UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

Edgar Corrêa de Amorim Filho

Topological transitions on Protein-Protein Interaction Networks

Recife
2019

Edgar Corrêa de Amorim Filho

Topological transitions on Protein-Protein Interaction Networks

Tese apresentada ao Programa de Pós-graduação em Matemática da Universidade Federal de Pernambuco como requisito parcial para obtenção do título de Doutor em matemática.

Área de Concentração: Geometria e Topologia

Orientador: Prof. Fernando Antonio Nóbrega Santos

Recife
2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

A524t Amorim Filho, Edgar Corrêa de

Topological transitions on protein-protein interaction networks / Edgar Corrêa de Amorim Filho. – 2019.

54 f.: il., fig., tab.

Orientador: Fernando Antônio Nóbrega Santo.

Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN, Matemática, Recife, 2019.

Inclui referências.

1. Matemática. 2. Característica de Euler. I. Santo, Fernando Antônio Nóbrega (orientador). II. Título.

510 CDD (23. ed.) UFPE- MEI 2019-082

EDGAR CORRÊA DE AMORIM FILHO

TOPOLOGICAL TRANSITIONS ON PROTEIN – PROTEIN INTERACTION NETWORKS

Tese apresentada ao Programa de Pós-graduação do Departamento de Matemática da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutorado em Matemática.

Aprovado em: 25/02/2019

BANCA EXAMINADORA

Prof. Dr. Fernando Antônio Nóbrega Santo (Orientador)
Universidade Federal de Pernambuco

Prof. Dr. Manoel José Machado Soares Lemos (Examinador Interno)
Universidade Federal de Pernambuco

Prof. Dr. César Augusto Rodrigues Castilho (Examinador Interno)
Universidade Federal de Pernambuco

Prof. Dr. Maurício Domingues Coutinho Filho (Examinador Externo)
Universidade Federal de Pernambuco

Prof.Dr. Borko Stosic (Examinador Externo)
Universidade Federal Rural de Pernambuco

To my wife and son!

ACKNOWLEDGEMENTS

Thank CAPES and CNPq for financial support in the early years of the doctorate.

*E*dgar Corrêa de Amorim Filho

ABSTRACT

In this work, we utilized concepts of applied algebraic topology to explore the very recent ideas of topological phase transitions in complex networks to the context of the Duplication Divergence model for protein-protein interaction Network. To do so, we used methods of topological data analysis to compute the Euler characteristic analytically, and the Betti numbers numerically for two variants of the Duplication Divergence model, namely the totally asymmetric model and the heterodimerization model. We contrast our theoretical results with experimental data freely available at online repositories of gene coexpression networks of *Saccharomyces cerevisiae*, also known as baker's yeast, as well as of the nematode *Caenorhabditis elegans*. We detected one topological phase transition in Yeast networks obtained according to different similarity measures, corresponding to phase transitions at close critical thresholds. Our results give evidence that the Euler characteristic can be interpreted as an intrinsic bio-marker for Yeast networks and reinforces the hypothesis of the possibility of using topological phase transitions to build topological and geometrical biomarkers for networks more generally.

Keywords: Euler characteristic. Homology. Phase transitions. Protein Interaction Networks. Topological Data Analysis.

RESUMO

Neste trabalho, utilizamos métodos de topologia algébrica aplicada para explorar o recente conceito de transições de fase topológicas em redes complexas em modelos de interação entre proteínas. Em particular, usamos métodos de análise topológica de dados para computar analiticamente a característica de Euler e numericamente os números de Betti para duas variações do modelo de Duplicação e Divergência em redes de interação entre proteínas, a saber, o modelo totalmente assimétrico e o modelo com heterodimerização. Contrastamos nossos resultados teóricos e numéricos com dados experimentais disponíveis em repositórios online para redes de coexpressão genética da *Saccharomyces cerevisiae*, a levedura utilizada na produção do pão e de cerveja, bem como para o nematoide *Caenorhabditis elegans*. Detectamos uma transição de fase nas redes de levedura, obtidas através de diferentes medidas de similaridade, que corresponde a transições de fase em um valor de correlação crítica com poucas flutuações. Nosso resultado dá evidências de que a característica de Euler pode ser interpretada como biomarcador intrínseco para redes de interação entre proteínas e reitera a possibilidade de utilizar transições de fase topológicas para construir biomarcadores topológicos e geométricos em redes de uma maneira geral.

Palavras-chave: Caracteristica de Euler. Homologia. Transições de Fase. Rede de Iteração de Froteinas. Análise Topológica de Dados.

LIST OF FIGURES

Figure 1 – Triangulation of the sphere $\mathbb{S}^2$ through a tetrahedron. 23

Figure 2 – Filtration process of brain networks. From [1]. 29

Figure 3 – Topological transitions of the random clique complex. Numerical evaluation obtained from [1]. 32

Figure 4 – Topological phase transitions on functional brain networks [1]. Observe that the brain network of each individual (gray) presents transitions inside a narrow interval, around the average (blue). 33

Figure 5 – Duplication step of the totally asymmetric model. For each duplication step a node is selected to be duplicated (red) within its edges. Each edge (dashed lines) that goes from the replica (pink) is activated with independent probability p. 37

Figure 6 – Euler entropy $S_\chi = \log|\langle\chi\rangle|$ as function of the retention probability p of the Duplication Divergence graph with 1000 nodes. S_χ was determined by (4.1). The singularities $S_\chi \to -\infty$ locate the topological phase transitions on the duplication divergence graphs. We put the number of nodes and edges in a logarithmic scale to let visible the change of dominance between the number of nodes and the number of edges that occurs at the same spot as the phase transition. 39

Figure 7 – Betti numbers β_0 and β_1 of DD graphs as functions of the retention probability p (N = 1000). At the first phase the graph is totally disconnected, with many components. As p increases β_1 (i.e, the number of cycles) surpasses β_0 exactly when p crosses the transition point, indicating an important change on the topology of the DD graph. 40

Figure 8 – Percolation transition on DD model. The three curves shows the fraction of nodes on the biggest component N_G/N as function of the retention probability p. 40

Figure 9 – Network representation of gene co-expression network for the yeast. This image was build using the software Cytoscape [2]. The presented network have 1893 nodes and 7955 links and it was collected from Yeastnet v3 [3]. 42

Figure 10 – Euler entropy S_χ as a function of the correlation threshold ε of 20 co-expression networks from Yeastnet database. Each gray line corresponds to a co-expression network, while the blue line is the average of the gray ones. Most of the topological phase transition was detected on those networks for $\varepsilon \approx 0.876$, except for the Yeast network with the DNA damaged, where the transition happened at $\varepsilon \approx 0.573$ (red line). This data was designed intended to measure the response to DNA damage in the network [4], and the shift in the threshold of the topological transition is apparently sensitive to that response. 43

Figure 11 – Average of the Betti numbers as a function of the correlation threshold ε. Observe that, as happened with the totally asymmetric model, the dominance of β_0 and β_1 shifts at the vicinity of the singularity of the Euler entropy (black line). 44

Figure 12 – Duplication step of the heterodimerization model. For each duplication step a node is selected to be duplicated (red) within its edges. Each edge (dashed lines) that goes from the replica (pink) is activated with independent probability p. Also, an edge going from the replica to the original node (dotted line) is created with probability q. 45

Figure 13 – Average of Euler Entropy as function of the retention probability p for different values o q. Observe that, differently than the previous model, there are more transitions and their number varies with the value of q. This happens because the heterodimerization step makes possible the appearance of cliques of different sizes as observed in real PPIN [5] . . 46

Figure 14 – (top) Euler entropy of PPIN from *C. elegans* in contrast with simulation with heterodimerization model (bottom). Both have two singularities . 47

LIST OF TABLES

Table 1 – Chain groups bases. 23

Table 2 – Homology of $\mathbb{S}^2$ 25

Table 3 – Homology groups of basic 2-manifolds. 25

Table 4 – Phase Transitions on $G(n, p)$ 30

Table 5 – Phase Transitions on $Y_d(n, p)$. 31

CONTENTS

1 INTRODUCTION 12

2 BASIC ALGEBRAIC TOPOLOGY CONCEPTS 15
2.1 Simplicial complexes 15
2.2 Homology Groups and Betti Numbers 18
2.3 Computing Homology Groups 21
2.4 Discussions 25

3 ALGEBRAIC TOPOLOGY AND PHASE TRANSITIONS 27
3.1 Topological Data Analysis of Weighted Networks 27
3.2 Phase Transitions on Erdös-Rényi Model 30
3.2.1 Erdös-Rényi model 30
3.2.2 The random d-complex 31
3.3 Topological Phase Transitions in Data Driven Systems 31

4 TOPOLOGICAL PHASE TRANSITIONS ON PPIN 34
4.1 Introduction 34
4.2 From Protein Interaction Networks to Duplication-Divergence . . . 35
4.3 Phase transition on Totally asymmetric Duplication-Divergence model 37
4.4 Phase Transitions on GCN 41
4.5 Heterodimerization model. 45
4.6 Conclusions 48

REFERENCES 49

1 INTRODUCTION

Topology is the area of mathematics that deals with shapes. Two significant problems in topology are identifying properties of shapes that are preserved through deformations, stretching, and twisting, and recognize whenever two objects are deformations of each other.

Given this invariance against deformation, very recently, ideas and concept of algebraic topology started to be used to analyze comprehend data, an effervescent area that nowadays is known as Topological Data Analysis (TDA). The recent advances in experimental methods and the increasing computer power generally, modern science have to deal with the fact that data is being produced at an unprecedented rate. Moreover, data is usually complex and noisy due to the inherent nature of the experimental data. In that sense, topology and geometry come as natural tools to be applied in order to understand data generically, i.e., avoiding reductionist approaches. Mostly because topology deals with qualitative geometric information, such as connectivity, topological invariants, etc., but also because topology is robust against noise or perturbations since noise can be seen as deformations on the shape of data [6]. This scientific revolution opened perspective to experiment several subtopics of topology into to data, which make TDA a very active and creative field nowadays.

Topology can also be an innovative tool to understand several systems whose dynamics is usually unknown, the so-called complex systems. In many cases, it is useful to represent complex systems as graphs or networks, referred here as complex networks, where the nodes represent the elements of a complex system, and the edges the relations between them. Examples of complex networks appear in many contexts, like the internet, financial system, biology, to name a few [7]. Given that the exact dynamics of such systems are unexplained, we often can extract significant, quantitative information by associating a network to a complex system. However, the benefits of associating a graph to data are limited to its low dimensional aspect. The recent use of algebraic topology, by associating a simplicial complex to a network, has been proved to be useful to address this limitation [6, 8]. With the use of algebraic topology in data, we are able to find real-world applications in various interdisciplinary fields, ranging from material science [9], breast cancer diagnosis [10], neuroscience [1], to computer vision [11], for example.

On the other hand, topology has contributed to the understanding of several phenomena in physics from quantum field [12] to condensed matter physics [13, 14]. One topic that is studied both in the context of theoretical physics and algebraic topology is phase transitions [15, 16]. In the context of statistical mechanics of Hamiltonian systems, phase transitions can also be detected using topology, particularly Morse theory. In fact, in several works, the Euler characteristic plays an essential whole to detect classical phase transitions [17–21]. From the mathematical point of view, in particular in graph theory, the most well-known phase transition is the formation of a giant component in random graphs [22], i.e., a percolation transition in a random graph. Although the percolation problem is well established in graph theory, the extension of this problem to more complex structures, in particular to a simplicial complex, is a very active area of the algebraic topology [23, chapter 22], in particular, there are very recent results that generalized the ideas of percolation transitions by Erdos and Reyni for simplicial complexes [24].

Very recently, concepts of TDA, network science, and theoretical physics were merged to develop a methodology to detect phase transitions on complex networks, particularly in neuroscience [1]. This recent work opened the perspective of obtaining reliable topological biomarkers for brain organization.

In this thesis, we will apply the ideas discussed above to the context of Protein-Protein Interaction Networks (PPIN), i.e., networks where the nodes represent proteins, and the edges represent relations between them. These relations are usually physical interactions, but can also be functional. Topological properties of PPIN have been subject to many relevant studies nowadays [25–27]. An interesting problem in genetics is to identify which genes correspond to a given phenotype. With co-expression networks, for example, we can identify groups of proteins that share common expression patterns through various experiments, raising the hypothesis that the genes regulated by them share a common functionality [28].

We will focus our attention on two different PPIN models, namely: totally asymmetric PPIN model and the heterodimerization model. In the first case, we computed the Euler characteristic analytically and detected a phase transition that corresponds with the percolation transition, as reported in different models by [22]. We also contrasted our results with gene co-expression networks from yeast freely available in the Yeastnet v.3 [3] database and observed topological transitions consistent with the model.

