] M. Eidi and J. Jost. Ollivier ricci curvature of directed hypergraphs. *Sci. Rep.*, 10:12466, 2020.

[86] S. Eilenberg and J. A. Zilber. Semi-simplicial complexes and singular homology. *Ann. of Math.*, 51:499–513, 1950.

[87] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290, 1959.

[88] E. Estrada. *The Structure of Complex Networks: Theory and Applications*. Oxford University Press, 2011.

[89] E. Estrada and N. Hatano. Communicability graph and community structures in complex networks. *Appl. Math. Comput.*, 214:500–511, 2009.

[90] E. Estrada and N. Hatano. Returnability in complex directed networks (digraphs). *Lin. Alg. Appl.*, 430:1886–1896, 2009.

[91] E. Estrada, P. A. Knight, and P. Knight. *A first course in network theory.* Oxford University Press, 2015.

[92] E. Estrada and G. J. Ross. Centralities in simplicial complexes: Applications to protein interaction networks. *Journal of theoretical biology*, 438:46–60, 2018.

[93] G. Fagiolo. Clustering in complex directed networks. *Phys. Rev., E Stat. Nonlinear Soft Matter Phys.*, 76:026107, 2007.

[94] F. Fallani and F. Babiloni. *The Graph Theoretical Approach in Brain Functional Networks: Theory and Applications.* Morgan & Claypool, 2010.

[95] F. Fallani, L. Costa, F. A. Rodriguez, L. Astolfi, G. Vecchiato, J. Toppi, G. Borghini, F. Cincotti, D. Mattia, S. Salinari, R. Isabella, and F. Babiloni. A graph-theoretical approach in brain functional networks. possible implications in eeg studies. *Nonlinear biomedical physics*, 4 Suppl 1(Suppl 1):S8, 2010.

[96] F. V. Farahani, W. Karwowski, and N. R. Lighthall. Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Front. Neurosci.*, 13:585, 2019.

[97] B. T. Fasy, Y. Qin, B. Summa, and C. Wenk. Comparing distance metrics on vectorized persistence summaries. *Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[98] A. Fattorusso, S. Matricardi, E. Mencaroni, G. B. Dell'Isola, G. Di Cara, P. Striano, and A. Verrotti. The pharmacoresistant epilepsy: An overview on existant and new emerging therapies. *Frontiers in neurology*, 12:674483, 2021.

[99] X. Fernández and D. Mateos. Topological biomarkers for real-time detection of epileptic seizures. *arXiv:2211.02523*, 2022.

[100] R. S. Fisher, J. H. Cross, C. D'Souza, J. A. French, S. R. Haut, N. Higurashi, E. Hirsch, F. E. Jansen, L. Lagae, S. L. Moshé, J. Peltola, E. Roulet Perez, I. E. Scheffer, A. Schulze-Bonhage, E. Somerville, M. Sperling, E. M. Yacubian, and S. M. Zuberi. Instruction manual for the ilae 2017 operational classification of seizure types. *Epilepsia*, 58(4):531–542, 2017.

[101] R. S. Fisher, J. H. Cross, J. A. French, N. Higurashi, E. Hirsch, F. E. Jansen, L. Lagae, S. L. Moshé, J. Peltola, E. Roulet Perez, I. E. Scheffer, and S. M. Zuberi. Operational classification of seizure types by the international league against

epilepsy: Position paper of the ilae commission for classification and terminology. *Epilepsia*, 58(4):522–530, 2017.

[102] R. Forman. Bochner's method for cell complexes and combinatorial ricci curvature. *Discrete Comput Geom*, 29:323–374, 2003.

[103] A. Fornito, A. Zalesky, and E. T. Bullmore. *Fundamentals of Brain Network Analysis*. Academic Press, Elsevier, 2016.

[104] J. Fox, N. Samudra, M. Johnson, M. J. Humayun, and B. W. Abou-Khalil. Temporal intermittent rhythmic theta activity (tirta): A marker of epileptogenicity? *eNeurologicalSci*, 29:100433, 2022.

[105] L. C. Freeman. Centrality in social networks: conceptual clarification. *Soc. Netw.*, 1:215–239, 1978.

[106] G. Friedman. An elementary illustrated introduction to simplicial sets. *Rocky Mountain Journal of Mathematics*, 42:353–423, 2012.

[107] J. Friedman. Computing betti numbers via combinatorial laplacians. *Algorithmica*, 21:331–346, 1998.

[108] K. J. Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.*, 2:56–78, 1994.

[109] K. J. Friston. Functional and effective connectivity: a review. *Brain connectivity*, 1(1):13–36, 2011.

[110] K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *Neuroimage*, 19:1273–1302, 2003.

[111] P. Frosini. Measuring shapes by size functions. *In D. P. Casasent, editor, Intelligent Robots and Computer Vision X: Algorithms and Techniques*, 1607:122–133, 1992.

[112] F. Fröhlich. *Network Neuroscience*. Academic Press, Elsevier, 2016.

[113] A. Fujita, D. Takahashi, J. Balardin, M. Vidal, and J. Sato. Correlation between graphs with an application to brain network analysis. *Computational Statistics & Data Analysis*, 109:76–92, 2016.

[114] A. Fujita, M. C. Vidal, and D. Y. Takahashi. A statistical method to distinguish functional brain networks. *Front. Neurosci.*, 11:66, 2017.

[115] X. Gao, B. Xiao, D. Tao, and X. Li. A survey of graph edit distance. *Pattern Analysis and Applications*, 13(1):113–129, 2010.

[116] J. Geweke. Measurement of linear dependence and feedback between multiple time series. *J. Am. Stat. Assoc.*, 77:304–324, 1982.

[117] R. Ghrist. *Elementary Applied Topology*. Createspace, 1st edition, 2014.

[118] E. N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[119] C. Giusti, R. Ghrist, and D. S. Bassett. Two's company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data. *Journal of Computational Neuroscience*, 41(1):1–14, 2016.

[120] C. Giusti, E. Pastalkova, C. Curto, and V. Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460, 2015.

[121] J. Gleick. *The Information: A History, a Theory, a Flood*. Pantheon Books, 2011.

[122] T. E. Goldberg. *Combinatorial Laplacians of Simplicial Complexes*. Annandale-on-Hudson, New York, 2002.

[123] D. Govc, R. Levi, and J. P. Smith. Complexes of tournaments, directionality filtrations and persistent homology. *J. Appl. Comput. Topol.*, 5:313–337, 2021.

[124] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[125] A. Grigor'yan. Advances in path homology theory of digraphs. *Notices of the ICCM*, 10(2):61–124, 2022.

[126] A. Grigor'yan, R. Jimenez, Y.V. Muranov, and S. Yau. On the path homology theory of digraphs and eilenberg–steenrod axioms. *Homology, Homotopy and Applications*, 20(2):179–205, 2018.

[127] A. Grigor'yan, Y. Lin, Y.V. Muranov, and S. Yau. Homologies of path complexes and digraphs. *arXiv:1207.2834v4*, 2013.

[128] A. Grigor'yan, Y. Lin, Y.V. Muranov, and S. Yau. Homotopy theory for digraphs. *Pure Appl. Math. Q.*, 10(4):619–674, 2014.

[129] A. Grigor'yan, Y. Lin, Y.V. Muranov, and S. Yau. Cohomology of digraphs and (undirected) graphs. *Asian Journal of Mathematics*, 19:887–932, 2015.

[130] A. Grigor'yan, Y. Lin, Y.V. Muranov, and S. Yau. Path complexes and their homologies. *Journal of Mathematical Sciences*, 248:564–599, 2020.

[131] A. Grigor'yan, Y.V. Muranov, and S. Yau. Graphs associated with simplicial complexes. *Homology, Homotopy and Applications*, 16:295–311, 2014.

[132] A. Grigor'yan, Y.V. Muranov, and S. Yau. On a cohomology of digraphs and hochschild cohomology. *J. Homotopy Relat. Struct.*, 11:209–230., 2016.

[133] A. Grigor'yan, Y.V. Muranov, and S. Yau. Homologies of digraphs and künneth formulas. *Comm. Anal. Geom.*, 25(5):969–1018, 2017.

[134] Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology and therapeutics*, 69(3):89–95, 2001.

[135] I. Gutman. The energy of a graph. *Ber. Math.-Stat. Sekt. Forschungszent.*, 103:1–22, 1978.

[136] D. Happel. Hochschild cohomology of finite-dimensional algebras. *in: S'em. d'Algèbre Paul Dubreil et Marie-Paul Malliavin, Lect. Notes Math.*, 1404:108–126, 1989.

[137] F. Harary. *Graph Theory.* Addison–Wesley, 1969.

[138] A. Hatcher. *Algebraic Topology.* Cambridge Univ. Press, Cambridge, 2002.

[139] B. He, L. Astolfi, P. A. Valdes-Sosa, D. Marinazzo, S. Palva, C. G. Benar, C. M. Michel, and T. Koenig. Electrophysiological brain connectivity: Theory and implementation. *IEEE transactions on bio-medical engineering*, 2019.

[140] S. Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in human neuroscience*, 3:31, 2009.

[141] K. Heurling, A. Leuzy, M. Jonasson, A. Frick, E. R. Zimmer, A. Nordberg, and M. Lubberink. Quantitative positron emission tomography in brain research. *Brain Research*, 1670:220–234, 2017.

[142] M. T. Horstmann, S. Bialonski, N. Noennig, H. Mai, J. Prusseit, J. Wellmer, H. Hinrichs, and K. Lehnertz. State dependent properties of epileptic brain networks: comparative graph-theoretical analyses of simultaneously recorded eeg and meg. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 121(2):172–185, 2010.

[143] B. Horwitz. The elusive concept of brain connectivity. *Neuroimage*, 19:466–470, 2003.

[144] Y. Hu, Q. Zhang, R. Li, T. Potter, and Y. Zhang. Graph-based brain network analysis in epilepsy: an eeg study. *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER), San Francisco, CA, USA*, pages 130–133, 2019.

[145] S. A. Huettel, A. W. Song, and G. McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer Associates, 3rd edition, 2014.

[146] M. D. Humphries and K. Gurney. Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PLoS ONE*, 3:e0002051, 2008.

[147] R. M. Hutchison et al. Dynamic functional connectivity: promise, issues, and interpretation. *NeuroImage*, 80:360–378, 2013.

[148] T. Jech. *Set Theory: The Third Millennium Edition*. Springer, 2003.

[149] X. Jiang, G. B. Bian, and Z. Tian. Removal of artifacts from eeg signals: A review. *Sensors (Basel, Switzerland)*, 19(5):987, 2019.

[150] J. Johnson. *Hypernetworks in the Science of Complex Systems*. Imperial College Press, 2013.

[151] J. Jonsson. *Simplicial Complexes of Graphs*. Lecture Notes in Mathematics, Springer, 2008.

[152] D. Joyner, M. V. Nguyen, and N. Cohen. *Algorithmic Graph Theory*. 2011. http://code.google.com/p/graph-theory-algorithms-book.

[153] M. J. Kaminski and K. J. Blinowska. A new method of the description of the information flow in the brain structures. *Biological cybernetics*, 65(3):203–210, 1991.

[154] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth. *Principles of Neural Science*. McGraw Hill, 5th edition, 2014.

[155] A. P. Kartun-Giles and G. Bianconi. Beyond the clustering coefficient: A topological analysis of node neighbourhoods in complex networks. *Chaos, Solitons & Fractals: X*, 1:100004, 2019.

[156] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

[157] C. L. Keown, M. C. Datko, C. P. Chen, J. O. Maximo, A. Jahedi, and R. A. Müller. Network organization is globally atypical in autism: a graph theory study

of intrinsic functional connectivity. *Biol. Psychiatry Cogn. Neurosci. Neuroimag.*, 2:66–75, 2017.

[158] J. M. Kleinberg. Authoritative sources in a hyper-linked environment. *J. ACM*, 46:604–632, 1999.

[159] D. J. Kleitman and D. L. Wang. Algorithms for constructing graphs and digraphs with given valences and factors. *Discrete Mathematics*, 6(1):79–88, 1973.

[160] W. Klonowski. Everything you wanted to ask about eeg but were afraid to get the right answer. *Nonlinear biomedical physics*, 3(1):2, 2009.

[161] O. Knill. The energy of a simplicial complex. *Linear Algebra and its Applications*, 600(1):96–129, 2020.

[162] D. Kozlov. *Combinatorial Algebraic Topology*. Algorithms and Computation in Mathematics, vol. 21, Springer-Verlag, Berlin–Heidelberg, 2008.

[163] X. Kramer and R. Laubenbacher. Combinatorial homotopy of simplicial complexes and complex information networks. *In Applications of Computational Algebraic Geometry (D. Cox and B. Sturmfels, Eds.), Proc. Sympos. in Appl. Math., Vol. 53, Amer. Math. Soc., Providence.*, 1998.

[164] E. W. Lang, A. M. Tomé, I. R. Keck, J. M. Górriz-Sáez, and C. G. Puntonet. Brain connectivity analysis: A short survey. *Computational Intelligence and Neuroscience*, 2012:412512, 2012.