Yeast is an ideal model for studying biological systems, specially because of the easiness to connect genes and proteins with the functions they represent in the cell. Since the publication of the genome sequence of *S. cerevisiae* on April 1996, has been responsible of several advances and for the establishment of research areas like Functional Genomics and Systems Biology. Also, nearly 1000 yeast genes are members of orthologous gene families associated with human disease, which turns useful for the assessment of function of human proteins associated disease state [29].

Later in this thesis, we analyze the phase transitions numerically for the heterodimerization models. In this case, we compared our results with data from *C. elegans* collected online on Wormnet database [30]. *C. elegans* a nematode which is a model system to study animal development and behavior [31].

We stress that It is not our intention to be exhaustive, since Its an interdisciplinary and active topic in the interface of Topology, Theoretical Physics, Data analysis, and Complex systems.

This thesis is written as follows. The next chapter is dedicated for algebraic topology concepts which will be of major importance for our analysis. In particular, we will focus on the homology groups of a simplicial complex, and how to compute it for data through the use of boundary matrices. In chapter 3 we will discuss briefly on how topology have been used to improve our understanding about data and digress on how some rigorous results on phase transitions in simplicial complexes and theoretical physics could be transposed to data-driven problems; finally, in chapter 4, we will apply the methodology developed to the PPIN models.

2 BASIC ALGEBRAIC TOPOLOGY CONCEPTS

In this chapter, we introduce basic algebraic topology concepts that will be used in this work. We will digress about simplicial complexes and its geometric realization, and later we introduce homology groups. These concepts will form the basis for our topological approach to data.

2.1 Simplicial complexes

The central concept of this section is the simplicial complexes. Here, we will define the fundamental structures of the simplicial complexes and present the Euler characteristic. The idea is that a simplicial complex is a simple way to visualize topological spaces. The "atoms" of a simplicial complex are the k-simplices, which are the simplest structures of the Euclidean space. We will start by the geometric definition.

Let $S = \{x_0, x_1, x_2, \ldots, x_k\}$ a discrete set of points in the Euclidean space $\mathbb{R}^d$. We say that a point $x \in \mathbb{R}^d$ is an **affine combination** of S if there exist real numbers $\lambda_0, \lambda_1, \ldots, \lambda_k$ that adds exactly 1, and x can be written as

$$x = \sum_{i=0}^{k} \lambda_i \cdot x_k.$$

An affine combination is said a **convex combination** if $\lambda_i \geqslant 0$ for every i. The numbers $\lambda_0, \lambda_1, \ldots, \lambda_k$ are called the **barycentric coordinates** of x. To the set of all points that are convex combinations of S, we give the name **convex hull** of S. If no point x_i of S is an affine combination of $S - \{x_i\}$, we say that S is an **affinely independent** set.

Example 2.1 (k=1)**.** Suppose that $S = \{x_0, x_1\} \subset \mathbb{R}^2$ where $x_0 \neq x_1$. The line that goes passes through x_0 and x_1 is the set of points that can be written as a function of a parameter λ as

$$x_0 + \lambda(x_1 - x_0) = (1 - \lambda)x_0 + \lambda x_1.$$

Observe that the points on the line segment connecting x_0 to x_1 are obtained by setting $0 \leqslant \lambda \leqslant 1$. If we name $\lambda_0 = \lambda$ and $\lambda_1 = 1 - \lambda$ it gets clear that $\lambda_0, \lambda_1 \geqslant 0$ and $\lambda_0 + \lambda_1 = 1$. So the convex hull of S is the closed segment line connecting x_0 to x_1.

Example 2.2 (k=2)**.** Now let's suppose that $S = \{x_0, x_1, x_2\} \subset \mathbb{R}^2$. By the previous example, we can see that if x_2 is in the line that contains x_0, x_1 ($x_1 \neq x_2$), for example, than x_2 would be an affine combination of $S - \{x_2\} = \{x_0, x_1\}$. So for S to be an affinely independent set, x_0, x_1 and x_2 cannot be in the same line, i.e., they need to be vertices of a triangle. In that case, it is an interesting linear algebra exercise to check that the convex hull of S is the set of points that are on the interior of the triangle with vertices at x_0, x_1

and x_2 (including the edges end the vertices of the triangle). Also, any set with 4 or more elements in $\mathbb{R}^2$ is not affinely independent.

A ***k*-simplex** is the convex hull of an affinely independent set S with $k+1$ points. Each point of S is said to be a **vertex** of the k-simplex. In simpler terms, a k-simplex is the smallest convex set of the Euclidean space $\mathbb{R}^d$ that contains $k+1$ points that do not belong to a k-dimensional hyperplane. A k-simplex is a k-dimensional topological subspace of a Euclidean space, i.e., the **dimension** of a k-simplex σ is $\dim \sigma = k$.

Name σ the k-simplex with vertices in S. We say that τ is a **face** of σ if τ is also a simplex whose vertices are at $T \subseteq S$. When τ is a face of σ it is often said that σ is a **coface** τ. This relation will be indicated by $\tau \leqslant \sigma$.

Now we have the necessary background to define a simplicial complex.

Definition 2.3. A **simplicial complex** is a finite set X of simplices that is closed under taking intersections and faces, i.e., that satisfies the conditions:

a. If $\sigma \in X$ and $\tau \leqslant \sigma$, then $\tau \in X$;

b. Whenever σ_1 and σ_2 are in X, then $\sigma_1 \cap \sigma_2$ is either empty or a face of σ_1 and σ_2.

The **dimension** of X is the dimension of the largest simplex in X, i.e.,

$$\dim X = max\{\dim \sigma | \sigma \in X\}.$$

A simplicial complex can be seen as a finite union of simplices that intersects themselves in other simplices (or nothing).

Although the above definition is very easy to be visualized, it is not very proper for computations. The abstract definition that we are going to present next is more suitable for that propose.

Definition 2.4. Given a set K, an **abstract simplicial complex** of K is a collection $\mathcal{S}$ of subsets of K such that:

a. $\{v\} \in \mathcal{S}$ for all $v \in K$;

b. If $\tau \subseteq \sigma$, then $\tau \in \mathcal{S}$.

The unitary subsets of K are called **vertices**. The elements of $\mathcal{S}$ are called **abstract simplices**. A simplex σ is said to be a ***k*-simplex**, or a simplex with **dimension** k ($\dim \sigma = k$) if σ has $k+1$ elements. The simplex τ is a **face** of σ (or σ is a **coface** of τ) if $\tau \subseteq \sigma$.

Given a geometric simplicial complex X, there is a natural way to build an abstract simplicial complex $\mathcal{S}$ associated with it. For every simplex σ of X, consider $\{v_0, v_1, \ldots, v_k\}$ the set o vertices of σ. Now define

$$\mathcal{S} = \{\{\text{vertices of } \sigma\} \mid \sigma \in X\}.$$

It is easy to check that $\mathcal{S}$ is an abstract simplicial complex. This natural representation of X through an abstract simplicial complex $\mathcal{S}$ is called a **vertex scheme** of X. The vertex scheme allow us to easily compare simplicial complexes.

We can use simplicial complexes to represent manifolds and other topological spaces by its underlying space. The **underlying space** $|X|$ of a simplicial complex X is $|X| = \cup_{\sigma \in K} \sigma$. By itself, $|X|$ is a topological space. When a topological space $\mathbb{X}$ is homeomorphic to the underlying space of a simplicial complex X, we say that $|X|$ is a **triangulation** of $\mathbb{X}$. A classical example of triangulation is the one of the sphere that can be triangulated as a tetrahedron (figure 1).

Definition 2.5. We say that the geometric simplicial complexes X_1 and X_2 are **isomorphic** if there is an homeomorphism between $|X_1|$ and $|X_2|$.

Definition 2.6. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ abstract simplicial complexes with vertices in V_1 and V_2, respectively. We say that $\mathcal{S}_1$ and $\mathcal{S}_2$ are **isomorphic** if there exists a $1-1$ function $\phi : V_1 \to V_2$ such that, changing all the vertices by their image by ϕ in each member of $\mathcal{S}_1$ we obtain $\mathcal{S}_2$. The function ϕ is said to be an **isomorphism**.

The next theorem, whose proof can be found in [32], will allow us to abandon the terms "geometric" and "abstract" and call everything **simplicial complex**.

Theorem 2.7 (Geometric realization theorem)**.** *Every abstract simplicial complex $\mathcal{S}$ is isomorphic to the vertex scheme of some geometric simplicial complex X. And two geometric simplicial complexes are isomorphic if and only if their vertex schemes are isomorphic as abstract simplicial complexes.*

If an abstract simplicial complex $\mathcal{S}$ is isomorphic to the vertex scheme of some geometric simplicial complex X, we say that X is a **geometric realization** of $\mathcal{S}$. It is uniquely determined up to linear isomorphism.

This way we can see the simplicial complexes as structures that can be seen by two manners. Formally, it is a purely combinatorial object that can be easily manipulated by computers. But it is also a map of the vertices into a space in which it is realized.

The concept of triangulation is very important for computation proposes, for example. If you want to render a surface into the computer, it would cost too an infinite amount of time and processing to compute the properties of every point in this surface. So, to turn this task possible and fast to achieve one can use a triangulation that approximates the surface in average and render the triangulation instead. Also, when you have a triangulation of a topological space $\mathbb{X}$, you have a purely combinatorial representation of $\mathbb{X}$ that, at least for our proposes, will be a lot easier to analyse.

An important problem is to classify spaces based on their fundamental properties. A common tool for differentiating between spaces is an **invariant**, i.e., a map that assigns the same object to spaces of the same topological type. It is possible that an invariant assign the same object to spaces of different types. But the power of the topological invariants lies on the fact that even though we cannot say much about spaces that are assigned to the same object, whenever two spaces are assigned to different objects you have sure that they belong to different topological types. The measure of "goodness" of an invariant is exactly the number of spaces that you are able to classify with it.

One important invariant is the so-called Euler characteristic, that is particularly important for the classification of surfaces, for example. It was originally defined for graph by Euler.

Definition 2.8. Let $\mathcal{S}$ be a simplicial complex. Let's call s_k the total number of k-complexes in K. The topological invariant called **Euler characteristic** is the number $\chi(\mathcal{S})$ defined as

$$\chi(\mathcal{S}) = \sum_{k=0}^{\dim \mathcal{S}} (-1)^k s_k.$$

It is a known fact that if $|\mathcal{S}_1|$ and $|\mathcal{S}_2|$ are triangulations of the same space $\mathbb{M}$, then $\chi(\mathcal{S}_1) = \chi(\mathcal{S}_2)$. So we can say that the Euler characteristic of a triangulation of a space $\mathbb{M}$ is the Euler characteristic of the space $\chi(\mathbb{M})$

2.2 Homology Groups and Betti Numbers

Before introducing homology groups, we must first define orientation in the context of simplicial complexes.

Definition 2.9. Let σ be a simplex with vertexes $v_0, v_1, \dots, v_k$. We say that two ordering of the vertices are equivalent, and indicate by

$$(v_{i_0}, v_{i_1}, \dots, v_{i_k}) \sim (v_{j_0}, v_{j_1}, \dots, v_{j_k}),$$

if they differ only by an even permutation of the vertices. This relation divides the set of all possible orderings into two equivalence classes. Each one of this classes is said to be an **orientation** of σ. An **oriented simplex** is a simplex together with an orientation of its vertices.

For simplicity, will use the indication

$$v_0 v_1 \dots v_k$$

for the simplex with vertices $v_0, v_1, \dots, v_k$ (independent), and

$$[v_0, v_1, \dots, v_k]$$

for the oriented simplex with the same vertices and orientation indicated by the given ordering of vertices.

Definition 2.10. The ***p*th chain group** of a simplicial complex $\mathcal{S}$ is the free Abelian group $C_p(\mathcal{S})$ on the oriented p-simplices, where $[\sigma] = -[\tau]$ whenever $\sigma = \tau$ and have different orientations. Every element of $C_p(\mathcal{S})$ is called a ***p*-chain**, and has always the form

$$\sum_q n_q[\sigma_q],$$

where the numbers n_q are integers and σ_q are p-simplices. For convenience, if $p < 0$ or $p > \dim \mathcal{S}$, we define $C_p(\mathcal{S})$ as the trivial group.

Homology groups are structures that concerns about the connectivity between two immediate dimensions. So it is necessary to build relations between those p-chain groups. For that we have the boundary operators.