[165] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Phys. Rev. Lett.*, 87:198701, 2001.

[166] W. Leal, G. Restrepo, P. F. Stadler, and J. Jost. Forman-ricci curvature for hypergraphs. *Advances in Complex Systems*, 24(1):2150003:1–2150003:24, 2021.

[167] H. Lee, M. Chung, H. Kang, B. Kim, and D. Lee. Discriminative persistent homology of brain networks. *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 841–844, 2011.

[168] H. Lee, H. Kang, M. Chung, B. Kim, and D. Lee. Weighted functional brain network modelling via network filtration. *NIPS Workshop on Algebraic Topology and Machine Learning*, 2012.

[169] W. Liao, Z. Zhang, Z. Pan, D. Mantini, J. Ding, X. Duan, C. Luo, G. Lu, and H. Chen. Altered functional connectivity and small-world in mesial temporal lobe epilepsy. *PloS one*, 5(1):e8525, 2010.

[170] M. H. Libenson. *Practical Approach to Electroencephalography*. 1e-Saunders, 2009.

[171] Y. Lin, S. Ren, C. Wang, and J. Wu. Weighted path homology of weighted digraphs and persistence. *arXiv:1910.09891*, 2019.

[172] J. Liu, M. Li, Y. Pan, W. Lan, R. Zheng, F.-X. Wu, and J. Wang. Complex brain network analysis and its applications to brain disorders: A survey. *Complexity*, 2017:1–27, 2017.

[173] Y. Liu, H. Wang, Y. Duan, J. Huang, Z. Ren, J. Ye, and et al. Functional brain network alterations in clinically isolated syndrome and multiple sclerosis: a graph-based connectome study. *Radiology*, 282:534–541, 2017.

[174] R. D. Luce and A. D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2):95–116, 1949.

[175] D. Lütgehetmann, D. Govc, J. P. Smith, and R. Levi. Computing persistent homology of directed flag complexes. *Algorithms*, 13(1):19, 2020.

[176] H. Lütkepohl. *New introduction to multiple time series analysis*. Springer, New York, 2005.

[177] S. Maletić and M. Rajković. Combinatorial laplacian and entropy of simplicial complexes associated with complex networks. *Eur. Phys. J. Spec. Top.*, 212:77–97, 2012.

[178] S. Maletić, M. Rajković, and D. Vasiljević. Complexes of networks and their statistical properties. *In: Bubak M., van Albada G.D., Dongarra J., Sloot P.M.A. (eds) Computational Science – ICCS 2008. ICCS 2008. Lecture Notes in Computer Science, vol 5102. Springer.*, 2008.

[179] S. L. Marple. *Digital Spectral Analysis*. Dover Publications, New York, 1987.

[180] A. Martino and A. Rizzi. (hyper)graph kernels over simplicial complexes. *Entropy*, 22(10):1155, 2020.

[181] P. Masulli and A. E .P. Villa. The topology of the directed clique complex as a network invariant. *SpringerPlus*, 5(388), 2016.

[182] E. Merelli, M. Piangerelli, M. Rucco, and D. Toller. A topological approach for multivariate time series characterization: the epileptic brain. *EAI Endorsed Transactions on Self-Adaptive Systems*, 2(7):5, 2016.

[183] A. Mheich, F. Wendling, and M. Hassan. Brain network similarity: methods and applications. *Network neuroscience (Cambridge, Mass.)*, 4(3):507–527, 2020.

[184] T. A. Milligan. Epilepsy: A clinical overview. *The American journal of medicin*, 134(7):840–847, 2021.

[185] J. Milnor. The geometric realization of a semi-simplicial complex. *Annals of Mathematics*, 65(2):357–362, 1957.

[186] R. Milo, S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[187] M. Mitchell. *Complexity: A Guided Tour*. Oxford University Press, Oxford, UK, 2009.

[188] E. Mones, L. Vicsek, and T. Vicsek. Hierarchy measure for complex networks. *PLoS ONE*, 7(3):e33799, 2012.

[189] M. A. Montenegro, F. Cendes, M. M. Guerreiro, and C. A. M. Guerreiro. *EEG na prática clínica*. Editora Revinter, 2nd edition, 2012.

[190] A. Mowshowitz. Entropy and the complexity of the graphs i: An index of the relative complexity of a graph. *Bull. Math. Biophys.*, 30:175–204, 1968.

[191] P. Mukherjee, J. I. Berman, S. W. Chung, C. P. Hess, and R. G. Henry. Diffusion tensor mr imaging and fiber tractography: theoretic underpinnings. *AJNR Am. J. Neuroradiol.*, 29:632–641, 2008.

[192] J. R. Munkres. *Elements of Algebraic Topology*. CRC Press, 1993.

[193] J. V. Nadler and D. D. Spencer. What is a seizure focus? *Adv Exp Med Biol.*, 813:55–62, 2014.

[194] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[195] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, New York, NY, 2010.

[196] V. Nikiforov. Beyond graph energy: norms of graphs and matrices. *Linear Algebra and its Applications*, 506:82–138, 2016.

[197] A. Omidvarnia, G. Azemi, B. Boashash, O'Toole, J. M., P. B. Colditz, and S. Vanhatalo. Measuring time-varying information flow in scalp eeg signals: orthogonalized partial directed coherence. *IEEE transactions on bio-medical engineering*, 61(3):680–693, 2014.

[198] J. P. Onnela, J. Saramaki, J. Kertesz, and K. Kaski. Intensity and coherence of motifs in weighted complex networks. *Phys. Rev., E Stat. Nonlinear Soft Matter Phys.*, 71(6 Pt 2):065103, 2005.

[199] A. C. Papanicolaou. *Magnetoencephalography and Magnetic Source Imaging*. Cambridge University Press, 2009.

[200] D. Papo, J. Buldú, and S. Boccaletti. Network theory in neuroscience. *In: Jaeger, D., Jung, R. (eds) Encyclopedia of Computational Neuroscience. Springer, New York, NY.*, 2014.

[201] D. Papo, M. Zanin, J. H. Martínez, and J. M. Buldú. Beware of the small-world neuroscientist! *Front. Hum. Neurosci.*, 10, 2016.

[202] J. Parvizi and S. Kastner. Promises and limitations of human intracranial electroencephalography. *Nature neuroscience*, 21(4):474–483, 2018.

[203] F. Passerini and S. Severini. Quantifying complexity in networks: The von neumann entropy. *International Journal of Agent Technologies and Systems*, 1(4):58–67, 2009.

[204] J. Pattillo, A. Veremyev, S. Butenko, and V. Boginski. On the maximum quasi-clique problem. *Discrete Appl. Math.*, 161(1-2):244–257, 2013.

[205] W. Penfield. *The Mystery of the Mind*. Princeton University Press, 1975.

[206] E. Pereda, R. Q. Quiroga, and J. Bhattacharya. Causal influence: Nonlinear multivariate analysis of neurophysical signals. *Prog Neurobiol.*, 77(1-2):1–37, 2005.

[207] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and et al. Homological scaffolds of brain functional networks. *J. R. Soc. Interface*, 11(101):20140873, 2014.

[208] M. Piangerelli, M. Rucco, L. Tesei, and et al. Topological classifier for detecting the emergence of epileptic seizures. *BMC Res Notes*, 11:392, 2018.

[209] H. Poincaré. Analysis situs. *Journal de l'École Polytechnique*, 2(1):1–123, 1895.

[210] H. Poincaré. *La Valeur de la Science*. Paris, Flammarion, 1904.

[211] H. Poincaré. *Science et Méthode*. Paris, Flammarion, 1947.

[212] S. C. Ponten, F. Bartolomei, and C. J. Stam. Small-world networks and epilepsy: graph theoretical analysis of intracerebrally recorded mesial temporal lobe seizures. *Clin. Neurophysiol.*, 118(4):918–927, 2007.

[213] S. Qian, G. Sun, Q. Jiang, K. Liu, B. Li, M. Li, and et al. Altered topological patterns of large-scale brain functional networks during passive hyperthermia. *Brain Cogn.*, 83:121–131, 2013.

[214] M. A. Quraan, C. McCormick, M. Cohn, T. A. Valiante, and M. P. McAndrews. Altered resting state brain dynamics in temporal lobe epilepsy can be observed in spectral power, functional connectivity and graph theory metrics. *PLoS One*, 8:e68609, 2013.

[215] A. Ramezanpour and V. Karimipour. Simple models of small world networks with directed links. *Physical Review E*, 66(3):036128, 2002.

[216] M. B. Rao, A. Arivazhagan, S. Sinha, R. D. Bharath, A. Mahadevan, M. Bhat, and P. Satishchandra. Surgery for drug-resistant focal epilepsy. *Annals of Indian Academy of Neurology*, 17(Suppl 1):S124–S131, 2014.

[217] N. Rashevsky. Life, information theory, and topology. *Bulletin of Mathematical Biophysics*, 17:229–235, 1955.

[218] M. W. Reimann, M. Nolte, M. Scolamiero, K. Turner, R. Perin, G. Chindemi, P. Dłotko, R. Levi, K. Hess, and H. Markram. Cliques of neurons bound into cavities provide a missing link between structure and function. *Frontiers in computational neuroscience*, 11:48, 2017.

[219] S. Ren, C. Wu, and J. Wu. Weighted persistent homology. *Rocky Mountain J. Math.*, 48(8):2661–2687, 2018.

[220] H. Riihimäki. Simplicial q-connectivity of directed graphs with applications to network analysis. *arxiv:2202.07307*, 2021.

[221] F. Rosenow and H. Luders. Presurgical evaluation of epilepsy patients. *Brain*, 124:1683, 2001.

[222] M. Roy, S. Schmid, and G. Trédan. Modeling and measuring graph similarity: The case for centrality distance. *arXiv:1406.5481*, 2014.

[223] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010.

[224] S. B. Rutkove. Introduction to volume conduction. *The clinical neurophysiology primer*, page 43–53, 2007.

[225] Y. Saito and H. Harashima. Tracking of information within multichannel eeg record. *Recent Advantages in EEG and EMG Data Processing*, pages 133–146, 1981.

[226] K. Sameshima and L. A. Baccalá (eds.). *Methods in Brain Connectivity Inference through Multivariate Time Series Analysis*. Boca Raton, FL, USA: CRC Press, Taylor & Francis Group, 2014.

[227] S. Sanei-Mehri, A. K. Das, and S. Tirthapura. Enumerating top-k quasi-cliques. *2018 IEEE International Conference on Big Data (Big Data)*, pages 1107–1112, 2018.

[228] E. Saucan, A. Samal, M. Weber, and J. Jost. Discrete curvatures and network analysis. *MATCH Commun. Math. Comput. Chem.*, 80:605–622, 2018.

[229] I. E. Scheffer, S. Berkovic, G. Capovilla, M. B. Connolly, J. French, L. Guilhoto, E. Hirsch, S. Jain, G. W. Mathern, S. L. Moshé, D. R. Nordli, E. Perucca, T. Tomson, S. Wiebe, Y. H. Zhang, and S. M. Zuberi. Ilae classification of the epilepsies: Position paper of the ilae commission for classification and terminology. *Epilepsia*, 58(4):512–521, 2017.

[230] A. Schlögl. A comparison of multivariate autoregressive estimators. *Signal Processing*, 86(9):2426–2429, 2006.

[231] D. L. Schomer and F. H. L. da Silva. *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Oxford University Press, 2018.

[232] J. Searle. *Minds, Brains and Science*. Harvard University Press, 1986.

[233] D. H. Serrano and D. S. Gómez. Centrality measures in simplicial complexes: Applications of topological data analysis to network science. *Applied Mathematics and Computation*, 382:125331, 2020.

[234] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[235] T. A. Shatto and E. K. Çetinkaya. Variations in graph energy: A measure for network resilience. *2017 9th International Workshop on Resilient Networks Design and Modeling (RNDM)*, pages 1–7, 2017.

[236] L. Sivakumar, M. Dehmer, and K. Varmuza. Uniquely discriminating molecular structures using novel eigenvalue-based descriptors. *Match Communications in Mathematical and in Computer Chemistry*, 67(1):147–172, 2012.

[237] A. E. Sizemore, C. Giusti, A. Kahn, J. M. Vettel, R. F. Betzel, and D. S. Bassett. Cliques and cavities in the human connectome. *Journal of Computational Neuroscience*, 44(1):115–145, 2018.

[238] D. Smilkov and L. Kocarev. Rich-club and page-club coefficients for directed graphs. *ArXiv, abs/1103.2264*, 2010.

[239] O. Sporns. Brain connectivity. *Scholarpedia*, 2(10):4695, 2007.

[240] O. Sporns. *Networks of the Brain*. MIT Press, 2010.

[241] O. Sporns. *Discovering the Human Connectome*. MIT Press, 2012.

[242] O. Sporns. The human connectome: Origins and challenges. *NeuroImage*, 80:53–61, 2013.

[243] O. Sporns. Contributions and challenges for network models in cognitive neuroscience. *Nature neuroscience*, 17(5):652–660, 2014.

[244] O. Sporns, C. J. Honey, and R. Kötter. Identification and classification of hubs in brain networks. *PloS one*, 2(10):e1049, 2007.

[245] O. Sporns and R. Kötter. Motifs in brain networks. *PLoS Biol*, 2(11):e369, 2004.

[246] O. Sporns, G. Tononi, and R. Kötter. The human connectome: A structural description of the human brain. *PLoS computational biology*, 1(14):e42, 2005.

[247] O. Sporns and J. D. Zwi. The small world of the cerebral cortex. *Neuroinformatics*, 2(2):145–162, 2004.

[248] K. Sreejith, J. Jost, E. Saucan, and A. Samal. Forman curvature for directed networks. *arXiv:1605.04662*, 2017.

[249] R. P. Sreejith, K. Mohanraj, J. Jost, E. Saucan, and A. Samal. Forman curvature for complex networks. *arXiv:1603.00386*, 2016.

[250] E. K. St. Louis, L. C. Frey, J. W. Britton, L. C. Frey, J. L. Hopp, P. Korb, M. Z. Koubeissi, W. E. Lievens, E. M. Pestana-Knight, and E. K. St. Louis. *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants*. American Epilepsy Society, Chicago, 2016.

[251] C. J. Stam. Modern network science of neurological disorders. *Nature reviews. Neuroscience*, 15(10):683–695, 2014.

[252] C. J. Stam and J. C. Reijneveld. Graph theoretical analysis of complex networks in the brain. *Nonlinear biomedical physics*, 1(1):3, 2007.

[253] M. L. Stanley, M. N. Moussa, B. M. Paolini, R. G. Lyday, J. H. Burdette, and P. J. Laurienti. Defining nodes in complex brain networks. *Front. Comput. Neurosci.*, 7:169, 2013.

[254] J. J. Steenbergen. *Towards a Spectral Theory for Simplicial Complexes*. PhD thesis, Duke University, 2013.

[255] A. Steger and N.C. Wormald. Generating random regular graphs quickly. *Combinatorics, Probability and Computing*, 8:377–396, 1999.

[256] Q. Sun, Y. Liu, and S. Li. Weighted directed graph-based automatic seizure detection with effective brain connectivity for eeg signals. *SIViP*, 18:899–909, 2024.

[257] Y. Sun, H. Zhao, J. Liang, and X. Ma. Eigenvalue-based entropy in directed complex networks. *PLoS ONE*, 16(6):e0251993, 2021.

[258] B. Tadić, M. Andjelković, B. M. Boshkoska, and Z. Levnajić. Algebraic topology of multibrain connectivity networks reveals dissimilarity in functional patterns during spoken communications. *PLoS ONE*, 11(11):e0166787, 2016.

[259] D. Y. Takahashi, L. A. Baccala, and K. Sameshima. Frequency domain connectivity: an information theoretic perspective. *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, page 1726–1729, 2010.

[260] D. Y. Takahashi, L. A. Baccalá, and K. Sameshima. Connectivity inference between neural structures via partial directed coherence. *J. Appl. Stat.*, 34:1259–1273, 2007.

[261] D. Y. Takahashi, L. A. Baccalá, and K. Sameshima. Information theoretic interpretation of frequency domain connectivity measures. *Biological cybernetics*, 103(6):463–469, 2010.

[262] T. P. Trappenberg. *Fundamentals of Computational Neuroscience*. Oxford University Press, 3rd edition, 2023.

[263] R. L. Utianski, J. N. Caviness, E. C. van Straaten, T. G. Beach, B. N. Dugger, H. A. Shill, E. D. Driver-Dunckley, M. N. Sabbagh, S. Mehta, C. H. Adler, and J. G. Hentz. Graph theory network function in parkinson's disease assessed with electroencephalography. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, 127(5):2228–2236, 2016.

[264] S. P. van den Broek, F. Reinders, M. Donderwinkel, and M. J. Peters. Volume conduction effects in eeg and meg. *Electroencephalography and Clinical Neurophysiolog*, 106(6):522–534, 1998.

[265] M. P. van den Heuvel, C. J. Stam, R. S. Kahn, and H. E. Hulshoff Pol. Efficiency of functional brain networks and intellectual performance. *J. Neurosci.*, 29:7619–7624, 2009.

[266] B. C. M. van Wijk, C. J. Stam, and A. Daffertshofer. Comparing brain networks of different size and connectivity density using graph theory. *PLoS ONE*, 5(10):e13701, 2010.

[267] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

[268] M. C. Vlooswijk, M. J. Vaessen, J. F. A. Jansen, M. C. F. T. M. de Krom, H. J. M. Majoie, P. A. M. Hofman, and et al. Loss of network efficiency associated with cognitive decline in chronic epilepsy. *Neurology*, 77:938–944, 2011.

[269] P. von Bunau, F. C. Meinecke, S. Scholler, and K. R. Muller. Finding stationary brain sources in eeg data. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, page 2810–2813, 2010.

[270] E. Výtvarová, J. Fousek, M. Bartoň, R. Mareček, M. Gajdoš, M. Lamoš, and et al. The impact of diverse preprocessing pipelines on brain functional connectivity. *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 2644–2648, 2017.

[271] B. Wang, H. Tang, and Z. Z. Xiu. Entropy optimization of scale-free networks robustness to random failures. *Physica A: Statistical Mechanics and its Applications*, 363(2):591–596, 2006.

[272] L. Wang, C. Zhu, Y. He, Y. Zang, Q. Cao, H. Zhang, and et al. Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. *Hum. Brain Mapp.*, 30:638–649, 2009.

[273] Z. Wang, F. Liu, S. Shi, S. Xia, F. Peng, L. Wang, S. Ai, and Z. Xu. Automatic epileptic seizure detection based on persistent homology. *Frontiers in physiology*, 14:1227952, 2023.

[274] M. Ward and E. L. Schofield. Biomarkers for brain disorders. *Therapy*, 7(4):321–336, 2010.

[275] V. S. Wasade and M. V. Spanaki (eds). *Understanding Epilepsy: A Study Guide for the Boards*. Cambridge University Press, 2019.

[276] D. J. Watts. *Small Worlds*. Princeton University Press, 1999.

[277] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[278] D. B. West. *Introduction to Graph Theory*. Pearson Education, 2000.

[279] U. Wilensky. Netlogo. *Center for Connected Learning and Computer-Based Modeling, Northwestern University*, 1999. http://ccl.northwestern.edu/netlogo.

[280] P. Wills and F. G. Meyer. Metrics for graph comparison: A practitioner's guide. *PLOS ONE*, 15(2):e0228728, 2020.

[281] C. Wu, S. Ren, J. Wu, and K. Xia. Weighted (co)homology and weighted laplacian. *arXiv:1804.06990*, 2021.

[282] M. Xia, J. Wang, and Y. He. Brainnet viewer: a network visualization tool for human brain connectomics. *PLoSONE*, 8:e68910, 2013.

[283] J. Xu. Simulating weighted, directed small-world networks. *Appl. Math. Comput.*, 216:2118–2128, 2010.

[284] T. Yamada. The ricci curvature on simplicial complexes. *Theory and Applications of Graphs*, 10(2), 2023.

[285] C. Ye, R. C. Wilson, C. H. Comin, L. D. Costa, and E. R. Hancock. Approximate von neumann entropy for directed graphs. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 89(5):052804, 2014.

[286] Z. Yin, J. Li, Y. Zhang, A. Ren, K. M. Von Meneen, and L. Huang. Functional brain network analysis of schizophrenic patients with positive and negative syndrome based on mutual information of eeg time series. *Biomedical Signal Processing and Control*, 31:331–338, 2017.

[287] J. Yoo, E. Y. Kim, Y. M. Ahn, and J. C. Ye. Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. *Journal of Neuroscience Methods*, 267:1–13, 2016.

[288] L. Zager. Graph similarity and matching. Master's thesis, MIT, Cambridge, MA, USA, 2005.

[289] A. Zalesky, A. Fornito, and E. T. Bullmore. On the use of correlation as a measure of network connectivity. *Neuroimage*, 60:2096–2106, 2012.

[290] A. Zalesky, A. Fornito, I. H. Harding, L. Cocchi, M. Yücel, C. Pantelis, and et al. Whole-brain anatomical networks: does the choice of nodes matter? *Neuroimage*, 50:970–983, 2010.

[291] Z. Zhang. Random walk on simplicial complexes. *Algebraic Topology. [math.AT]. Université Paris-Saclay*, 2020.

[292] S. Zhou and R. J. Mondragón. The rich-club phenomenon in the internet topology. *IEEE Communications Letters*, 8:180–182, 2003.

[293] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005.

Appendix A

Software Review

In this appendix, we present an overview of the software used in the examples of Chapter 4 and in the analysis of Chapter 7 and also a brief description of the `DigplexQ` package, a novel Python package that was developed during this thesis.

A.1 Software Review

There are several free/open-source software available for brain connectivity estimation, graph theoretical analysis, topological data analysis, and statistical analysis, written in several different programming languages. In what follows, we present a brief description of the software used in this thesis.

- **asympPDC:** asympPDC is a MATLAB/Octave package for the analysis of time series data via partial directed coherence (PDC) and/or directed transfer function (DTF) (and their variants - information/generalized PDC/DTF) to infer directed interactions in the frequency domain between structures.
  Available at https://github.com/asymppdc

- **DigplexQ:** DigplexQ is a Python package to perform computations with digraph-based complexes (e.g. directed flag complexes and path complexes).
  Available at https://github.com/heitorbaldo/DigplexQ

- **Giotto-TDA:** Giotto-TDA is a Python package built on top of scikit-learn for topological data analysis. It provides functions to compute Betti numbers, persistence diagrams, barcodes, Betti curves, bottleneck distance, etc., from simplicial complexes or graphs/digraphs given as input data.
  Available at https://giotto-ai.github.io/gtda-docs/0.5.1/library.html

- **Flagser:** Flagser is a C++ library built on top of Ripser to compute homologies of directed flag complexes. It is also implemented within the Giotto-TDA package.
  Available at https://github.com/luetge/flagser

- **HodgeLaplacians:** HodgeLaplacians is a Python package created to compute the Hodge Laplacian matrices from simplicial complexes given as input data. Available at https://github.com/tsitsvero/hodgelaplacians

- **NetworkX:** NetworkX is a Python package to perform quantitative analysis in undirected and directed graphs. It provides several graph measures, graph distance measures, random models, visualization functions, etc. Available at https://networkx.org.

- **Persim:** Persim is a Python package for topological data analysis. It provides functions to compute persistence diagrams, persistence landscapes, bottleneck distance, heat kernel, etc., from simplicial complexes given as input data. Available at https://persim.scikit-tda.org/en/latest/

- **Pingouin:** Pingouin is a Python package based on NumPy and Pandas to perform statistical analysis. It provides standard parametric and non-parametric statistical tests, such as t-test, ANOVAs, Kruskal-Wallis test, Mann-Whitney test, Wilcoxon signed-rank test, etc. Available at https://pingouin-stats.org/build/html/index.html

A.2 DigplexQ

As part of this thesis, the Python package `DigplexQ` (released as free software under the MIT license[1]) was developed to perform computations with digraph-based complexes. It is based on other well-known Python packages, such as `NetworkX` (graph measures), `Giotto-TDA/Flagser` (persintence diagrams, Betti numbers, and topological distances), `Persim` (topological distances), and `HodgeLaplacians` (Hodge Laplacians).

`DigplexQ` is an "adjacency matrix-centered" package since it was designed so that the user can perform all computations just by entering an adjacency matrix as input. It contains the implementation of almost all simplicial characterizations measures and simplicial similarity comparision distances introduced in Chapter 4.

The `DigplexQ` package has been tested under Windows and Linux-Ubuntu platforms running Python version 3.7 and higher. It is available in the PyPi repository at https://pypi.org/project/digplexq (current version 0.0.7), and it can be installed through the PIP package manager via the command `pip3 install digplexq`.