Definition 2.11. Let $\mathcal{S}$ be a simplicial complex and $\sigma \in \mathcal{S}$ the simplex given by $\sigma[v_0, v_1, \dots, v_p]$. The **boundary** of σ, denoted by $\partial\sigma$ is the element of $C_{p-1}(\mathcal{S})$ given by

$$\partial\sigma = \sum_{i=0}^{p} (-1)^i [v_0, v_1, \dots, \hat{v}_i, \dots, v_p],$$

where $\hat{v}_i$ indicates that v_i was deleted from the sequence. The boundary operator defines an homomorphism $\partial_p = \partial|_{C_p(\mathcal{S})} : C_p(\mathcal{S}) \to C_{p-1}(\mathcal{S})$.

It is easy to check that the boundary of a simplex, as defined above, does not depend on the particular choice of orientation.

Example 2.12 (Boundary of some simplices)**.** :

- $\partial[a, b] = [b] - [a]$;
- $\partial[a, b, c] = [b, c] - [a, c] + [a, b]$;
- $\partial[a, b, c, d] = [b, c, d] - [a, c, d] + [a, b, d] - [a, b, c]$.

Observe that the boundary of a p-simplex is exactly the $(p-1)$-dimensional faces of it, with an orientation that is naturally given by the formula. If we take the boundary of the boundary, we have the result given by the next theorem.

Theorem 2.13. $\partial(\partial\sigma) = 0$ *for every oriented simplex* σ.

Proof. Let $\sigma = [v_0, v_1, \dots, v_k]$. So

$$\begin{aligned} \partial(\partial\sigma) &= \partial\left(\sum_i (-1)^i [v_0, \dots, \hat{v}_i, \dots, v_k]\right) \\ &= \sum_{j<i} (-1)^i(-1)^j [v_0, \dots, \hat{v}_j, \dots, \hat{v}_i, \dots, v_k] \\ &\quad + \sum_{j>i} (-1)^i(-1)^{j-1} [v_0, \dots, \hat{v}_j, \dots, \hat{v}_i, \dots, v_k]. \end{aligned}$$

Observe that switching i and j at the second sum, every term at the first sum will also be on the second, but with a minus sign. So $\partial(\partial\sigma) = 0$. □

If $\mathcal{S}$ is an n-dimensional simplicial complex, the boundary operator give birth to the chain of homomorfisms that are illustrated bellow:

$$0 \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \dots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0.$$

The theorem 2.13 show that the composition map of two consecutive homomorfisms is trivial, i.e., $\partial_k \partial_{k+1} = 0$. Observe that since $\mathcal{S}$ has no $(n+1)$-dimensional simplices then C_{n+1} is the trivial group. Observe also that since vertices have no boundary, we could set the map ∂_0 as the trivial homomorphism. Such sequence of homomorphisms between chain groups is called a **chain complex**.

Follow immediately from theorem 2.13, and from properties of homomorphisms, the next theorem.

Theorem 2.14. *The sets* $\operatorname{Im}\partial_{k+1}$ *and* $\operatorname{Ker}\partial_k$ *are free Abelian normal subgroups of* C_k. *Also* $\operatorname{Im}\partial_{k+1}$ *is a normal subgroup of* $\operatorname{Ker}\partial_k$.

Proof. The proof of the first statement comes from a collection of properties of homomorphisms and will be omitted.

For the second part, observe that the boundary of each $(k+1)$-simplex is a k-chain that goes to 0 by ∂_k. So the homomorphism ∂_{k+1} sends $k+1$-chains into the kernel of ∂_k, i.e., $\operatorname{Im}\partial_{k+1} \subseteq \operatorname{Ker}\partial_k$. Since those sets are free Abelian groups of C_k, turns out that $\operatorname{Im}\partial_{k+1}$ is actually a normal subgroup of $\operatorname{Ker}\partial_k$. □

Definition 2.15. The **kth cycle group** is $Z_k = \operatorname{Ker}\partial_k$. The elements of Z_k a called **k-cycles**. The **kth boundary group** is $B_k = \operatorname{Im}\partial_{k+1}$. The elements of B_k are called *$k-$boundaries*.

Definition 2.16. The **kth homology group** of a simplicial complex, H_k, is the quotient of kth cycle group over the kth boundary group, i.e.,

$$H_k = Z_k/B_k = \operatorname{Ker} \partial_k / \operatorname{Im} \partial_{k+1}.$$

If the k-cycles z_1 and z_2 belong to the same class of Z_k, we say that they are **homologous** and denoted by $z_1 \sim z_2$.

As they are factor groups of two free Abelian groups, the homology groups are finitely generated Abelian. Therefore, the fundamental theorem of finitely generated Abelian groups applies. So we can use the homology groups to describe spaces through topological invariants called Betti numbers.

Definition 2.17. The ***k*th Betti number** β_k of a simplicial complex K is the rank of the free part of its kth homology group, i.e.,

$$\beta_k = \operatorname{Rank} H_k = \operatorname{Rank} Z_k - \operatorname{Rank} B_k.$$

The Betti numbers and the Euler characteristic of a simplicial complex $\mathcal{S}$ are related by the formula

$$\chi(\mathcal{S}) = \sum_{k=0}^{n} (-1)^k \beta_k.$$

So, the Euler characteristics is the alternating sum of the Betti numbers. This result is known as Euler-Poincaré Formula.

2.3 Computing Homology Groups

Once the main goal of this work is to use Euler characteristic and Betti numbers to obtain information on real data, an important task is to compute the homology groups. For that we use the so called **boundary matrices**.

Let $\mathcal{S}$ be a simplicial complex. As the p-chains are free Abelian groups, the set of oriented p-simplices form the **standard basis** for for it. The **standard matrix representation** of ∂_p is a representation of the boundary operator $\partial_p : C_p \to C_{p-1}$ relative to the standard bases of the chain groups C_p and C_{p-1}. The boundary matrix $M_p = [a_i^j]$ is a $(s_{p-1} \times s_p)$ matrix with coefficients in $\{-1, 0, 1\}$. Given standard bases of oriented simplices, lets say $(\sigma_1, \sigma_2, \dots, \sigma_{s_p})$ of p-simplices and $(\tau_1, \tau_2, \dots, \tau_{s_{p-1}})$ of $(p-1)$-simplices, and given a p-chain $c = \sum a_i \sigma_i$, the boundary of c will be of the form $\partial_p c = \sum b_j \tau_j$, where the b_j are given by

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{s_{p-1}} \end{bmatrix} = \begin{bmatrix} a_1^1 & a_1^2 & a_1^3 & \dots & a_1^{s_p} \\ a_2^1 & a_2^2 & a_2^3 & \dots & a_2^{s_p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{s_{p-1}}^1 & a_{s_{p-1}}^2 & a_{s_{p-1}}^3 & \dots & a_{s_{p-1}}^{s_p} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{s_p} \end{bmatrix}.$$

The null-space of M_p corresponds to Z_p, while its range-space corresponds to B_{p-1}. The next step is to reduce M_p to its diagonal form. For that we use a **reduction algorithm** that consists of applying the following **elementary row (or column) operations**:

- switch the places of two rows;
- multiply a row by -1;
- replace row i by (row i) $+ q$(row j), for an integer q and $j \neq i$.

Each of this elementary row operations are equivalent to making a change of basis for C_{p-1}. The reduction algorithm repeats systematically those elementary operations until M_p is reduced to the form:

$$\tilde{M}_p = \left[\begin{array}{cccc|c} b_1 & 0 & \dots & 0 & \\ 0 & b_2 & \dots & 0 & \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & b_{l_p} & \\ \hline & & & & \\ \multicolumn{4}{c|}{0} & 0 \\ & & & & \end{array}\right].$$

$\tilde{M}_p$ is called the **(Smith) normal form** of M_p. The number of non-zero rows l_p of $\tilde{M}_p$ is the rank of M_p. The normal form indicates explicitly what we need to compute the homology groups.

- The **torsion coefficients** of H_{p-1} are the b_i's that are greater than 1;
- Rank Z_p is the number of null rows of $\tilde{M}_p$. So Rank $Z_p = s_p - l_p$;
- The first l_p columns forms a basis for B_{p-1}. Therefore, Rank $B_p =$ Rank $M_{p+1} = l_{p+1}$.

Thus, by definition 2.17 we have

$$\beta_p = \text{Rank}\, Z_p - \text{Rank}\, B_p = s_p - l_p - l_{p+1}. \tag{2.1}$$

In example 2.18 bellow we will see how to compute the boundary matrices the homology groups of the sphere $\mathbb{S}^2$.

Example 2.18. Recall that the sphere $\mathbb{S}^2$ can be triangulated as a tetrahedron. So, at first, we have to describe the tetrahedron by its oriented simplices. Figure 1 shows the triangulation of the sphere through a tetrahedron.

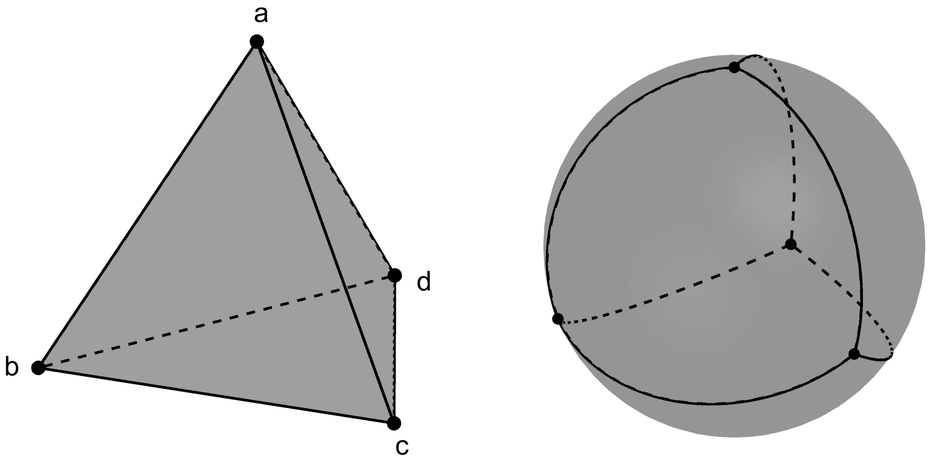

Figure 1 – Triangulation of the sphere $\mathbb{S}^2$ through a tetrahedron.

C_0	$\{[a],[b],[c],[d]\}$
C_1	$\{[a,b],[a,c],[a,d],[b,c],[b,d],[c,d]\}$
C_2	$\{[a,b,c],[a,b,d],[a,c,d],[b,c,d]\}$

Table 1 – Chain groups bases.

Table 1 presents a basis for the 0, 1 and 2-chain groups. Computing the boundary of each simplex on the bases, we can build the boundary matrices:

$$M_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}, \qquad M_1 = \begin{bmatrix} -1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & -1 & 0 \\ 0 & 1 & 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix},$$

$$M_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \qquad M_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Applying elementary row operations we reduce the boundary matrices to their normal forms that are, respectively:

$$\tilde{M}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}, \quad \tilde{M}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\tilde{M}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \tilde{M}_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The python algorithm below describes how we can achieve computationally the boundary matrices of a simplicial complex.

```
import numpy as np
def Boundary_matrix(SComplex, k):
  # k-th Chain complex
  source = tuple([s for s in SComplex if len(s) == k+1])
  # (k-1)-th Chain complex
  target = tuple([s for s in SComplex if len(s) == k])
  if k == 0:
    M = [[0] for simplex in source]
  elif k >= max([len(simplex) for simplex in SComplex]):
    M = [[0 for simplex in target]]
  else:
    M = []
    for simplex in source:
      # Boundary initiated with all coefficients being 0
      boundary = dict()
      for i in target:
        boundary[i] = 0
      # Variable to control the coefficients of the boundary
      sign = -1
      for i in range(len(simplex)):
        # At each step, remove the ith element of simplex
        aux = simplex[:i]+simplex[i+1:]
        # and multiply the coefficient by -1.
        sign *= -1
```

```
            boundary[aux]=sign
        M.append(list(boundary.values()))
    return np.array(M).T.tolist()
```

So, with the boundary matrices we can easily calculate the rank of the boundary and cycle groups and, with formula 2.1, obtain the Betti numbers of $\mathbb{S}^2$, as presented on the table bellow:

	$p = 0$	$p = 1$	$p = 2$
Rank Z_p	4	3	1
Rank B_p	3	3	0
β_p	1	0	1

Table 2 – Homology of $\mathbb{S}^2$

Even though the necessary information to compute the Betti numbers comes from the normal form of the boundary matrix, it is not actually necessary to achieve the normal form through elementary row operations. For computational proposes, the python's package **numpy** provides us with optimized linear algebra algorithms to compute the rank of the boundary matrices and the Betti numbers.

The next section will be dedicated to discuss about the description that the homology groups provide.