[1] https://opensource.org/license/mit

Appendix B

Supplementary Information

B.1 Summary of the Simplicial Measures

Table B.1: Summary of the directed simplicial measures used in the analysis and/or in the examples. For simplicity's sake, we omitted the terms "directed simplicial" from the nomenclature of the measures.

id	Measure	Notation	Equation
1	Average Shortest q-Walk Length	$\vec{L}_q(\mathcal{G}_q)$	[4.1]
2	Global q-Efficiency	$\vec{E}^q_{glob}(\mathcal{G}_q)$	[4.3]
3	q-Returnability	$K_{r,q}(\mathcal{G}_q)$	[4.5]
4	In-q-Degree Centrality	$C^-_{deg_q}(\sigma)$	[4.7]
5	Out-q-Degree Centrality	$C^+_{deg_q}(\sigma)$	[4.8]
6	q-Harmonic Centrality	$\vec{HC}_q(\sigma)$	[4.12]
7	q-Betweenness Centrality	$\vec{B}_q(\sigma)$	[4.13]
8	Global q-Reaching Centrality	$GRC_q(\mathcal{G}_q)$	[4.16]
9	Average q-Clustering Coefficient	$\vec{C}_q(\mathcal{G}_q)$	[4.17]
10	in-q-Rich-Club Coefficient	$\phi^-_{q,k}(\mathcal{G}_q)$	[4.22]
11	out-q-Rich-Club Coefficient	$\phi^+_{q,k}(\mathcal{G}_q)$	[4.23]
12	q-Structural Entropy	$H^{str}_q(\mathcal{G}_q)$	[4.26]
13	in-q-Degree Distribution Entropy	$H^-_q(\mathcal{G}_q)$	[4.29]
14	out-q-Degree Distribution Entropy	$H^+_q(\mathcal{G}_q)$	[4.30]
15	in-q-Forman-Ricci Curvature	$F^-_q(\sigma)$	[4.33]
16	out-q-Forman-Ricci Curvature	$F^+_q(\sigma)$	[4.34]
17	q-Energy	$\varepsilon_q(\mathcal{G}_q)$	[4.36]
18	in-q-Katz Centrality	$\vec{K}^-_q(\sigma)$	[4.37]

B.2 Summary of the Simplicial Distances

Table B.2: Summary of the simplicial and persistence distances used in the analysis and/or in the examples.

id	Distance	Notation	Equation
1	Bottleneck Distance	$d_{W_\infty}(P,Q)$	[3.23]
2	p-Wasserstein Distance	$d_{W_p}(P,Q)$	[3.24]
3	Betti Distance	$d_{\mathcal{B}}(\mathcal{B}_P,\mathcal{B}_Q)$	[3.25]
4	First Topological Structure Distance	$\widehat{T}^1_{tsd}(\mathcal{X}_1,\mathcal{X}_2)$	[4.43]
5	Fourth Topological Structure Distance	$\widehat{T}^4_{tsd}(\mathcal{X}_1,\mathcal{X}_2)$	[4.43]
6	Fifth Topological Structure Distance	$\widehat{T}^5_{tsd}(\mathcal{X}_1,\mathcal{X}_2)$	[4.43]
7	Histogram Cosine Kernel	$K_{HC}(\mathcal{X}_1,\mathcal{X}_2)$	[4.45]
8	Jaccard Kernel	$K_J(\mathcal{X}_1,\mathcal{X}_2)$	[4.46]
9	Simplicial n-Spectral Distance	$\widehat{D}_{\mathcal{L}_n}(\mathcal{X}_1,\mathcal{X}_2)$	[4.50]

Appendix C

Results

Table C.1: W-statistics and p-values (in parentheses) for simplicial measures showing significant differences ($p < 0.05$, Wilcoxon paired test) between the pre-ictal phase and ictal phase for the original iPDC networks of the left and right hemispheres in the delta and theta bands. See Table B.1 for the measure ids.

	Left		**Right**	
Measure	**Delta**	**Theta**	**Delta**	**Theta**
2	3.0 (0.039)	-	2.0 (0.023)	0.0 (0.008)
4	-	-	-	-
5	-	-	-	-
6	-	-	1.0 (0.016)	3.0 (0.039)
8	-	-	1.0 (0.016)	3.0 (0.039)
9	2.0 (0.023)	-	-	0.0 (0.008)
13	-	-	2.0 (0.023)	0.0 (0.008)
14	-	-	3.0 (0.039)	1.0 (0.016)
17	3.0 (0.039)	-	2.0 (0.023)	0.0 (0.008)

Table C.2: W-statistics and p-values (in parentheses) for simplicial measures showing significant differences ($p < 0.05$, Wilcoxon paired test) between the pre-ictal phase and ictal phase for the 0-digraphs of the left and right hemispheres in the delta and theta bands. See Table B.1 for the measure ids.

	Left		**Right**	
Measure	**Delta**	**Theta**	**Delta**	**Theta**
2	2.0 (0.023)	-	3.0 (0.039)	-
4	2.0 (0.023)	-	-	-
5	2.0 (0.023)	-	-	-
6	2.0 (0.023)	-	2.0 (0.023)	-
8	2.0 (0.023)	-	1.0 (0.016)	1.0 (0.016)
9	2.0 (0.023)	-	-	-
13	-	-	3.0 (0.039)	-
14	-	-	3.0 (0.039)	-
17	2.0 (0.023)	-	2.0 (0.023)	-

Table C.3: W-statistics and p-values (in parentheses) for simplicial measures showing significant differences ($p < 0.05$, Wilcoxon paired test) between the pre-ictal phase and ictal phase for the 1-digraphs of the left and right hemispheres in the delta and theta bands. See Table B.1 for the measure ids.

	Left		**Right**	
Measure	**Delta**	**Theta**	**Delta**	**Theta**
2	2.0 (0.023)	-	3.0 (0.039)	-
4	3.0 (0.039)	-	3.0 (0.039)	3.0 (0.039)
5	2.0 (0.023)	-	0.0 (0.022)	3.0 (0.039)
6	2.0 (0.023)	-	-	-
8	2.0 (0.023)	-	-	0.0 (0.008)
9	1.0 (0.016)	2.0 (0.023)	0.0 (0.008)	1.0 (0.016)
13	2.0 (0.023)	-	-	3.0 (0.039)
14	2.0 (0.023)	-	-	2.0 (0.023)
17	2.0 (0.023)	-	-	-

Table C.4: W-statistics and p-values (in parentheses) for simplicial measures showing significant differences ($p < 0.05$, Wilcoxon paired test) between the pre-ictal phase and ictal phase for the 2-digraphs of the left and right hemispheres in the delta and theta bands. See Table B.1 for the measure ids.

	Left		**Right**	
Measure	**Delta**	**Theta**	**Delta**	**Theta**
2	2.0 (0.023)	2.0 (0.023)	0.0 (0.008)	-
4	1.0 (0.016)	2.0 (0.023)	0.0 (0.008)	1.0 (0.016)
5	1.0 (0.016)	2.0 (0.023)	0.0 (0.008)	0.0 (0.008)
6	1.0 (0.016)	2.0 (0.023)	0.0 (0.008)	3.0 (0.039)
8	1.0 (0.016)	3.0 (0.039)	0.0 (0.008)	1.0 (0.016)
9	1.0 (0.035)	1.0 (0.016)	3.0 (0.039)	1.0 (0.035)
13	1.0 (0.016)	3.0 (0.039)	0.0 (0.008)	3.0 (0.039)
14	1.0 (0.016)	3.0 (0.039)	0.0 (0.008)	2.0 (0.023)
17	1.0 (0.016)	2.0 (0.023)	0.0 (0.008)	-

Table C.5: W-statistics and p-values (in parentheses) for simplicial measures showing significant differences ($p < 0.05$, Wilcoxon paired test) between the pre-ictal phase and ictal phase for the 3-digraphs of the left and right hemispheres in the delta and theta bands. See Table B.1 for the measure ids.

	Left		**Right**	
Measure	**Delta**	**Theta**	**Delta**	**Theta**
2	0.0 (0.022)	3.0 (0.039)	-	-
4	0.0 (0.022)	3.0 (0.039)	-	1.0 (0.035)
5	0.0 (0.022)	3.0 (0.039)	-	1.0 (0.035)
6	0.0 (0.022)	3.0 (0.039)	-	-
8	0.0 (0.022)	-	2.0 (0.023)	-
9	-	1.0 (0.035)	-	1.0 (0.035)
13	0.0 (0.022)	-	2.0 (0.023)	1.0 (0.035)
14	0.0 (0.022)	-	2.0 (0.023)	1.0 (0.035)
17	0.0 (0.022)	3.0 (0.039)	3.0 (0.039)	1.0 (0.035)

Table C.6: W-statistics and p-values (in parentheses) for simplicial distances showing significant differences ($p < 0.05$, Wilcoxon paired test) between the means of the distributions $d(G^L_{ic}, G^L_{pre})$ and $d(G^L_{pre}, G^L_{pre})$, for each frequency band. See Table B.2 for the distance ids.

Distance	**Delta**	**Theta**	**Alpha**
1	-	-	-
2	-	0.0 (0.008)	-
3	-	2.0 (0.023)	-
4	-	3.0 (0.039)	-
5	-	-	-
6	-	-	-
7	-	-	-
8	0.0 (0.008)	0.0 (0.008)	0.0 (0.008)

Table C.7: W-statistics and p-values (in parentheses) for simplicial distances showing significant differences ($p < 0.05$, Wilcoxon paired test) between the means of the distributions $d(G^L_{ic}, G^L_{pos})$ and $d(G^L_{pos}, G^L_{pos})$, for each frequency band. See Table B.2 for the distance ids.

Distance	**Delta**	**Theta**	**Alpha**
1	-	-	-
2	-	1.0 (0.016)	-
3	-	-	-
4	-	-	-
5	-	-	-
6	-	2.0 (0.023)	-
7	-	-	-
8	0.0 (0.008)	0.0 (0.008)	0.0 (0.008)

Table C.8: W-statistics and p-values (in parentheses) for simplicial distances showing significant differences ($p < 0.05$, Wilcoxon paired test) between the means of the distributions $d(G^R_{ic}, G^R_{pre})$ and $d(G^R_{pre}, G^R_{pre})$, for each frequency band. See Table B.2 for the distance ids.

Distance	**Delta**	**Theta**	**Alpha**
1	-	3.0 (0.039)	-
2	2.0 (0.023)	0.0 (0.008)	-
3		2.0 (0.023	-
4	-	-	-
5	-	-	-
6	-	-	-
7	-	-	-
8	0.0 (0.008)	0.0 (0.008)	0.0 (0.008)

Table C.9: W-statistics and p-values (in parentheses) for simplicial distances showing significant differences ($p < 0.05$, Wilcoxon paired test) between the distributions $d(G^R_{ic}, G^R_{pos})$ and $d(G^R_{pos}, G^R_{pos})$, for each frequency band. See Table B.2 for the distance ids.

Distance	**Delta**	**Theta**	**Alpha**
1	-	-	-
2	-	-	2.0 (0.023)
3	-	-	-
4	0.0 (0.008)	1.0 (0.016)	1.0 (0.016)
5	-	-	-
6	2.0 (0.023)	-	-
7	2.0 (0.023)	0.0 (0.008)	1.0 (0.016)
8	0.0 (0.008)	0.0 (0.008)	0.0 (0.008)

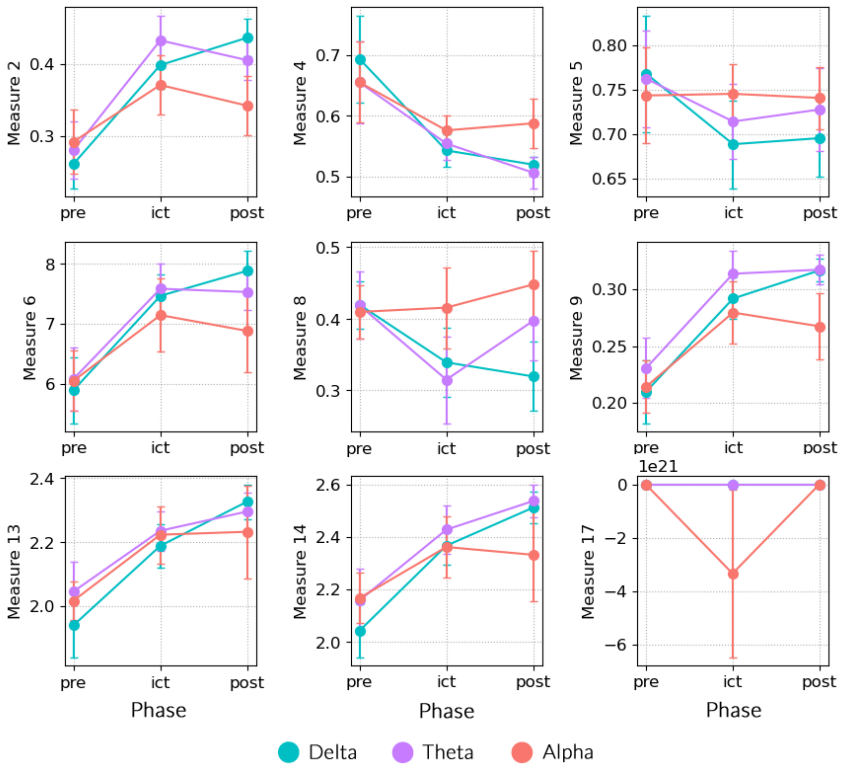


Figure C.1: Grand means and standard deviations of the simplicial measures computed in the original iPDC networks of the left hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

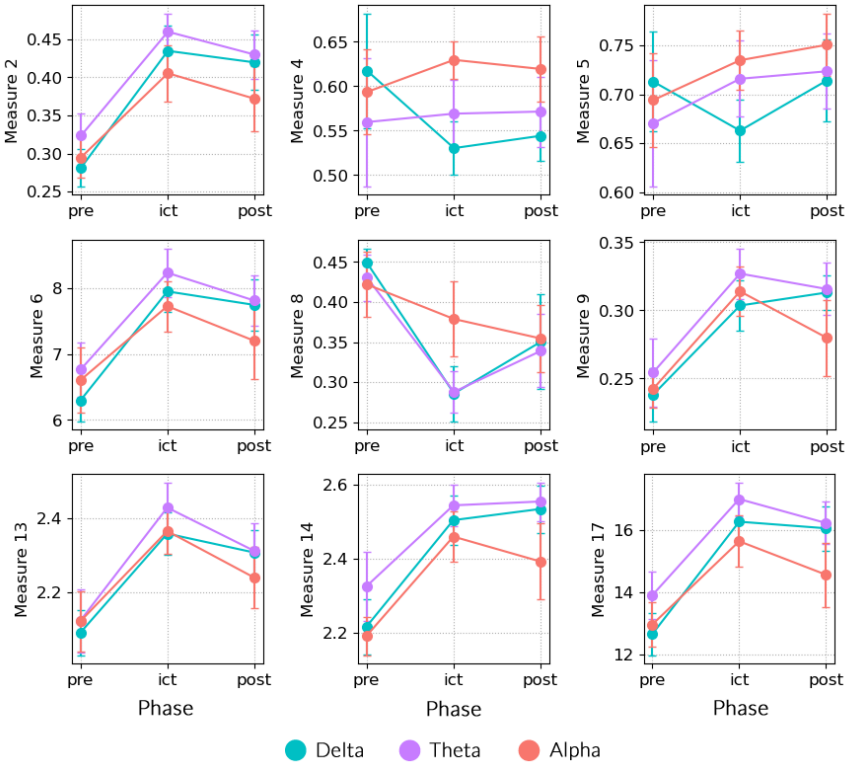


Figure C.2: Grand means and standard deviations of the simplicial measures computed in the original iPDC networks of the right hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

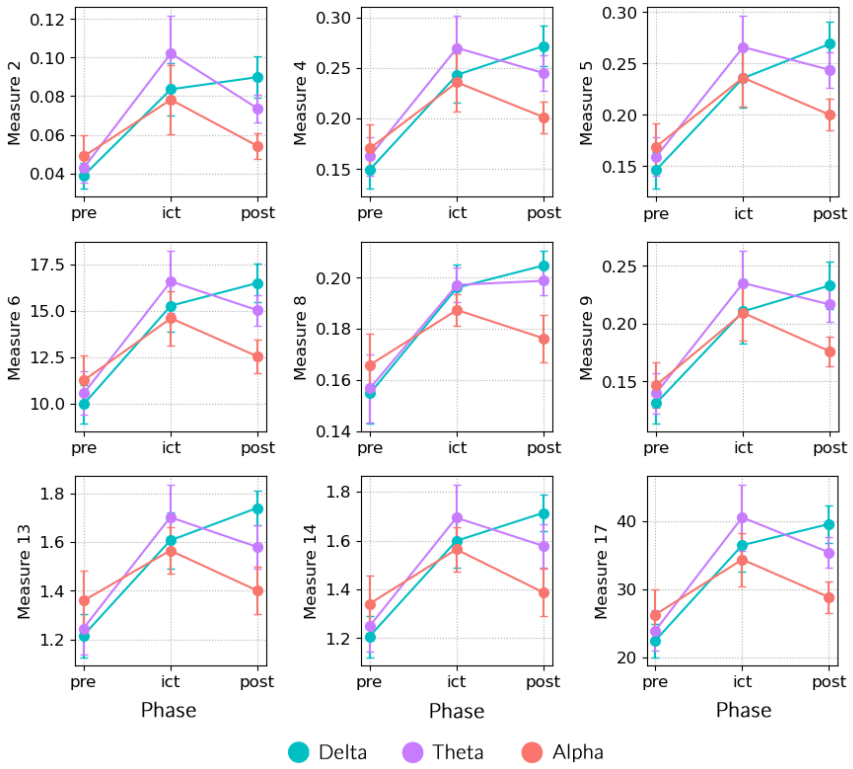


Figure C.3: Grand means and standard deviations of the simplicial measures computed in the 0-digraphs of the left hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

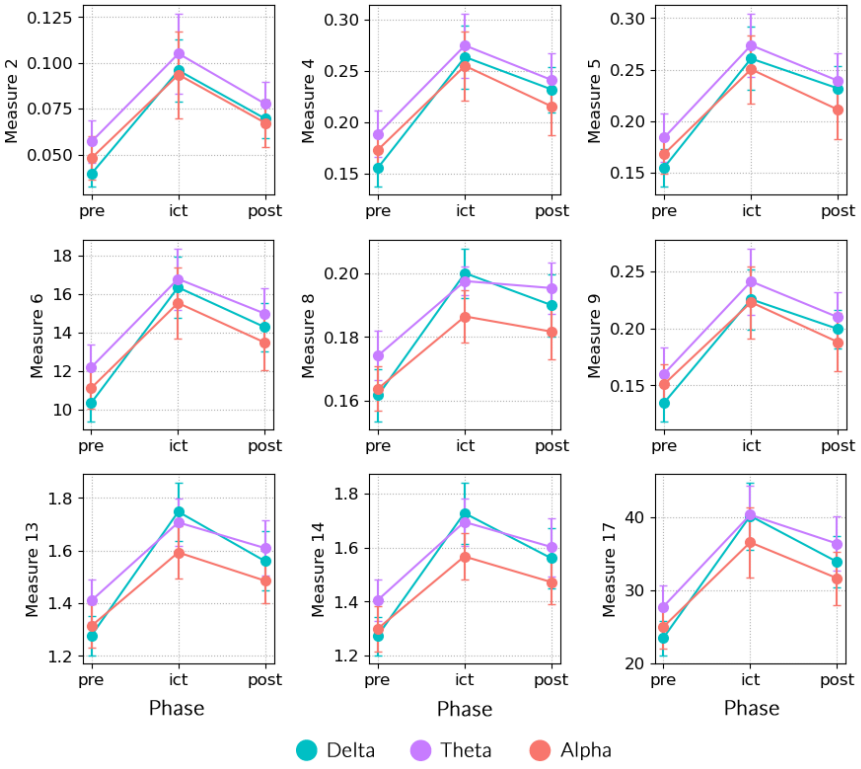


Figure C.4: Grand means and standard deviations of the simplicial measures computed in the 0-digraphs of the right hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

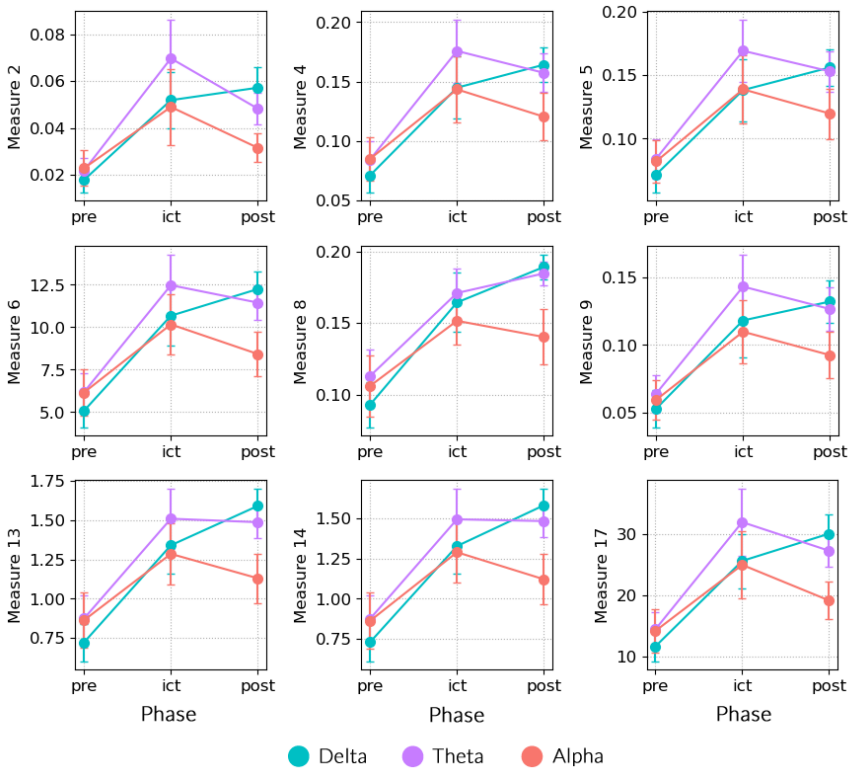


Figure C.5: Grand means and standard deviations of the simplicial measures computed in the 1-digraphs of the left hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

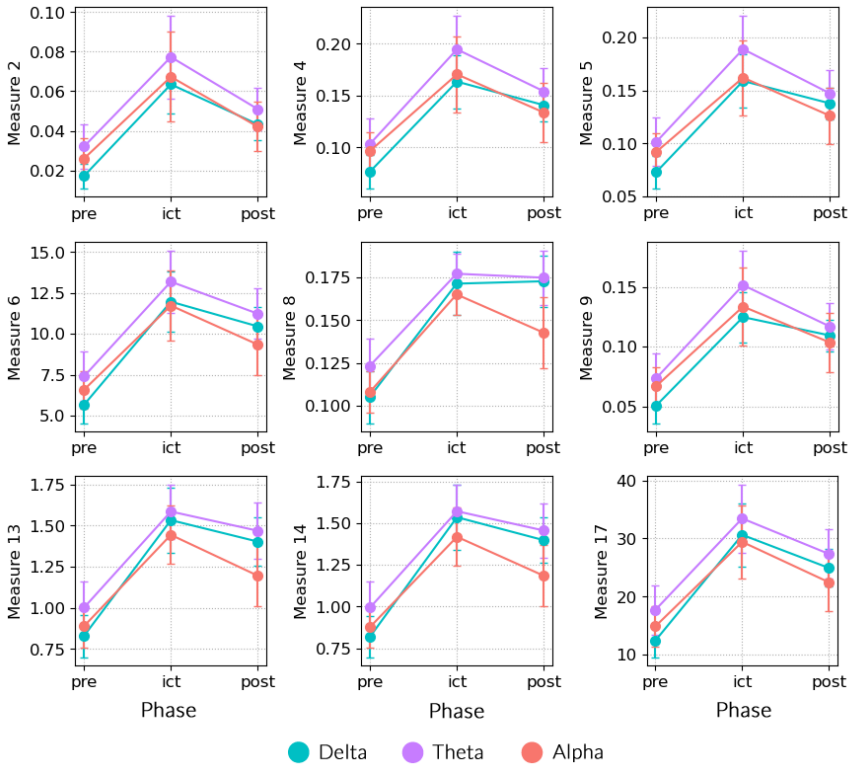


Figure C.6: Grand means and standard deviations of the simplicial measures computed in the 1-digraphs of the right hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

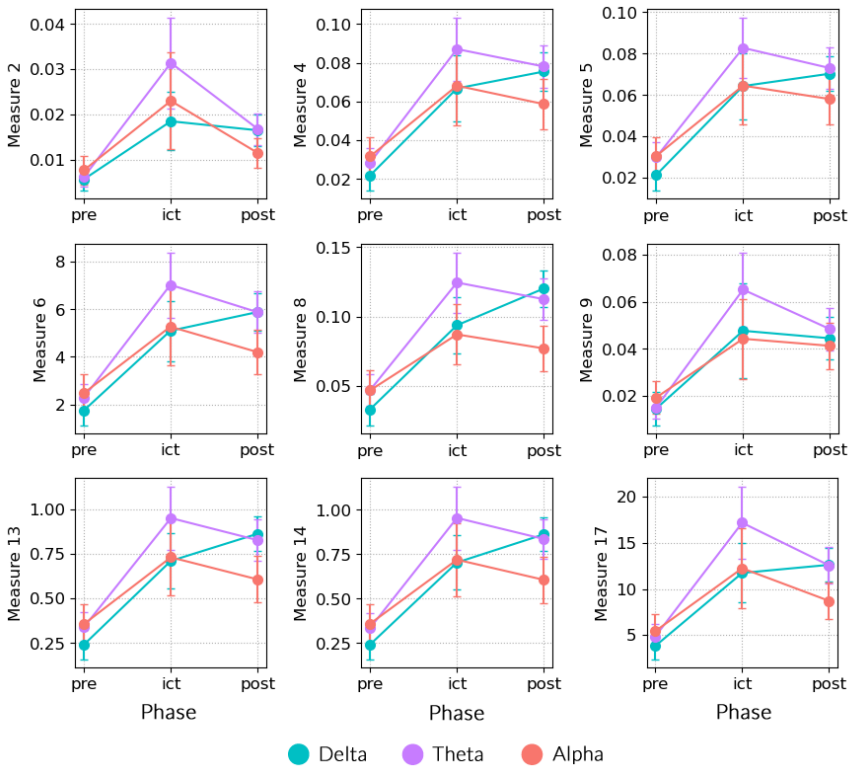


Figure C.7: Grand means and standard deviations of the simplicial measures computed in the 2-digraphs of the left hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