2.4 Discussions

Established the definitions and methods to calculate the homology group of simplicial complex, let's understand what does homology tells us. With that in mind, let's check the examples on table 3 bellow. The table shows the homology groups of triangulations of a few basic 2-manifolds. Those groups can be easily calculated using the method presented above and definition 2.16 and does not depend on the chosen triangulation.

2-manifold	H_0	H_1	H_2
Sphere	$\mathbb{Z}$	$\{0\}$	$\mathbb{Z}$
Torus	$\mathbb{Z}$	$\mathbb{Z} \times \mathbb{Z}$	$\mathbb{Z}$
Projective plane	$\mathbb{Z}$	$\mathbb{Z}_2$	$\{0\}$
Klein bottle	$\mathbb{Z}$	$\mathbb{Z} \times \mathbb{Z}_2$	$\{0\}$

Table 3 – Homology groups of basic 2-manifolds.

Because they are all 2-manifolds, their homology groups will be all trivial, except for the ones presented.

Torsion-free spaces are the ones the have homology that does not have terms that are finite cyclic groups $\mathbb{Z}_m$. For torsion-free spaces in three dimensions (like the

2-manifolds) the Betti numbers have intuitive interpretations that comes from **Alexander duality**. β_0 measures the number of components of the complex. β_1 is the rank of a basis for the **tunnels**, i.e., the number of cycles that are not boundary. β_2 is the number of **voids**, i.e, empty spaces that are enclosed by the complex.

Now, looking back at the table 3, observe that all of the spaces have a single component, so $H_0 = \mathbb{Z}$ and $\beta_0 = 1$. Both the sphere and the torus enclose an empty space, so $H_2 = \mathbb{Z}$ and $\beta_2 = 0$. The projective space and the Klein bottle otherwise are nonorientable, which means that those surfaces have a single side. So these surfaces do not enclose any void and their 2nd homology groups are trivial ($\beta_2 = 0$).

For H_1 observe that every simple closed curve drawn on the sphere is isomorphic to the boundary of a 2-simplex. So its first homology group is trivial ($\beta_1 = 0$). But the torus, otherwise, have two classes of cycles that are not boundaries. Therefore $H_1 = \mathbb{Z} \times \mathbb{Z}$ and $\beta_1 = 2$. Even though its harder to visualize it for harder dimensions, the Betti numbers β_n can be seen as the number of n-dimensional holes of the space.

3 ALGEBRAIC TOPOLOGY AND PHASE TRANSITIONS: FROM THEORY TO DATA

In this chapter, we discuss about a recent approach to detect topological phase transitions in experimental data, particularly to weighted networks. In a nutshell, one can associate a weighted graph (or network) to experimental data. Later, we can compute the simplicial complex associated to data. With the simplicial complex in hand, we can make use of several tools and results from algebraic topology directly into data. That's the main idea of Topological Data Analysis (TDA). Since TDA is a very recent topic in applied mathematics, several subtopics of Algebraic Topology are still under-exploited into the context of data. On the other hand, some methodologies are already widely spread into the artificial intelligence entrepreneurship, for instance, by using of the Mapper algorithm and persistent homology techniques. Very recently, phase transitions in random simplicial complexes [23, chapter 22], a very active area in Algebraic topology, was also explored in the context of data [1], with a deep inspiration in results relating phase transitions in theoretical physics [17, 20, 33]. This chapter will pave the way to implement our approach to protein-protein interaction networks in the next chapter.

3.1 Topological Data Analysis of Weighted Networks

Data is everywhere. From medicine to the stock market, or from astrophysics to social media, the amount of data we produce is growing exponentially over the last years. According to [34], in 2018, humanity produced about 2.5 quintillion bytes of data every day. This data revolution gave us several benefits, such as predictability on the weather, diagnostic of diseases, music streaming, etc. On the other hand, this data revolution sometimes gives us with the impression that the scientific method is not being successful to achieve a theory for data in the same way we have established theories in physics and mathematics [35]. Topological data analysis (TDA) is a generic approach to process and understand data and, therefore, is a strong candidate to circumvent this causality issue in modeling data, which can make data science closer to the standard scientific method. This large amount of data requires consistent techniques to be explained to obtain information from it. Although other approaches have been reported in the past, TDA became famous as a research topic through the works of Edelsbrunner [8], Zomorodian [36], and Carlsson [6]. TDA consists of giving shape to data, that usually comes in the form of a set of points, and read the data through the topological and geometrical properties of this shape.

Most of the methods inspired by topological and geometrical properties adopt the following pipeline:

1. The data is assumed to be a set of point distributed on a metric space, or with intrinsic

pair-wise connectivity defined by an incidence matrix. The metric or connectivity defined for the data is usually given experimentally or built by some procedure.

2. A chain of simplicial complexes called a **filtration**, is build on top of the data. This structure reflects the data on different scales. The main difficulty here is how to define the filtration in order to achieve relevant information about the data in a way that can be easy to built and manipulate it.

3. Topological invariants and geometric information are extracted from the data, and this knowledge can be seen as data's fingerprint. They can be used to understand the data better or to be combined with other features for further analysis.

The data that we will analyze in this work comes as a **weighted network**, i.e., a graph that for each edge we associate a real number, often referred to as a **weight**. This weight is usually a normalized measure of similarity or the force of some functional interaction between the nodes. Data like this appears in a vast range of contexts that go from biological networks, like protein interaction or brain networks, to social networks.

A filtration process can be naturally built on weighted networks. In order to implement it, we can see the weights as a distance between the nodes and apply the pipeline described above. Let $G = (N, E)$ be a weighted graph with vertices in $N = \{1, 2, 3, 4, \ldots, n\}$. Suppose that each edge (i, j) have weight $w(i, j) \in [0, 1]$. For each $\varepsilon \in [0, 1]$ define $G_\varepsilon = (N, E_\varepsilon)$ as the graph with same vertices as G and $E_\varepsilon = \{(i, j) \in E \mid w(i, j) \leqslant \varepsilon\}$. To each G_ε we can associate a simplicial complex $X_\varepsilon = X(G_\varepsilon)$ that is the **clique complex** (or **flag complex**) of G_ε. A **clique** in a graph G_ε is a subset of the vertices of G_ε in which every vertex is connected with each other by an edge of G_ε. The clique complex of G_ε is the simplicial complex $X(G_\varepsilon)$ in which every clique of G_ε is a simplex in $X(G_\varepsilon)$.

The filtration process generates a function $\mathcal{F} : \varepsilon \mapsto X_\varepsilon$ that traces the topology of the data on different scales. $\mathcal{F}(0) = X_0$ is the complex containing only the vertices (0-simplices). As ε increases the edges of G are being added, increasing the connectivity of the network and building higher dimensional simplices. Figure 2 illustrates the process.

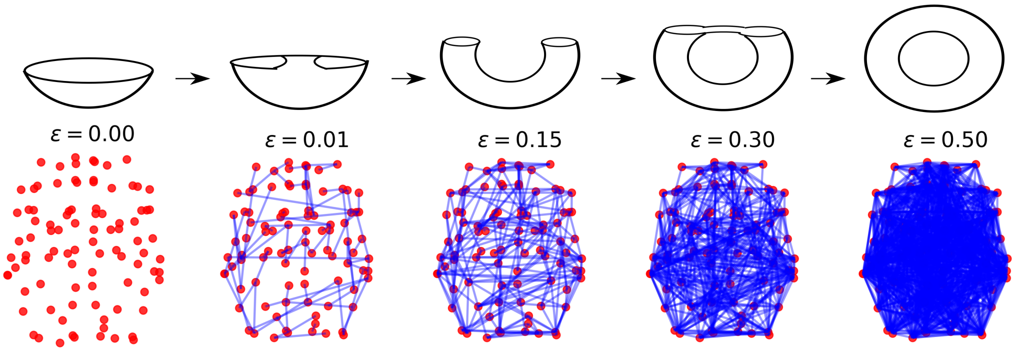

Figure 2 – Filtration process of brain networks. From [1].

The topological properties of X_ε as ε increases can give us reliable information about the data. Given the geometric realization theorem 2.7, this process is equivalent to building a hyperpolyhedra adding faces from lower to higher dimensions, where each k-clique in the network can be seen as k-dimensional face. One may ask, how the topology of this geometrical simplicial complex evolve during the filtration.

In this work, we propose the usage of the techniques of TDA in order to detect phase transitions on real data, that comes in the form of a weighted network. Phase transitions have been studied in a vast range of systems and techniques in theoretical physics [15, 16], and particularly for classical Hamiltonian systems using methods of Morse theory [17, 20, 33]. Particularly, the association between Euler characteristics and percolation transitions appear in many theoretical models [37–42].

In graph theory, the most well-known phase transition is the so-called giant component transition in random graphs studied by Erdös and Rényi in [22]. This transition is also known as percolation transition and is already applied in several data-driven problems [43]. Although the percolation problem is well established in graph theory, the extension of this problem to a simplicial complex is quite recent in algebraic topology [23, chapter 22] and even to data. In fact, there are quite new results that generalized the ideas of percolation transitions for simplicial complexes [24].

The current, state of art, most spread methodologies in TDA, are the Mapper algorithm [44] and persistent homology [45]. Several subtopics of algebraic topology, however, are potential candidates for new topological approaches to understand and interpret data. Very recently, the ideas of TDA, network theory, phase transitions in Hamiltonian systems, and phase transition in simplicial complexes were merged in order to detect phase transitions in data generally, particularly in functional brain networks [1].

In the next section, we will briefly discuss phase transitions from the Erdös-Rényi model to some random simplicial complexes, in order to move forward to detect phase transitions in data.

3.2 Phase Transitions on Erdös-Rényi Model

Random shapes appear naturally in many contexts and the concept of **phase transition** have been broadly studied for theoretical models. In this section we will resume some topological and geometrical properties of random simplicial complexes. We will start by briefly introducing some commonly studied models and point a few topological transitions detected on those models.

The study of random graphs was started by Erdös and Rényi in the 1960's and had impact on many areas of science, like discrete mathematics, computer science and engineering. During the following years many topological properties had been analysed for the Erdös-Rényi graph (that will be defined latter), and since graphs are purely 1-dimensional simplicial complexes, people started to think about what properties should emerge if we consider more general structures like d-dimensional random simplicial complexes. This section is devoted to present the Erdös-Rényi graph and some usual generalizations to higher dimensional simplices.

3.2.1 Erdös-Rényi model

Let n be a positive integer and $p \in \mathbb{R}$, $0 \leqslant p \leqslant 1$. The **Erdös-Rényi graph** is a graph with n vertices in which every edge is added independently with probability p. We usually think of p a function of n to study asymptotic properties of $G(n,p)$. An event is said to happens **with high probability (w.h.p.)** if its probability goes to 1 as $n \to \infty$.

One of the main topics about $G(n,p)$ is the search for **threshold** probabilities. Given a property $\mathcal{P}$, we want to find a critical value $\bar{p} = \bar{p}(n)$ for which if $p > \bar{p}$, then property $\mathcal{P}$ holds w.h.p., and if $p < \bar{p}$, then property $\mathcal{P}$ does not hold w.h.p.

Erdös-Rényi [22] found a sharp threshold for connectivity. They proved that for $\bar{p} = \log n/n$ the following statement holds:

Proposition 3.1. *For every $\varepsilon > 0$, if $p \geqslant (1+\varepsilon)\bar{p}$ then $G(n,p)$ is connected w.h.p. Also, if $p \leqslant (1-\varepsilon)\bar{p}$ then $G(n,p)$ is disconnected w.h.p.*

Some other important topological threshold found on $G(n,p)$ are listed bellow.

Property	Threshold ($\bar{p}$)	Reference
$H_1(G) \neq 0$	$1/n$	[46]
G is not planar	$1/n$	[47]
G is purely 1-dimensional	$\log n/n$	[22]
G have a giant component	$1/n$	[22]

Table 4 – Phase Transitions on $G(n,p)$

3.2.2 The random d-complex

Many generalizations of the Erdös-Rényi model have been studied, but since we are interested in topological properties of graphs in the context of simplicial complexes the following models should fit better.

The **random 2-complex** $Y_2(n,p)$ consists of a simplicial complex with n vertices, all $\binom{n}{2}$ possible 1-simplices and each of the $\binom{n}{3}$ possible 2-simplices is included independently with probability p. For **random d-complex** $Y_d(n,p)$ we start with with n vertices and all $(d-1)$-simplices. Every d-dimensional face appears with probability p.

Property	Threshold ($\bar{p}$)	Reference
$H_{d-1}(Y_d) \neq 0$	$O(\log n/n)$	[48]
Y_d is purely d-dimensional	$d \log n/n$	[49]

Table 5 – Phase Transitions on $Y_d(n,p)$.

3.3 Topological Phase Transitions in Data Driven Systems

Given the previous explanation on phase transitions in simplicial complex and on how to associate a simplicial complex to data, we are now ready to study topological phase transitions in simplicial complexes created from data. In fact, as far as we know, this approach was implemented very recently for random graphs and for functional brain networks [1]. The potential use of such approach is based on the universal character of a phase transitions, which roughly speaking, means that the same "material" would have the same transition temperature, at similar conditions. In other words, one could say that something that boils at 100^oC at sea level is very likely to be water.

Since we have simplicial complexes build from experimental data, we do not expect that the distribution of the simplices would match the theoretical models explained here. However, the behavior of the systems discussed here is somehow analogous to the experimental systems that we are going to discuss in the next chapter. In fact, as was reported in [1], the distribution of the Betti numbers of brain networks resembles the results explained here. In order to transpose this idea to data driven systems, we can hypothesize that some sort of experimental data can be seen as an stochastic realization of an abstract simplicial complex with a given distribution of simplices. Moreover, knowing several realizations of the simplicial complexes associated to data, one could find topological phase transitions. Finally, one could interpret the critical points of the topological phase transitions as a fingerprint of the data, in the same way the melting, and boiling, etc., temperatures of a material at sea level, are signatures of the material.

The random clique complex $X(n,p)$ is simply a way to associate a simplicial complex to the Erdös-Rényi graph $G(n,p)$. $X(n,p)$ is the simplicial complex whose simplices are the cliques of $G(n,p)$. It is the maximal simplicial complex compatible with a given graph.

In ref. [1], $X(n,p)$ was studied in the context of phase transitions in simplicial complexes and It's relations to physics. In short, both the logarithm density of the absolute value of the Euler characteristic, i.e., the Euler entropy, and the distribution of the Betti numbers are topological indicators for phase transition in this system. The first had inspiration in physics, but it intimately related to the second, since

$$\chi = \sum_k (-1)^k \cdot \beta_k.$$

Figure 3 shows the Euler characteristic and the Betti numbers of the random clique complex as function of the probability p obtained in ref. [1]. It's easy to see that the distribution of the Betti number is concentrated in a narrow probability interval, corroborating qualitatively with the rigorous results on phase transitions in simplicial complexes [23, chapter 22]. It's important to notice that even though the Euler characteristic [50] can be computed analytically, we cannot do the same for the Betti numbers. So the homology was computed numerically and its average is plotted analytically for 500 realizations in a random clique complex. Observe that the picks of the Euler characteristics points the exact threshold for the shift in the dominance of the Betti numbers.

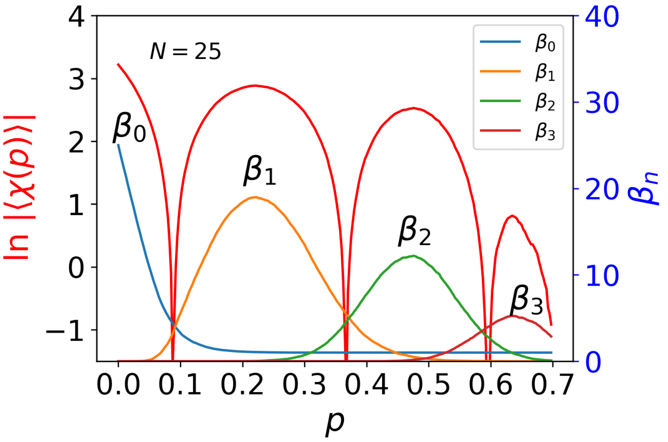

Figure 3 – Topological transitions of the random clique complex. Numerical evaluation obtained from [1].

Once we know how to address numerically phase transitions in $X(n,p)$ we now proceed to one exemplar of this approach to real data, particularly phase transitions in functional brain networks. We illustrate a filtration in a brain-network, obtained in [1]. In this case, the filtration parameter was the absolute value of the Pearson correlation coefficient, in descending order, i.e., the threshold. In figure 4 it was observed that, in similar conditions, the Euler characteristic of those brain networks (as a function of the correlation ε between brain regions) present similar behavior. Each individual, considered here as an stochastic realization of random simplicial complex, is illustrated in gray, while the average is illustrated in blue. Several transitions were detected on the brain networks, all of them in a very narrow interval. This discovery raises the hypothesis that the geometry and topology of functional brain networks are intrinsic properties that can be seen as biomarkers that could help on the detection of unhealthy brains. In the last chapter of this work, we will move forward in this field, trying to check this hypothesis to a different, and also relevant, type of data, namely protein-protein interaction networks (PPIN).

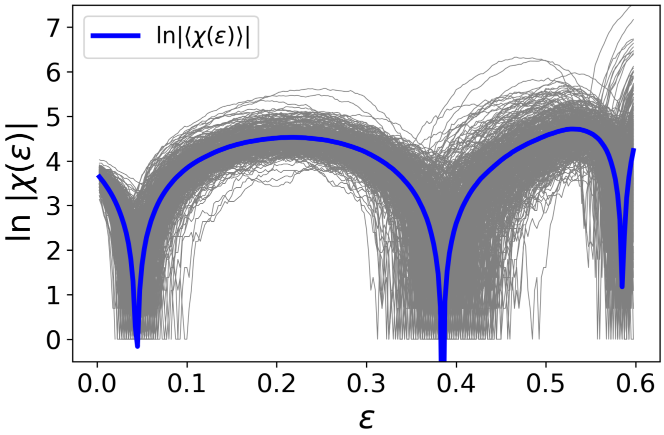

Figure 4 – Topological phase transitions on functional brain networks [1]. Observe that the brain network of each individual (gray) presents transitions inside a narrow interval, around the average (blue).

4 TOPOLOGICAL PHASE TRANSITIONS ON PPIN

In this chapter we will use the approach presented in the previous chapters of this work, to understand the topological properties of protein interaction networks. In particular, we will digress on standard models of protein interaction networks and will detect topological phase transitions observed in such models. Finally, we will see that those transitions also happens on real data, as we are going to illustrate for gene co-expression network from *Saccharomyces cerevisiae* and *Caenorhabditis elegans*.

4.1 Introduction

Nowadays, due to the increase in the availability of technology and new experimental methods that comes with it, we have massive data production that needs tons of processing power and new methods to analyze it. Therefore, it is necessary to develop techniques to obtain information from it [6]. In that sense, topology is a natural tool to be applied.

Topology is the branch of mathematics that concerns in finding properties of objects that are independent of a coordinate system. Topological properties of a system are usually global properties that are independent of perspective or frame of reference. One important problem in topology is to identify and characterize phase transitions in a given system. Association between phase transitions and topological changes in the configuration space have been reported in some model systems in theoretical physics using methods of Morse theory [17, 18, 20].

One of the topological invariants associated with phase transitions is the Euler characteristics. In fact, the Euler Characteristic (EC) is an essential topological invariant [51] and it has been observed that logarithm density of the EC presents singular behavior around the thermodynamic phase transition in theoretical models [17, 33].

For complex networks, EC is computable associating a simplicial complex to it, i.e., a set of simplices (node, edges, triangles, tetrahedrons, etc.), whose intersections between each other are also simplices. If we denote N_k as the total number of k-dimensional simplices, the EC is given by

$$\chi = \sum_k (-1)^k \cdot N_k.$$

The Euler entropy, recently defined in the literature by [20], is the logarithmic density of the EC. Many works associated, through the Euler entropy, phase transitions on a range of continuous systems in physics [38–40].

EC is also related to the percolation transition [52]. Recent works on complex networks verified that the first transition of the Euler entropy is related to the appearance of a giant component, i.e., components that involve a significant share of the elements of the system.

Lately, in [1] the EC was used to detect topological phase transition on functional brain networks and observed that those transitions could be seen as topological bio-markers for brain networks.

In this chapter we will use the same techniques proposed at [1] for Protein-Protein Interaction Networks (PPIN), i.e., networks in which the nodes are proteins the edges are graphical representations of the interactions between them. Topological properties of PPIN have been subject of studies on many relevant works [25–27]. In short, we found topological transitions analogous to the ones obtained at [1] on two models of growth of PPIN: the totally asymmetric and the heterodimerization duplication-divergence models.

To check if those topological transitions are consistent with real data of PPIN, we applied the same approach to data of Gene Co-expression Networks (GCN) of yeast freely available online from project Yeastnet v3 [3]. A GCN is a PPIN in which the edges are based on experimental measures of correlation on the activity of genes over different conditions. Each node on the network corresponds to a gene. An edge connects a pair of genes if they present similar expression patterns over all the experimental conditions [53]. GCN are often used to identify groups of genes that, over many experimental conditions, displays a similar expression profile. Based on the "guilt-by-association" principle [54], it is possible to hypothesize that those genes share a common functionality. Therefore, to understand and compare topological features between co-expression networks may provide useful information about the strengths (and weaknesses) of the association model used to infer the co-expression network [55]. We were able to detect a topological phase transition on the data of GCN from Yeastnet and observed that this transition happen in a narrow interval for several Yeast datasets, suggesting that these transitions are some sort of biomarker, i.e., intrinsic properties of those Yeast networks. Moreover, for Yeast networks with DNA damage [4], we observed that the phase transition happens for a threshold much lower than the ones detected for other yeast networks, which can be possibly seen as a sign of network reorganization due to DNA damage.

4.2 From Protein Interaction Networks to Duplication-Divergence

Proteins are macromolecules that participate in a vast range of functions in a cell, like replication of DNA, response to stimuli or even the transport of molecules, among others. The interaction between proteins represents a central role in almost every cellular process. The proper understanding of how proteins interact within a network can be a crucial advance on the identification of cells' physiology between normal or disease state [56]. The interactions between proteins in a cell can be mathematically represented as a graph, a mathematical structure composed by nodes (or vertices) and edges. Proteins are the nodes and links are given to pairs of protein that interact. This representation is often referred to as a Protein-Protein Interaction Network (**PPIN**).

In [25] it was first observed that PPIN are scale-free networks, i.e., a few nodes (proteins), the so-called Hubs, are the most important nodes, participating in the majority of the cellular functions, i.e., with high degree. In fact, the scale-free property can be seen in a range of other complexes systems, like world wide web, social networks, etc. [7]. The emergence of such systems is characterized by a preferential attachment principle. New nodes added to the system are more propense to attach to nodes with higher degree resulting in a power law degree distribution, i.e., the fraction $p(k)$ of nodes with degree k in the network is inversely proportional to some power of k. In PPIN, this scale-free topology is reported to be a consequence of gene duplication [57].

Gene duplication is the process that generates new genetic material during molecular evolution. The appearance of vertebrates and mammals from unicellular organisms is one of the exemplars of phenomena that would never be possible without gene duplication [58].

We can assume that duplicated genes produce indistinguishable proteins, that is, proteins that interact with the same proteins. Every time a gene duplicates, the proteins that are linked with the product of this gene has one extra link on the network. Thus, proteins with more links are more likely to have a neighbor to be duplicated. In this section, we will discuss models for gene duplication. Gene duplication is one of the most essential factors in evolution. The discovery of a scale-free property in PPIN gave rise to many models for generating PPIN based on the gene duplication principle [5, 59–62], most of them relying on divergence, a process in which genes generated by the same ancestral through duplication accumulate independent changes on their genetic profile over time.

The Duplication-Divergence (DD) model firstly proposed in [59] simulates the growth of a PPIN by gene duplication and divergence. There are some variants of this model, which include heterodimerization, arbitrary divergence [5], random mutations [60], among others.

In this work, we will analyze the simplicial complex associated with two of them. The first one is the totally asymmetric model of duplication and divergence network growth. It involves a single parameter p, called *retention* probability, proposed by [61]. The second one was proposed by [5] and is quite similar to the first one but with a second parameter q that will control the heterodimerization process (that will be defined below). We will also present data from real gene co-expression networks whose behavior corresponds to what is observed on those models.

4.3 Phase transition on Totally asymmetric Duplication-Divergence model

The totally asymmetric model is a duplication-divergence model in which the divergence process is assumed to happen only on the replica of the duplicated node. This model, proposed by [61], is based on the hypothesis that, after duplication, there is a slight chance that the replicate node starts to develop different functions than the original one. In the model, this change is indicated by the deletion of some edges that goes from the replica and is supposed to happen a single time during the growth of the network.

The model is defined as follows: Given a number N and a probabilistic parameter p ($0 \leqslant p \leqslant 1$), the model generates a graph with N nodes from a single edge following the presented algorithm:

- Duplication: One node is randomly selected to duplicate with its edges.
- Divergence: Each edge that goes from the replica activates with a retention probability p. The non-activated edges are removed.