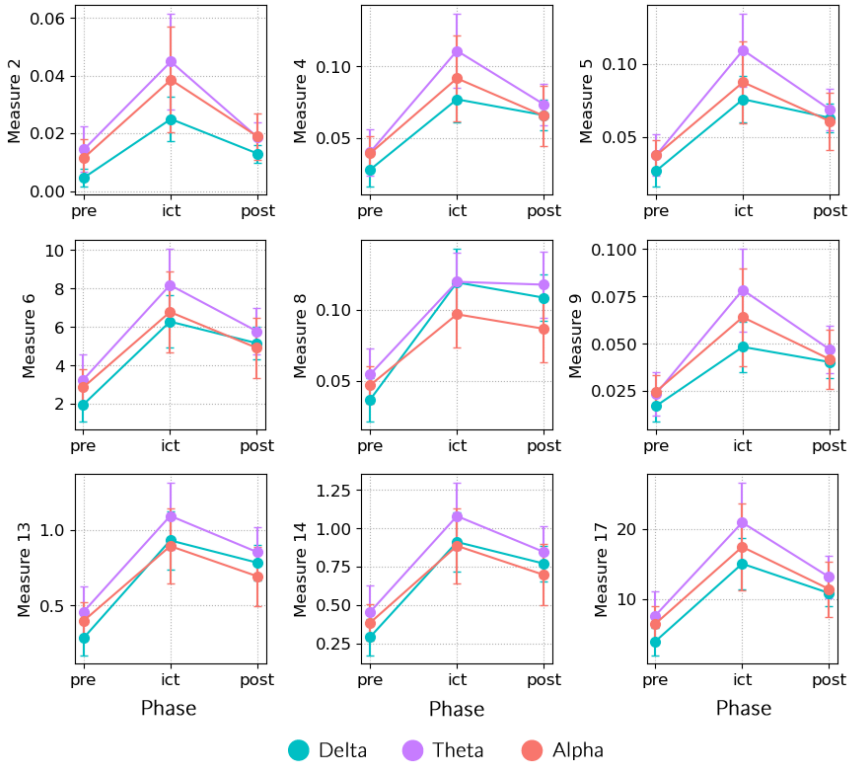


Figure C.8: Grand means and standard deviations of the simplicial measures computed in the 2-digraphs of the right hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

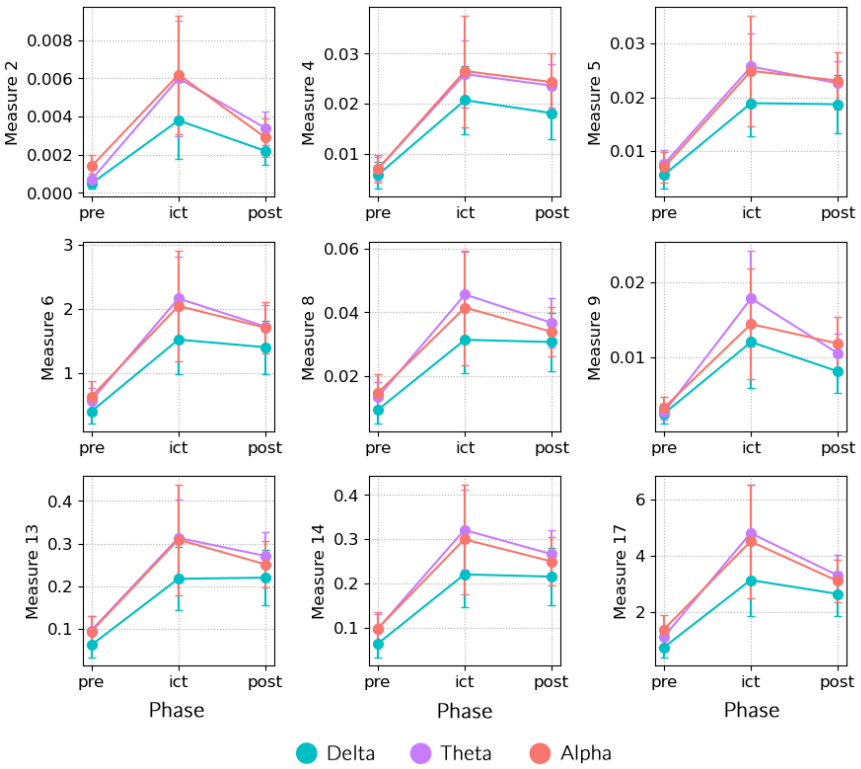


Figure C.9: Grand means and standard deviations of the simplicial measures computed in the 3-digraphs of the left hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

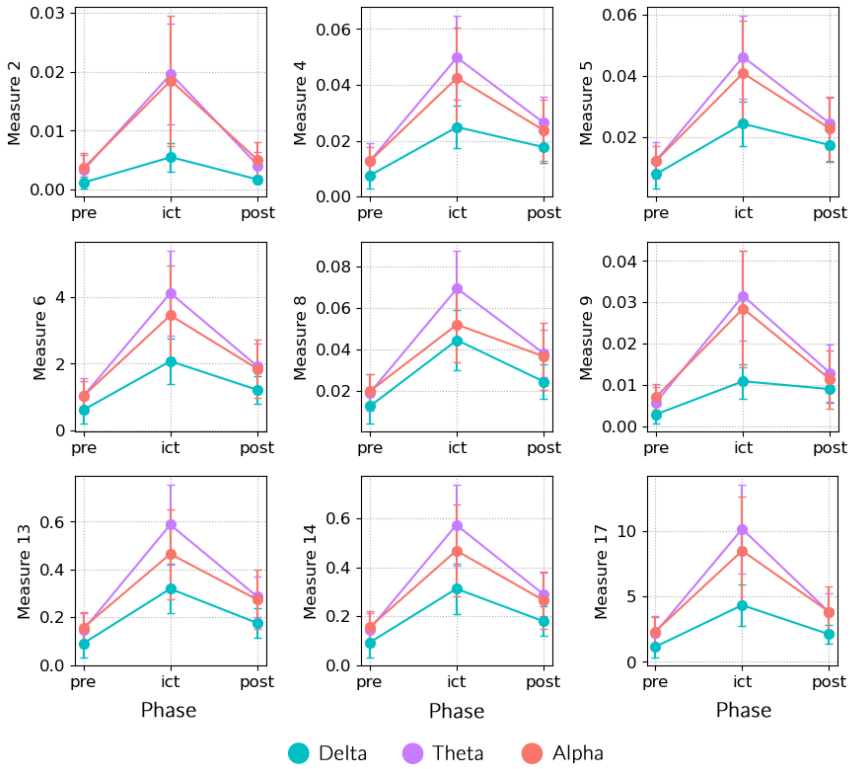


Figure C.10: Grand means and standard deviations of the simplicial measures computed in the 3-digraphs of the right hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

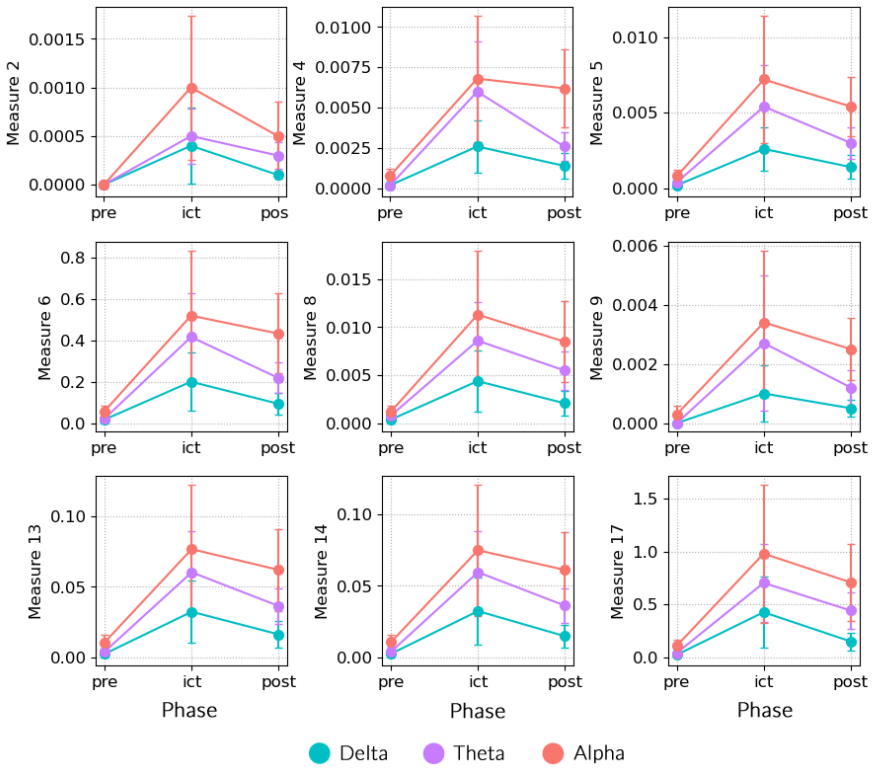


Figure C.11: Grand means and standard deviations of the simplicial measures computed in the 4-digraphs of the left hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

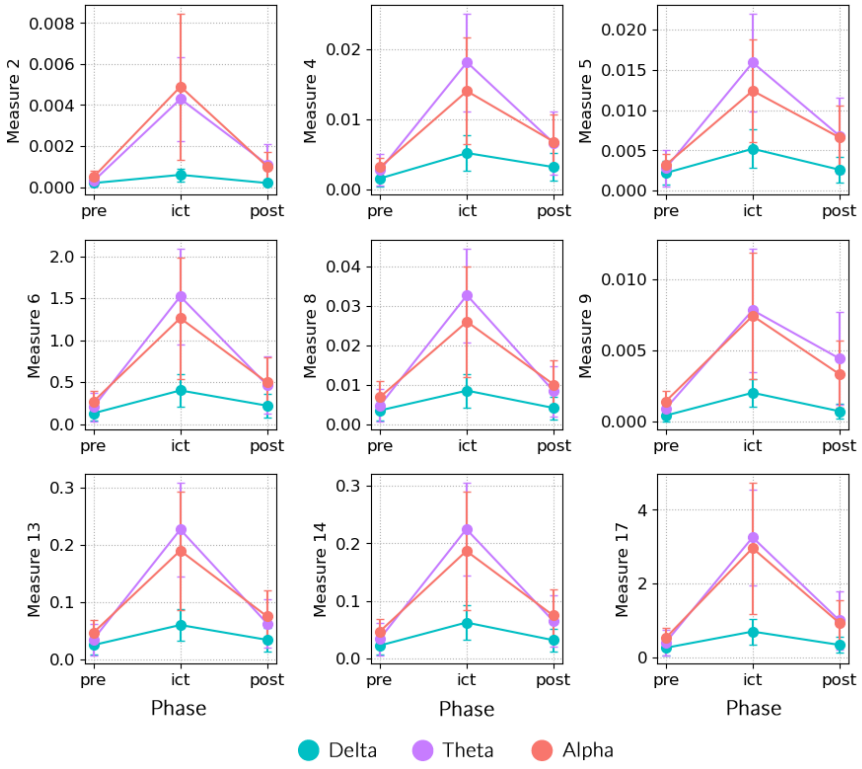


Figure C.12: Grand means and standard deviations of the simplicial measures computed in the 4-digraphs of the right hemisphere for each seizure phase (pre-ictal, ictal, post-ictal) and for each frequency band (delta, theta, alpha). See Table B.1 for the measure ids.

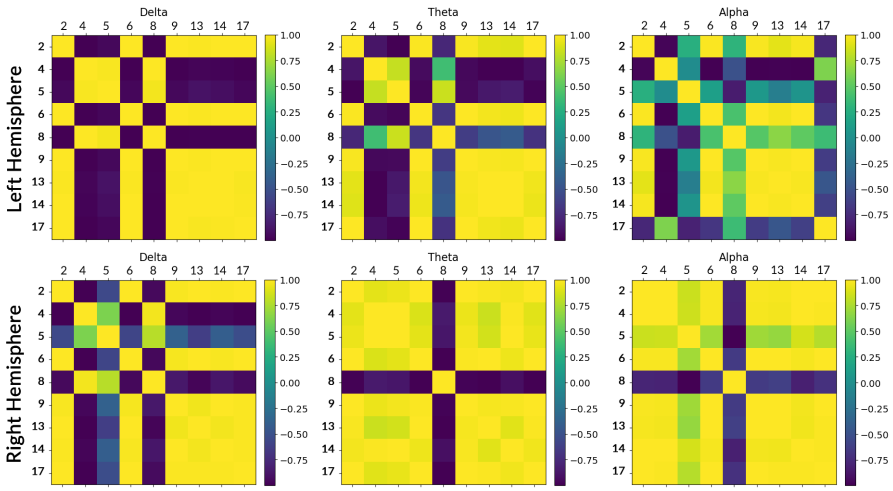


Figure C.13: Pearson's correlation coefficients between the simplicial measures across the pre-ictal, ictal, and post-ictal phases for the original iPDC networks (level $q = -1$). See Table B.1 for the measure ids.