Figure 5 bellow illustrates the process.

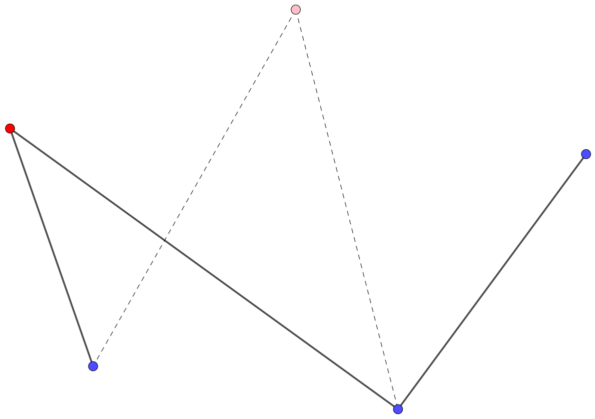

Figure 5 – Duplication step of the totally asymmetric model. For each duplication step a node is selected to be duplicated (red) within its edges. Each edge (dashed lines) that goes from the replica (pink) is activated with independent probability p.

The duplication step simulates genetic replication in a cell, while the divergence step simulates the possibility of a mutation after duplication, which can generate new proteins performing different functions than the original. Some authors [61] consider that the algorithm should treat as a failure any duplication that produces an isolated node and remove that node, but for experimental consistency, simplicity, and to make the computations faster, we will keep the isolated nodes. In fact, this approach can be seen as a step forward in the understanding of gene co-expression data modeling. As we will verify further, experimental data shows isolated nodes for lower levels of co-expression ($\ll 1$). Therefore, keeping the apparent duplication "failures" will make the modeling more appropriate to match with the experiments.

For the computation of the Euler entropy, observe that this algorithm will only produce bipartite networks, which implies that the networks produced by this model will not have cliques with size 3 or greater. The EC, therefore, will be given by the formula

$$\chi = N - E.$$

In here, N represents the total number of nodes, and E the total number of edges. Consequently, we are able to obtain the expected mean value of the EC in terms of the parameter p for a graph with N nodes, by

$$\langle \chi \rangle = N - \langle E \rangle. \tag{4.1}$$

The expected value for the number of edges as function of p for a totally asymmetric model, is calculated using recurrence method, which after some algebra is given by:

$$\langle E \rangle(p) = \prod_{k=1}^{N-2} \left(\frac{2}{k+1} p + 1 \right)$$

Since the Euler entropy is given by

$$S_\chi = \log |\langle \chi \rangle|$$

we have a singularity on S_χ at the values of p where $\langle \chi \rangle = 0$, i.e., where

$$\langle E \rangle(p) = \prod_{k=1}^{N-2} \left(\frac{2}{k+1} p + 1 \right) = N.$$

A graph that has more edges than nodes certainly has cycles [63]. So, the transition point marks the exact retention probability where it is highly probable that the network has cycles, i.e., in the vicinity of the emergence of a Giant component in the Network [22].

We define, therefore, phase transitions as the singularities of the Euler entropy. This definition finds a basis on the obtained on a range of continuous system [17], and was introduced to data in [1]. In figure 6, we observe the behavior of the S_χ as a function of the parameter p of the totally asymmetric model, for a network with 1000 nodes. One can observe that there is a single singularity when p reaches the critical value of ≈ 0.56 (dashed black line).

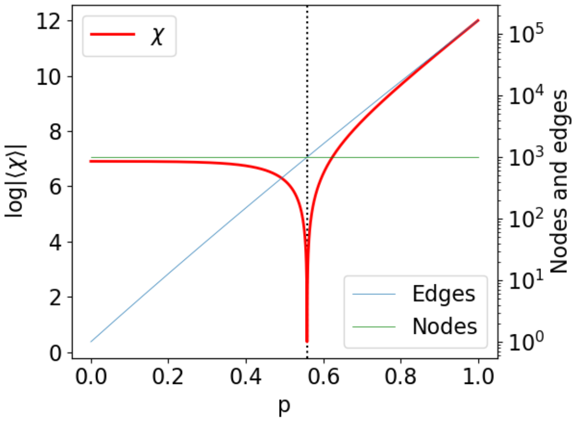

Figure 6 – Euler entropy $S_\chi = \log|\langle\chi\rangle|$ as function of the retention probability p of the Duplication Divergence graph with 1000 nodes. S_χ was determined by (4.1). The singularities $S_\chi \to -\infty$ locate the topological phase transitions on the duplication divergence graphs. We put the number of nodes and edges in a logarithmic scale to let visible the change of dominance between the number of nodes and the number of edges that occurs at the same spot as the phase transition.

To understand the implications and characterize this phase transition, we will analyze the model through other topological parameters, known as Betti numbers. In general, Betti numbers, which are usually indicated by β_n, and are defined as follows. β_0 is the total number of connected components. β_1 is the number of cycles on the graph. For $n > 1$ it starts to get tricky to understand the structure, but, roughly speaking, β_n indicates the number of n-dimensional "holes" on the network.

Figure 7 presents the behaviour of the Betti numbers β_0 and β_1 as functions of p compared with the Euler characteristics. Since we were not able to achieve an analytic expression for the Betti numbers, they were obtained numerically by generating 1000 simulations for each value of p (from 0 to 1 at steps of 10^{-2}), 1000 nodes each.

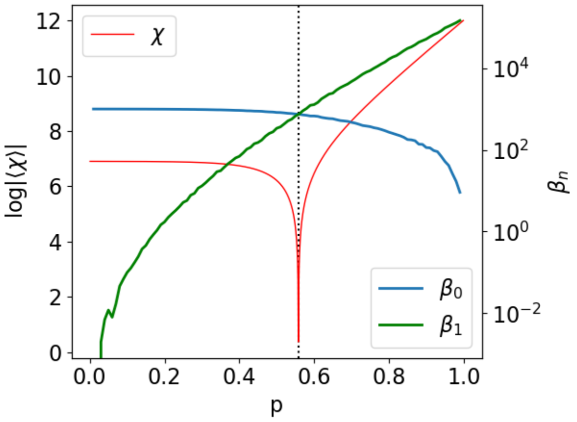

Figure 7 – Betti numbers β_0 and β_1 of DD graphs as functions of the retention probability p (N = 1000). At the first phase the graph is totally disconnected, with many components. As p increases β_1 (i.e, the number of cycles) surpasses β_0 exactly when p crosses the transition point, indicating an important change on the topology of the DD graph.

Notice that the topological phase transition occurs in the vicinity of the value of p_c where the Betti curves β_0 and β_1 overlap and, consequently, $\chi = 0$. For $p < p_c$ the network is divided into many components because there is a high chance that duplicated nodes does not connect with any neighbor of the original nodes. As p increases, the number of components decreases and we have the abundance of cycles in the network. Near p_c, the expected number of cycles is as big as the expected number of components, so there is a high chance that the graph has a cluster with $\mathcal{O}(N)$ nodes.

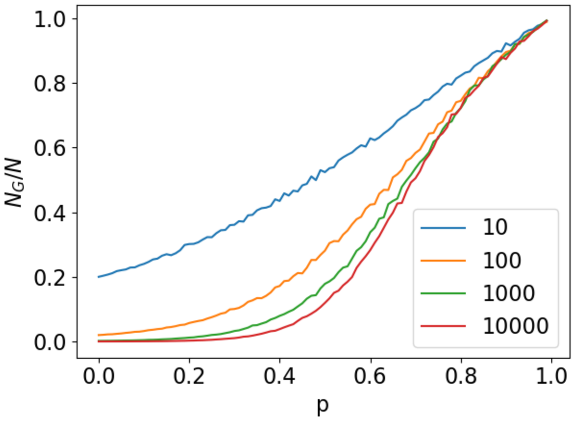

Figure 8 – Percolation transition on DD model. The three curves shows the fraction of nodes on the biggest component N_G/N as function of the retention probability p.

Another explanation for the obtained phase transition from the totally asymmetric model comes from percolation theory. It is a known fact that the first change of sign of the EC on complex networks is related with the appearance of a giant component, i.e., of a connected component on the graph that involves a significant part of the nodes on the system. Then, for the totally asymmetric model, we observe that the transition obtained is the percolation transition of the system, i.e., the value on the probabilistic parameter that maintains most of the proteins interacting as a whole.

Figure 8 shows the expected fraction of nodes that belong to the largest component as a function of the retention probability p for different numbers of nodes. Observe that as the number of nodes increases, the point where a significant fraction of nodes are in the giant component gets closer and closer to the transition in the thermodynamics limit ($N \gg 1$).

4.4 Phase Transitions on GCN

In a biological context, interactions among proteins can be physical interactions, indicated by the physical contact between them as a result of biochemical events guided by electrostatic forces [64], but also functional interactions. It is possible for a group of proteins to perform a common biological function without actually being in direct contact, regulating some process or even making common use of some other molecule [28].

Driving our focus to functional interactions, there are many methods to identify such interactions. A common method consists of analyzing gene co-expression patterns. This method is based on the principle that genes with correlated activity produce interacting protein [65]. Mapping the correlation on the activity of genes, we build the so-called co-expression networks. Co-expression networks are PPIN where the links between the proteins measure the similarity on the expression pattern of its producing gene. These measures are obtained analyzing the correlation of the activity of genes through different experiments [28].

We now contrast the theory and numerical simulations presented in the previous section with gene co-expression networks obtained experimentally from data for yeast networks in which we can observe topological phase transitions with a similar profile. Therefore, it makes sense to test if we will also find phase transitions on real data of GCN, and if such transitions can give us useful information about the network. In this section, we analyzed 48 GCN data of *Saccharomyces cerevisiae*, also known as baker's yeast, available online from the project YeastNet v3 [3]. The whole data-set covers around 97% of the yeast coding genome (5730 genes). Each data corresponds to a network with $\approx 800 - 3000$ nodes and between 7000 to 64000 links. The data is a collection generated through diverse experimental approaches.

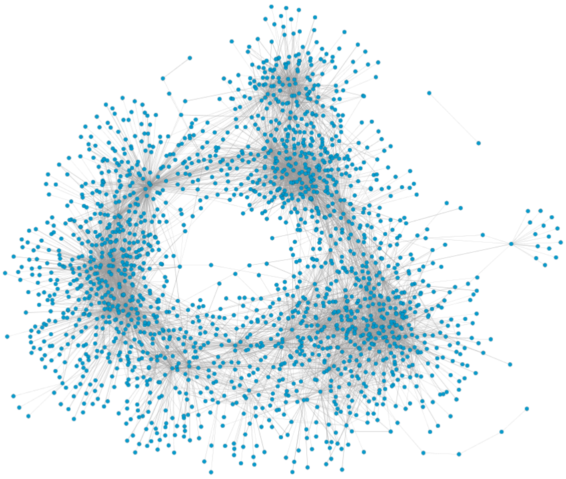

Figure 9 – Network representation of gene co-expression network for the yeast. This image was build using the software Cytoscape [2]. The presented network have 1893 nodes and 7955 links and it was collected from Yeastnet v3 [3].

Each of the data-sets analyzed consists of an adjacency matrix for the Yeast network. Each node on this network corresponds to a gene. Two genes are connected by an edge if there is significant co-expression relationship between the genes. We associate to each edge a weight that is the normalized correlation between the expression of the nodes involved.

The GCN then passed to a numerical process of filtration. This process consists of, for a given $\varepsilon \in [0,1]$, to remove all the edges with correlation at most $1-\varepsilon$. For $\varepsilon = 0$, for example, we would have an empty network, with all the nodes disconnected. As ε increases, we include the link with a correlation greater than $1-\varepsilon$, until ε reaches 1, when the graph becomes fully connected, with all its edges. In this work, we used the filtration process to follow the topology of the GCN as a function of the correlation between the genes.

In contrast with the results of the previous case, where the EC was given as function of the probabilistic parameter p of the totally asymmetric model, the EC is given as function of the threshold ε by

$$\chi(\varepsilon) = \sum_{k=1}^{N} (-1)^{k+1} Cl_k(\varepsilon),$$

where $Cl_k(\varepsilon)$ is the total number of k-cliques that are on the network for a given filtration level ε.

We stress that the computation of topological invariants is an NP-complete problem, and therefore, here, we only illustrate data-sets where the computation was feasible, namely, to 20 Yeast networks. Figure 10 shows the Euler entropy as a function of ε of 20 of the networks on Yeastnet's database. The time scale for the numerical computation of the numbers of k-cliques increases exponentially with the size of the network. As ε increases, the network becomes denser, and the computations become extensive. Because of that, for some of the datasets, we could not compute the Euler characteristics to a threshold where one could detect a topological phase transition. The Euler entropy averaged over the data sets (blue line), clearly shows the presence of a singularity.