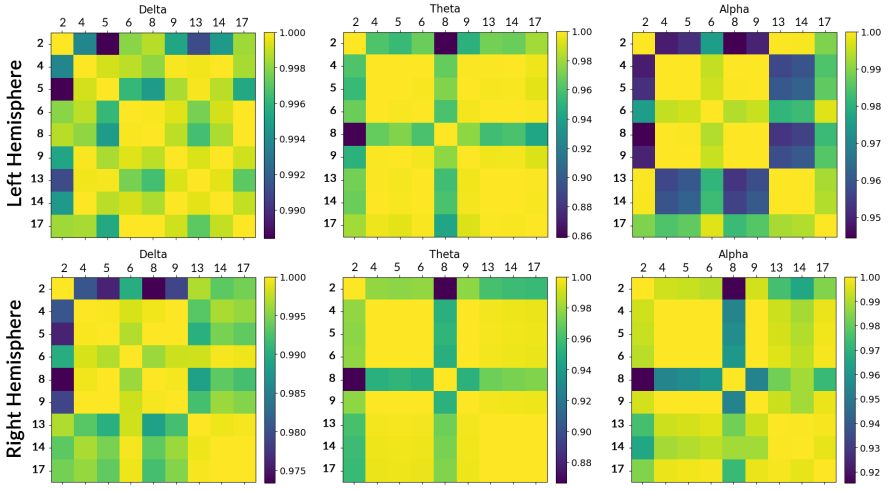


Figure C.14: Pearson's correlation coefficients between the simplicial measures across the pre-ictal, ictal, and post-ictal phases at level $q = 0$. See Table B.1 for the measure ids.

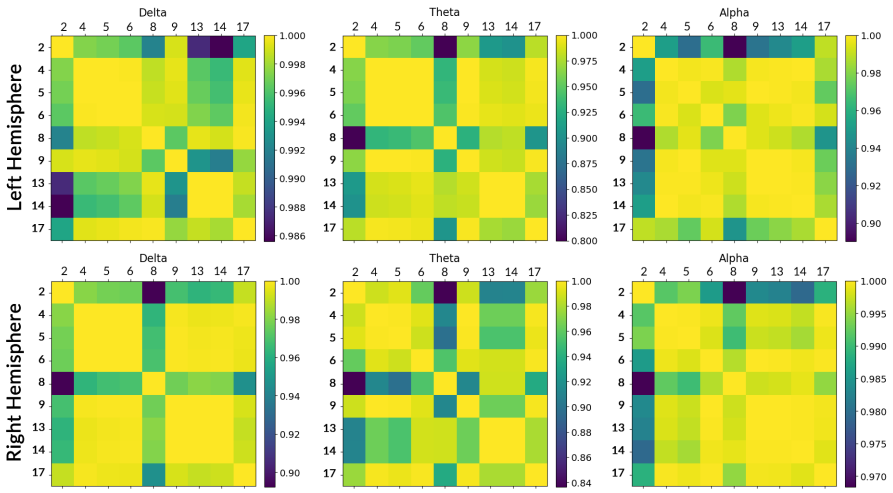


Figure C.15: Pearson's correlation coefficients between the simplicial measures across the pre-ictal, ictal, and post-ictal phases at level $q = 1$. See Table B.1 for the measure ids.

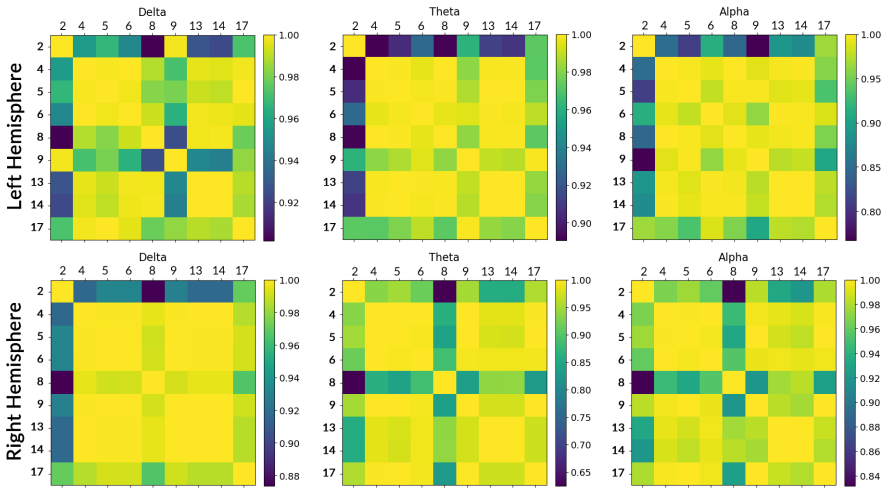


Figure C.16: Pearson's correlation coefficients between the simplicial measures across the pre-ictal, ictal, and post-ictal phases at level $q = 2$. See Table B.1 for the measure ids.

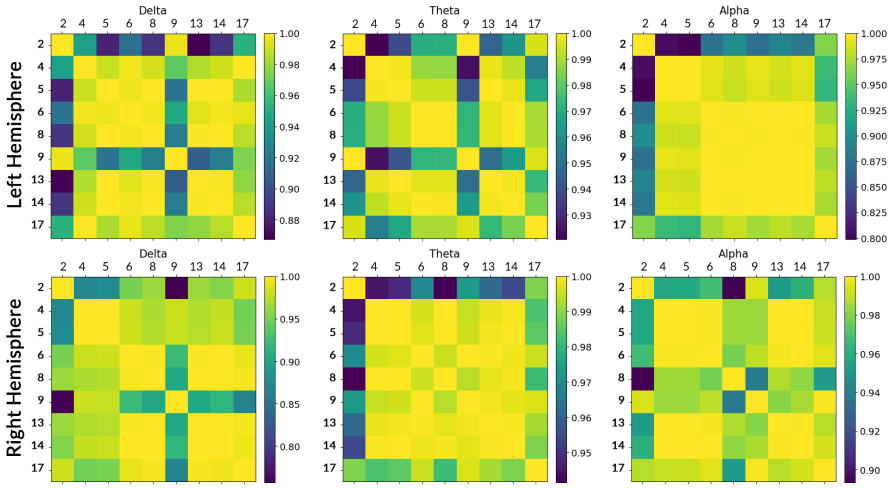


Figure C.17: Pearson's correlation coefficients between the simplicial measures across the pre-ictal, ictal, and post-ictal phases at level $q = 3$. See Table B.1 for the measure ids.

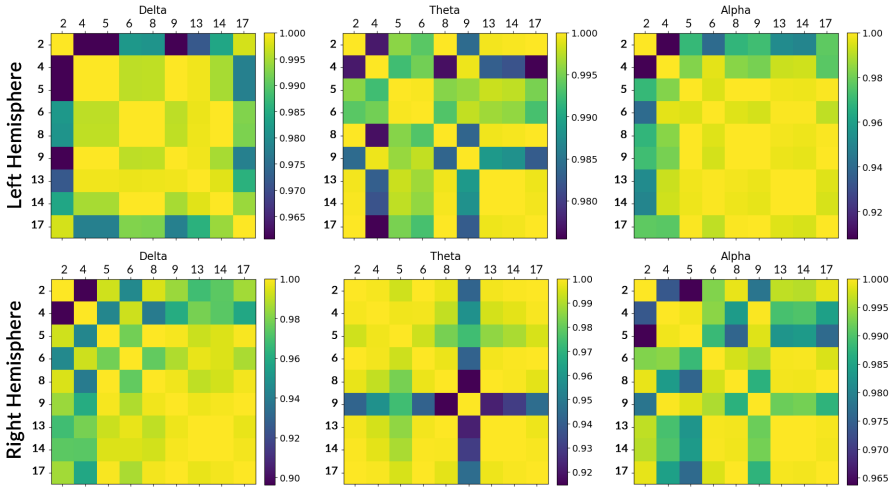


Figure C.18: Pearson's correlation coefficients between the simplicial measures across the pre-ictal, ictal, and post-ictal phases at level $q = 4$. See Table B.1 for the measure ids.

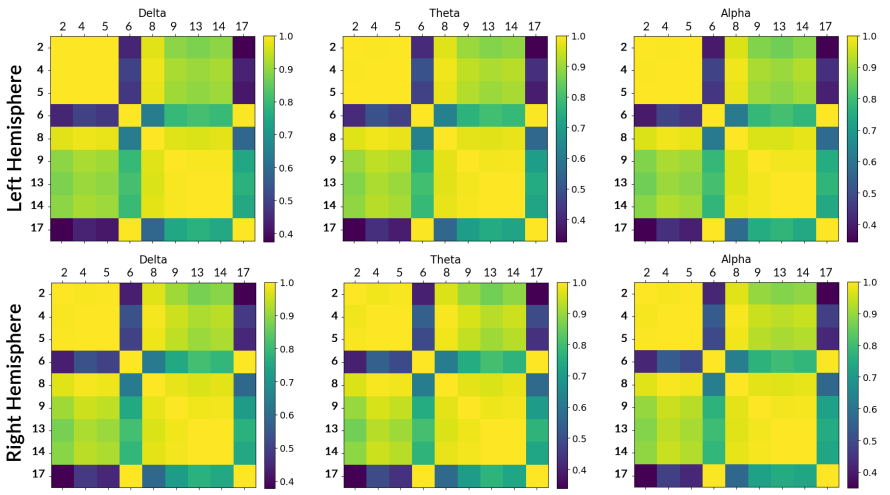


Figure C.19: Pearson's correlation coefficients between the simplicial measures across the levels $q = -1, 0, 1, 2, 3, 4$ in the pre-ictal phase. See Table B.1 for the measure ids.

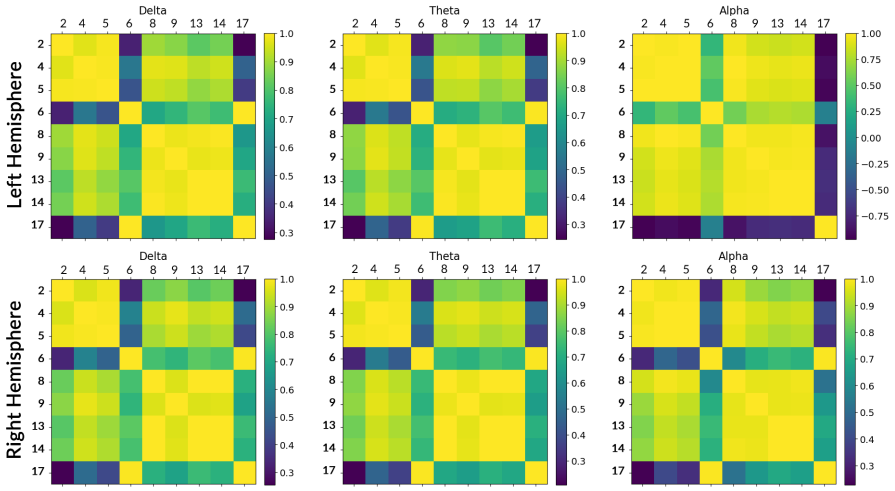


Figure C.20: Pearson's correlation coefficients between the simplicial measures across the levels $q = -1, 0, 1, 2, 3, 4$ in the ictal phase. See Table B.1 for the measure ids.

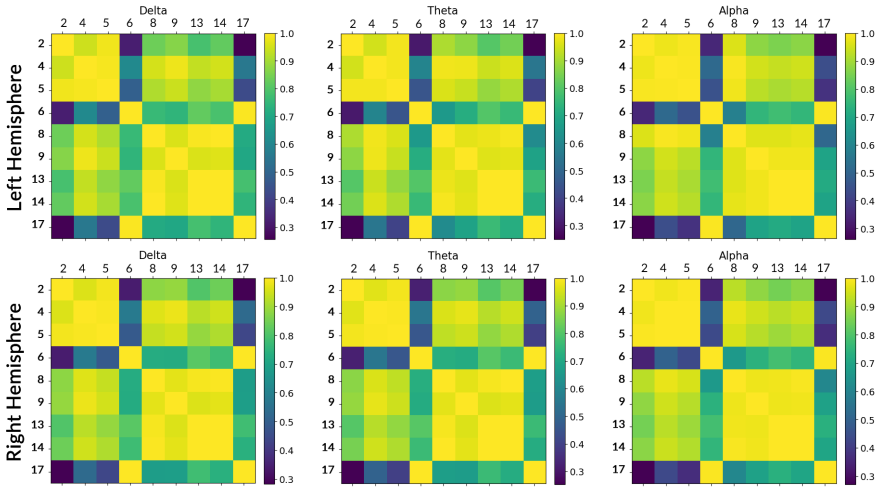


Figure C.21: Pearson's correlation coefficients between the simplicial measures across the levels $q = -1, 0, 1, 2, 3, 4$ in the post-ictal phase. See Table B.1 for the measure ids.