Observe that most of the singularities happen near $\varepsilon \approx 0.876$ except for one of the networks, where the singularity happened at $\varepsilon \approx 0.573$. This network comes from an experiment whose goal was to evaluate response to DNA damage. It is a known fact that eukaryotic cells respond to DNA damage by rearranging its cycle and modulating gene expression to ensure efficient DNA repair [4]. Therefore, our analysis suggests that the Euler entropy is sensible to this reorganization process in the damaged Yeast network and could also be seen as an intrinsic bio-marker with significant potential to be a comparative measure for GCN under DNA damage. Thus, an analysis is desirable to verify the relation between DNA damage and percolation transitions on a solid basis.

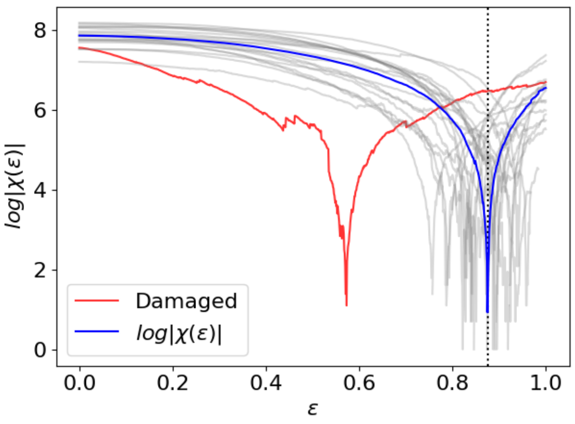

Figure 10 – Euler entropy S_χ as a function of the correlation threshold ε of 20 co-expression networks from Yeastnet database. Each gray line corresponds to a co-expression network, while the blue line is the average of the gray ones. Most of the topological phase transition was detected on those networks for $\varepsilon \approx 0.876$, except for the Yeast network with the DNA damaged, where the transition happened at $\varepsilon \approx 0.573$ (red line). This data was designed intended to measure the response to DNA damage in the network [4], and the shift in the threshold of the topological transition is apparently sensitive to that response.

In order to provide a characterization of the topological phase transition of GCN, we also calculated the Betti curves β_0 and β_1 for the same Yeast networks used before. For $n \geq 1$, β_n did not presented significant behaviour to be considered. In Figure 11, we illustrate the Betti curves for the data Yeast network dataset.

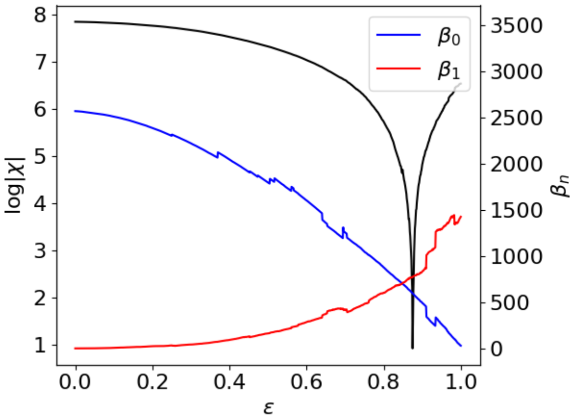

Figure 11 – Average of the Betti numbers as a function of the correlation threshold ε. Observe that, as happened with the totally asymmetric model, the dominance of β_0 and β_1 shifts at the vicinity of the singularity of the Euler entropy (black line).

In the whole data sets it we can observe agreement between the threshold transition of S_χ and the threshold where β_1 becomes greater than β_0, in analogy with the topological transitions reported on theoretical models for random simplicial complexes [23, 24]. For values of the threshold below the transition, since β_0 is greater than β_1, the network is fragmented in many components. As ε gets closer to the transition threshold, more edges are added to the network, lowering the number of components, and changing the network to a denser one and with more cycles. Once the number of components gets smaller, almost every new edge creates more loops, and the topological transition happens precisely when the number of loops is greater than the number of components.

Those observations about the behavior of the Betti numbers are compatible with the results observed in [66], and provides some possible interpretation for this shift in the critical points of the topological phase transition. In [66] it was observed a linear correlation of $R = -0.55$ between the chance of survival of cancer patients and the number of cycles in the network, also known as the complexity of the cancer PPIN. In fact, the complexity was measured using persistence homology, specifically by the magnitude of the Betti number β_1, that counts the number of one-dimensional cycles on the network. The higher the β_1, the higher is the complexity, that leads to lower chances of survival. This study gives evidence that, for cancerous cells, the complexity of the PPIN is associated with a health state of the cell.

Now, returning to our analysis of gene co-expression networks of Yeast, we observed that, for one of the networks, the transition indicated by the Euler characteristics happened at a distinct threshold (≈ 0.573). The analysis of the Betti numbers indicates that this transition is characterized by a shift of dominance between β_0 and β_1, as theoretically described for random simplicial complexes [24]. Thus, during the filtration process, the network corresponding to Yeast with "DNA damage" reaches the topological phase transition earlier than the other data (on which the transitions happened near $\varepsilon \approx 0.876$), indicating a rapid appearance of cycles on the network, i.e., the increase of network complexity. The arguments proposed in [66], together with the observations above reinforce the hypothesis that topological phase transitions have the potential to be used as intrinsic biomarkers for protein interaction networks more generally.

4.5 Heterodimerization model.

Here, we move forward to a model where the protein interactions display more complex dynamics. This model is a variation of the totally asymmetric model discussed in the previous sections. Given a number N and two probabilistic parameters p and q, the model generates a graph with N nodes from a single edge following the steps below:

- Duplication: One node is randomly selected to duplicate with its edges.
- Divergence: Each edge that goes from the replica activates with probability p.
- Heterodimerization: The replica and the original nodes connect with probability q.

The heterodimerization step mimics the probability that the original node is a dimer, i.e., two molecules joined by bonds that can be either strong or weak. This step is important for clustering and is observed in real PPIN [67]. It is also known that this model produces cliques with similar size and quantity, of those observed in some real PPIN [5], contrasting with the totally asymmetric model where we could only observe low clique size.

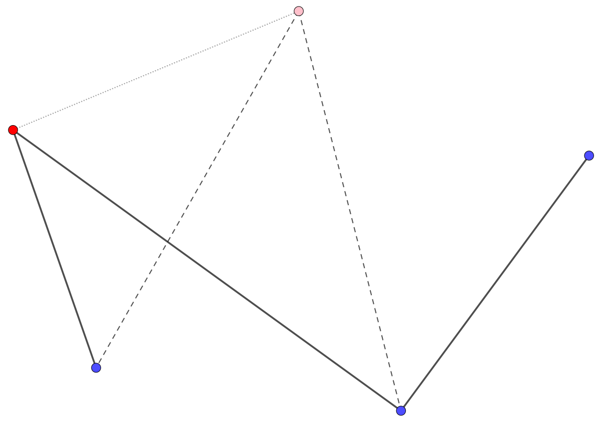

Figure 12 – Duplication step of the heterodimerization model. For each duplication step a node is selected to be duplicated (red) within its edges. Each edge (dashed lines) that goes from the replica (pink) is activated with independent probability p. Also, an edge going from the replica to the original node (dotted line) is created with probability q.

As previously defined, the phase transitions are the singularities of the Euler entropy. Here, given the complexity for generating these networks, we implemented it numerically and compared with experimental data. In figure 13, we observe the average of the Euler entropy S_χ as a function of the retention probability p for a graph with 1000 nodes and different values of q. Figure 13 shows the presence of up to three topological phase transitions and the positions of those transitions depends on the value of q. More transitions happen mostly because the heterodimerization step turns possible the appearance of cliques of a higher order. The higher the value o q, more probable it becomes for the appearance o cliques of bigger sizes. In some real PPIN, it was reported the presence in abundance of large cliques [68].

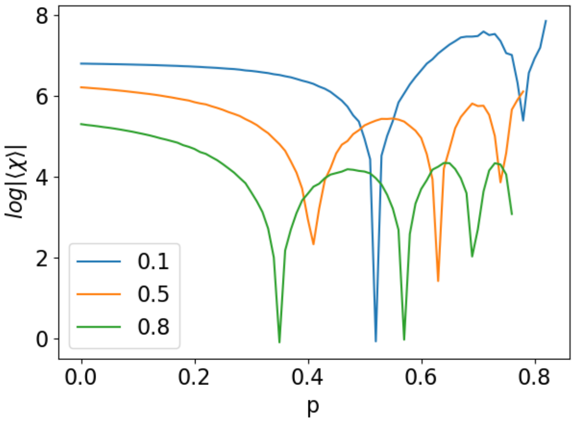

Figure 13 – Average of Euler Entropy as function of the retention probability p for different values o q. Observe that, differently than the previous model, there are more transitions and their number varies with the value of q. This happens because the heterodimerization step makes possible the appearance of cliques of different sizes as observed in real PPIN [5]

.

Because of the higher number of cliques, we were not able to achieve an analytic expression for the Euler entropy. Also, since the network becomes denser due to the heterodimerization step, we could not compute the Betti numbers for this model to make similar conclusions about the transitions, as we did for the totally asymmetric model. Nevertheless, to reinforce the significance of our analysis, we present experimental data for PPIN that presents similar phase transitions profile.

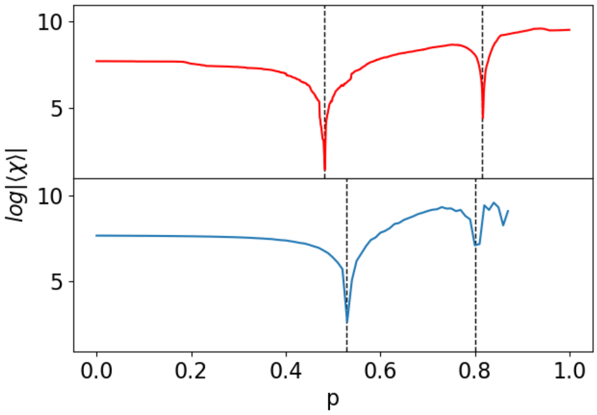

Figure 14 – (top) Euler entropy of PPIN from *C. elegans* in contrast with simulation with heterodimerization model (bottom). Both have two singularities

.

In figure 14 (top) we illustrated the Euler entropy of a PPIN for the *C. elegans*. The data was obtained from Wormnet v.3 database for this nematode [30]. For this network, which consists of 2,219 genes and 53,683 links, each link was inferred by analysis of bacterial and archaeal orthologs, i.e., homologous gene sequences of bacteria and archaea related to *C. elegans* by linear descent.

Differently from the data of GCN of yeast presented previously, the Euler entropy of this network presents two singularities at the vicinity of $\varepsilon_1 = 0.48$ and $\varepsilon_2 = 0.82$. This behavior could not be observed if we would consider the growth of a PPIN through duplication and divergence only. However, with the heterodimerization model, we could achieve a similar profile of the Euler entropy by simply setting the correct parameters, as can be seen in the bottom of figure 14. Figure 14 (bottom) shows the expected Euler entropy as a function of the retention probability p, averaged over $1,000$ simulations for each value of p (ranging from 0 to 1 with steps of 10^{-2}). For this simulation, we set the number of nodes as $2,219$ in a way that we can compare with the *C. elegans* data. We also set the value of the heterodimerization probability q as 0.05. This value was only set for comparison proposes, and probably is not the best value to fit the data. This choice of q was 0.05 because for higher values of q we have more chance of appearance of larger cliques, which would lead to more transitions.

This analysis suggests the presence of "dimmers" during the growth of the *C. elegans* network. Therefore, the topological phase transitions displayed by the heterodimerization model should fit better for this data. Moreover, other models for PPIN could give us other insights for those transitions.

4.6 Conclusions

In this work, we used the recent concept of topological phase transitions, in order to detect phase transitions in Yeast and co-expression networks. We verified that such transitions correspond to the emergence of a giant component in the network, as observed on networks generated by the totally asymmetric duplication-divergence model. Given that several Yeast datasets had a Topological Phase transitions at a very narrow threshold interval between 0.756 and 0.958, our results give strong support to the hypothesis that percolation transitions are strong topological bio-marker in a network.

Besides that, we propose to analyze PPIN through the Euler characteristics, which is a more straightforward parameter that relates to the Betti numbers. By analyzing a network for Yeast under DNA damage, we found that the transition point shifted to a value of 0.573, quite far from the interval $[0.756, 0.958]$, where the topological phase transitions of the other Yeast network datasets takes place. Therefore, through this work, we have evidence that zeros of the Euler characteristic, or the singularities of its Euler entropy, can be seen as suitable topological invariant to distinguish macroscopic properties of the Yeast networks.