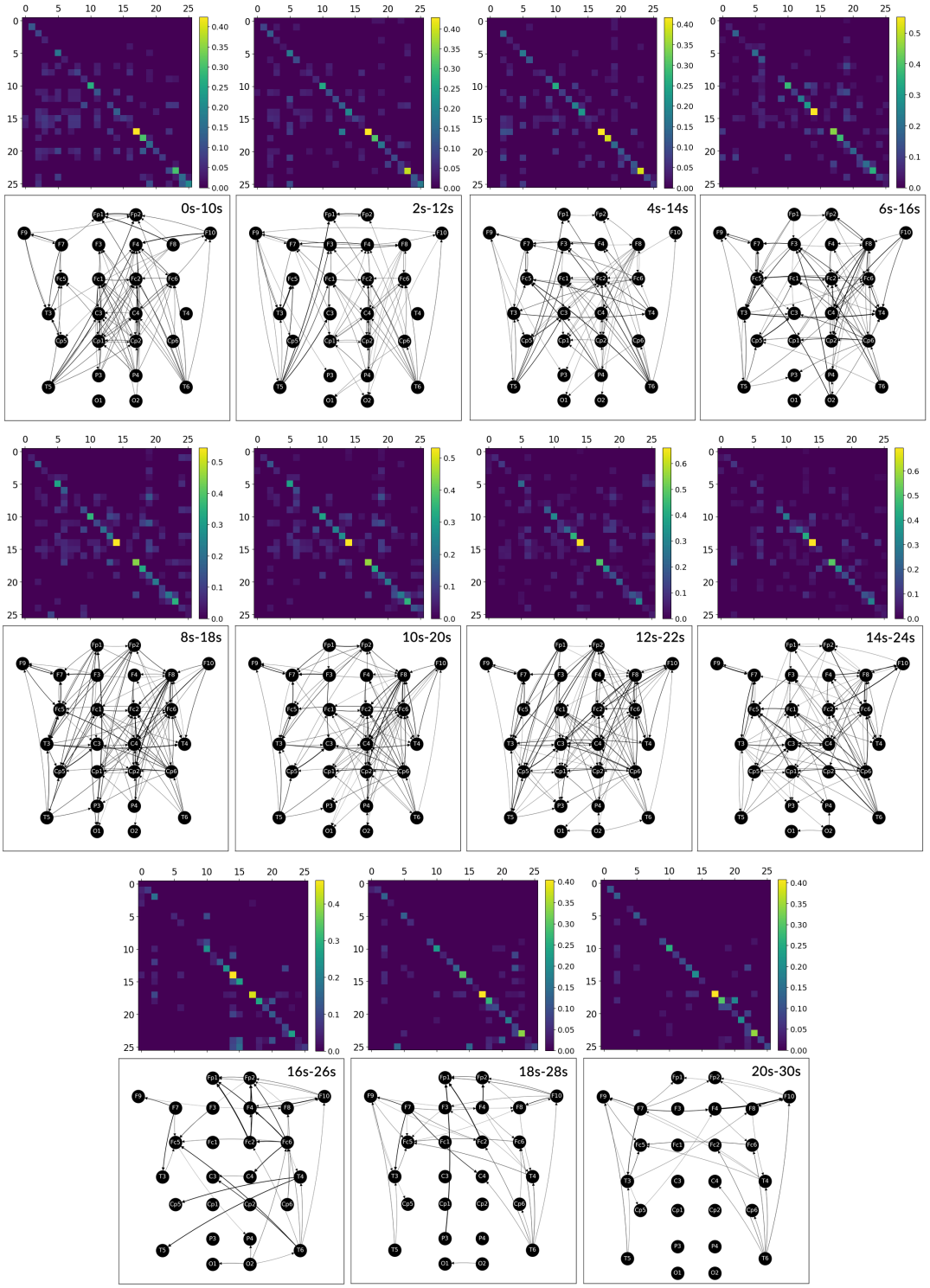


Figure C.22: iPDC networks, together with their weighted adjacency matrices (only statistically significant ($p < 0.001$) connections were considered), whose entries correspond to $|\iota\pi_{ij}(f)|^2$ (squared modulus), computed in the delta band of the pre-ictal phase of patient PN01. The networks were obtained using the sliding window technique with fixed-size windows of 10s and 80% overlap (i.e. 8s) in a 30s interval immediately before the seizure. The nodes in the adjacency matrices are numbered from 0 to 25, and correspond to the following electrodes, starting from bottom to top, right to left: 0 (O2), 1 (O1), 2 (T6), 3 (P4), ..., 25 (Fp1).

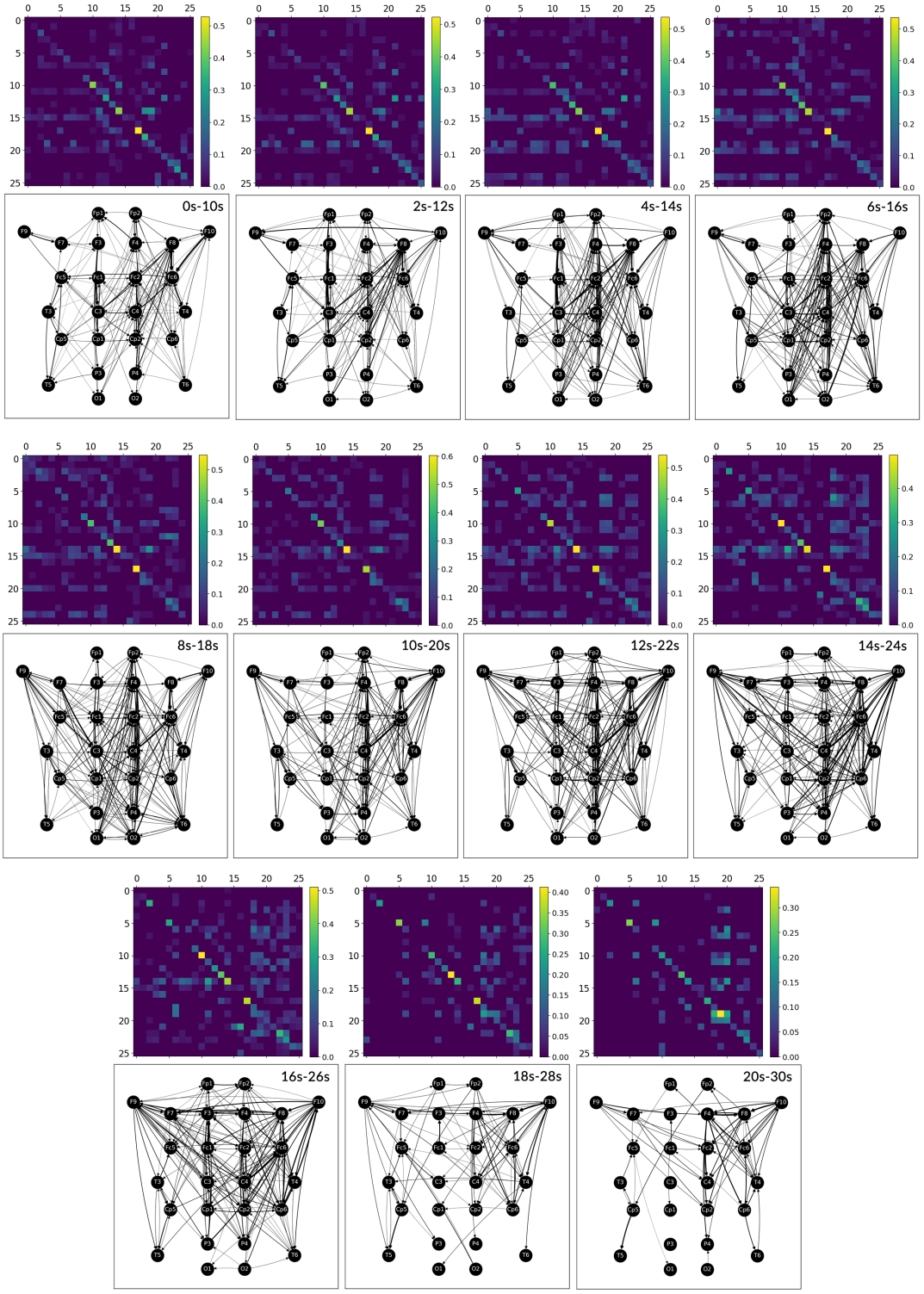


Figure C.23: iPDC networks, together with their weighted adjacency matrices (only statistically significant ($p < 0.001$) connections were considered), whose entries correspond to $|\iota\pi_{ij}(f)|^2$ (squared modulus), computed in the delta band of the ictal phase of patient PN01. The networks were obtained using the sliding window technique with fixed-size windows of 10s and 80% overlap (i.e. 8s) in a 30s interval starting on the seizure onset. The nodes in the adjacency matrices are numbered from 0 to 25, and correspond to the following electrodes, starting from bottom to top, right to left: 0 (O2), 1 (O1), 2 (T6), 3 (P4), ..., 25 (Fp1).

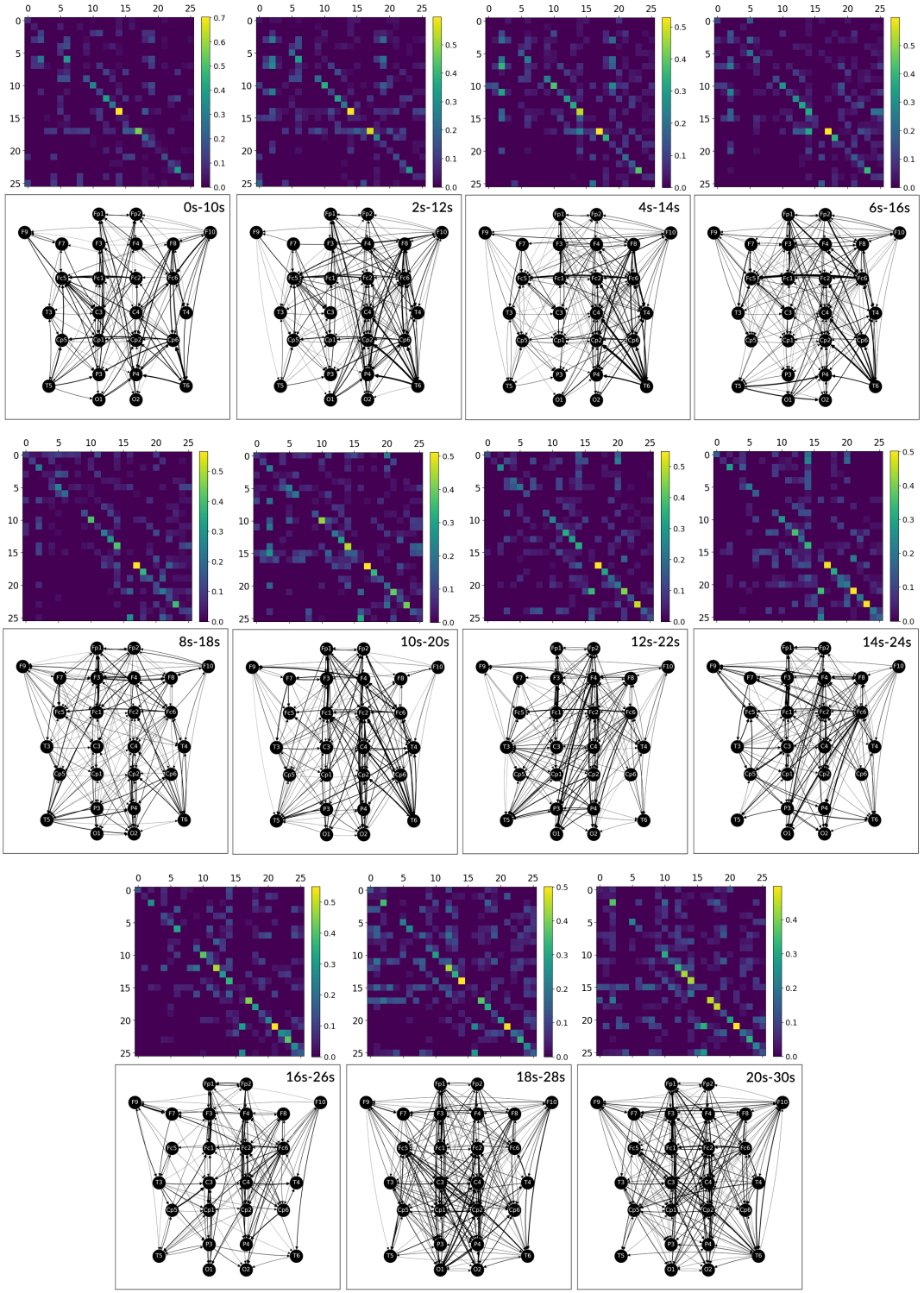


Figure C.24: iPDC networks, together with their weighted adjacency matrices (only statistically significant ($p < 0.001$) connections were considered), whose entries correspond to $|\iota\pi_{ij}(f)|^2$ (squared modulus), computed in the delta band of the post-ictal phase of patient PN01. The networks were obtained using the sliding window technique with fixed-size windows of 10s and 80% overlap (i.e. 8s) in a 30s interval immediately after the seizure. The nodes in the adjacency matrices are numbered from 0 to 25, and correspond to the following electrodes, starting from bottom to top, right to left: 0 (O2), 1 (O1), 2 (T6), 3 (P4), ..., 25 (Fp1).