For the *C. elegans* PPIN, on the other hand, the same approach gave rise to two topological phase transitions, indicating that the totally asymmetric model does is not sufficient to capture the topological aspects of the *C. elegans* PPIN data properly. It suggests that there should be some other process that leads to the growth of a PPIN. One possibility could be the presence of dimmers, i.e., two nodes that are parts of the same protein (or gene) that are connected. To test this hypothesis, we computed the expected Euler entropy of the heterodimerization duplication-divergence for the model and observed that, for a suitable set of parameters, we could obtain a similar profile of the Euler entropy as the one observed on *C. elegans* network. It is important to emphasize that, in order to match our theory and numerical simulations to the experimental data, we considered that the nodes that got isolated after duplication were kept in the network. Further studies are desired to a proper biological interpretation of this assumption.

It is important to discuss that other models propose different processes for the growth of a PPIN, which deserves investigation under our approach. In ref. [5] for example, the authors presents a model with arbitrary divergence in which they could replicate with high confidence the same number of cliques observed at PPIN data. Many other models simulate different aspects of the growth of a PPIN [5, 60, 62] and the analysis of such models through phase transitions of the Euler entropy would give us precious information about PPIN. In short, this work contributes to a better understanding of the topological properties of PPIN and gives us the perspective to use topological phase transitions as a methodology for classifying protein interaction networks more generally.

REFERENCES

1 SANTOS, F. A. N.; RAPOSO, E. P.; COUTINHO-FILHO, M. D.; COPELLI, M.; STAM, C. J.; DOUW, L. Topological phase transitions in functional brain networks. Disponível em: <http://dx.doi.org/10.1101/469478>.

2 SHANNON, P.; MARKIEL, A.; OZIER, O.; BALIGA, N.; WANG, J.; RAMAGE, D.; AMIN, N.; SCHWIKOWSKI, B.; IDEKER, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, v. 13, n. 11, p. 2498–2504, 2003.

3 KIM, H.; SHIN, J.; KIM, E.; KIM, H.; HWANG, S.; SHIM, J. E.; LEE, I. YeastNet v3: A public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. *Nucleic Acids Research*, 2014. ISSN 03051048.

4 GASCH, A. P.; HUANG, M.; METZNER, S.; BOTSTEIN, D.; ELLEDGE, S. J.; BROWN, P. O. Genomic Expression Responses to DNA-damaging Agents and the Regulatory Role of the Yeast ATR Homolog Mec1p. *Molecular Biology of the Cell*, v. 12, n. 10, p. 2987–3003, 10 2001. ISSN 1059-1524. Disponível em: <http://www.molbiolcell.org/doi/10.1091/mbc.12.10.2987>.

5 ISPOLATOV, I.; KRAPIVSKY, P. L.; MAZO, I.; YURYEV, A. Cliques and duplication–divergence network growth. *New Journal of Physics*, IOP Publishing, v. 7, n. 1, p. 145–145, 6 2005. ISSN 1367-2630. Disponível em: <http://stacks.iop.org/1367-2630/7/i=1/a=145?key=crossref.9173be6c5c84f98aff1f2bcd76806bcc>.

6 CARLSSON, G. Topology and data. *Bulletin of the American Mathematical Society*, v. 46, n. 2, p. 255–308, 1 2009. ISSN 0273-0979. Disponível em: <http://www.ams.org/journal-getitem?pii=S0273-0979-09-01249-X>.

7 BARABASI, A.-L.; PÓSFAI, M. *Network Science by Albert-László Barabási.* 1. ed. Cambridge: Cambridge University Press, 2016. 475 p. Disponível em: <http://networksciencebook.com/>.

8 EDELSBRUNNER, H.; HARER, J. J. *Computational topology : an introduction.* [S.l.]: American Mathematical Society, 2010. 241 p. ISBN 9780821849255.

9 LEE, Y.; BARTHEL, S. D.; DŁOTKO, P.; MOOSAVI, S. M.; HESS, K.; SMIT, B. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, Nature Publishing Group, v. 8, p. 15396, 5 2017. ISSN 2041-1723. Disponível em: <http://www.nature.com/doifinder/10.1038/ncomms15396>.

10 NICOLAU, M.; LEVINE, A. J.; CARLSSON, G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, v. 108, n. 17, p. 7265–7270, 4 2011. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21482760http://www.ncbi.nlm.nih.gov/pubmed/21482760http://www.ncbi.nlm.nih.gov/pubmed/21482760>.

11 PETRI, G.; SCOLAMIERO, M.; DONATO, I.; VACCARINO, F. Topological Strata of Weighted Complex Networks. *PLoS ONE*, v. 8, n. 6, p. 66506, 2013. Disponível em: <www.plosone.org>.

12 ATIYAH, M. F. Topological quantum field theory. *Publications Mathématiques de l'IHÉS*, v. 68, p. 175–186, 1988. Disponível em: <http://www.numdam.org/item/PMIHES_1988___68___175_0/>.

13 HALDANE, F. D. M. Nobel Lecture: Topological quantum matter. *Reviews of Modern Physics*, American Physical Society, v. 89, n. 4, p. 040502, 10 2017. ISSN 0034-6861. Disponível em: <https://link.aps.org/doi/10.1103/RevModPhys.89.040502>.

14 KOSTERLITZ, J. M. Nobel Lecture: Topological defects and phase transitions. *Reviews of Modern Physics*, American Physical Society, v. 89, n. 4, p. 040501, 10 2017. ISSN 0034-6861. Disponível em: <https://link.aps.org/doi/10.1103/RevModPhys.89.040501>.

15 STANLEY, H. E. H. E. *Introduction to phase transitions and critical phenomena.* [S.l.]: Oxford University Press, 1987. 308 p. ISBN 0195053168.

16 DOMB, C. *The critical point : a historical introduction to the modern theory of critical phenomena.* [S.l.]: Taylor & Francis, 1996. 376 p. ISBN 074840435X.

17 PETTINI, M. *Geometry and topology in Hamiltonian dynamics and statistical mechanics.* New York: [s.n.], 2007. ISBN 978-0-387-30892-0.

18 KASTNER, M. Phase transitions and configuration space topology. *Reviews of Modern Physics*, v. 80, p. 167–187, 2008. Disponível em: <https://journals.aps.org/rmp/pdf/10.1103/RevModPhys.80.167>.

19 BUCHANAN, M. It's just a phase... *Nature Physics*, v. 4, n. 1, p. 5–5, 1 2008. ISSN 1745-2473. Disponível em: <http://www.nature.com/articles/nphys819>.

20 SANTOS, F. A. N.; SILVA, L. C. B. da; COUTINHO-FILHO, M. D. Topological approach to microcanonical thermodynamics and phase transition of interacting classical spins. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2017, n. 1, p. 013202, 1 2017. ISSN 1742-5468. Disponível em: <http://stacks.iop.org/1742-5468/2017/i=1/a=013202?key=crossref.6284bdda4a6c2e768f85850c94d27803>.

21 PETTINI, G.; GORI, M.; FRANZOSI, R.; CLEMENTI, C.; PETTINI, M. On the origin of Phase Transitions in the absence of Symmetry-Breaking. 2018. Disponível em: <https://arxiv.org/pdf/1706.07950.pdf>.

22 ERDÖS, P.; RÉNYI, A. On Random Graphs I. *Publicationes Mathematicae (Debrecen)*, v. 6, p. 290–297, 1959. Disponível em: <http://snap.stanford.edu/class/cs224w-readings/erdos59random.pdf>.

23 TOTH, C. D.; O'ROURKE, J.; GOODMAN, J. E. *Handbook of Discrete and Computational Geometry, Third Edition.* 3. ed. [S.l.]: Chapman and Hall/CRC, 2017. ISBN 9781498711395.

24 LINIAL, N.; PELED, Y. On the phase transition in random simplicial complexes. *Annals of Mathematics*, v. 184, n. 3, p. 745–773, 11 2016. ISSN 0003-486X. Disponível em: <http://annals.math.princeton.edu/2016/184-3/p03>.

25 JEONG, H.; MASON, S. P.; BARABÁSI, A.-L.; OLTVAI, Z. N. Lethality and centrality in protein networks. *Nature*, Nature Publishing Group, v. 411, n. 6833, p. 41–42, 5 2001. ISSN 0028-0836. Disponível em: <http://www.nature.com/articles/35075138>.

26 MASLOV, S.; SNEPPEN, K. Specificity and stability in topology of protein networks. *Science (New York, N.Y.)*, American Association for the Advancement of Science, v. 296, n. 5569, p. 910–3, 5 2002. ISSN 1095-9203. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/11988575>.

27 YOOK, S.-H.; OLTVAI, Z. N.; BARABÁSI, A.-L. Functional and topological characterization of protein interaction networks. *PROTEOMICS*, John Wiley & Sons, Ltd, v. 4, n. 4, p. 928–942, 4 2004. ISSN 1615-9853. Disponível em: <http://doi.wiley.com/10.1002/pmic.200300636>.

28 VELLA, D.; ZOPPIS, I.; MAURI, G.; MAURI, P.; SILVESTRE, D. D. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *EURASIP Journal on Bioinformatics and Systems Biology*, v. 2017, p. 6, 2017. Disponível em: <https://bsb-eurasipjournals.springeropen.com/track/pdf/10.1186/s13637-017-0059-z>.

29 BOTSTEIN, D.; FINK, G. R. Yeast: an experimental organism for 21st Century biology. *Genetics*, Genetics Society of America, v. 189, n. 3, p. 695–704, 11 2011. ISSN 1943-2631. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/22084421http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3213361>.

30 CHO, A.; SHIN, J.; HWANG, S.; KIM, C.; SHIM, H.; KIM, H.; KIM, H.; LEE, I. WormNet v3: a network-assisted hypothesis-generating server for Caenorhabditis elegans. *Nucleic Acids Research*, Oxford University Press, v. 42, n. W1, p. W76–W82, 7 2014. ISSN 1362-4962. Disponível em: <http://academic.oup.com/nar/article/42/W1/W76/2435686/WormNet-v3-a-networkassisted-hypothesisgenerating>.

31 GIRARD, L. R.; FIEDLER, T. J.; HARRIS, T. W.; CARVALHO, F.; ANTOSHECHKIN, I.; HAN, M.; STERNBERG, P. W.; STEIN, L. D.; CHALFIE, M. WormBook: the online review of Caenorhabditis elegans biology. *Nucleic Acids Research*, Oxford University Press, v. 35, n. Database, p. D472–D475, 1 2007. ISSN 0305-1048. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl894>.

32 MUNKRES, J. R. *Elements of algebraic topology.* [S.l.]: Perseus Books, 1984. 454 p. ISBN 9780201627282.

33 SANTOS, F. A.; COUTINHO-FILHO, M. D. Topology, symmetry, phase transitions, and noncollinear spin structures. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 2009. ISSN 15393755.

34 MARR, B. *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.* 2018. Disponível em: <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#2d0edaf560ba>.

35 BAKER, R. E.; PEÑA, J.-M.; JAYAMOHAN, J.; JÉRUSALEM, A. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology letters*, The Royal Society, v. 14, n. 5, p. 20170660, 5 2018. ISSN 1744-957X. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/29769297>.

36 ZOMORODIAN, A. J. *Topology for computing.* [S.l.]: Cambridge University Press, 2005. 243 p. ISBN 9780511546945.

37 TOMITA, H.; MURAKAMI, C. Percolation Pattern in Continuous Media and Its Topology. *Researsh of Pattern Formation*, p. 197–203, 1994. Disponível em: <http://www.scipress.org/e-library/rpf/pdf/chap4/0197.PDF>.

38 BLANCHARD, P.; FORTUNATO, S.; GANDOLFO, D. Euler–Poincaré characteristic and phase transition in the Potts model on Z2. *Nuclear Physics B*, North-Holland, v. 644, n. 3, p. 495–508, 11 2002. ISSN 0550-3213. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0550321302006818>.

39 AKSIMENTIEV, A.; HOLYST, R. Application of the Euler characteristic to the study of homopolymer blends and copolymer melts. *Polimery*, v. 46, n. 05, p. 307–322, 5 2001. ISSN 00322725. Disponível em: <http://en.www.ichp.pl/Application-of-the-Euler-characteristic>.

40 BLANCHARD, P.; GANDOLFO, D.; RUIZ, J.; SHLOSMAN, S. *On the Euler-Poincaré Characteristic of the Random Cluster Model.* [S.l.]. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.583.4298&rep=rep1&type=pdf>.

41 SPEIDEL, L.; HARRINGTON, H. A.; CHAPMAN, S. J.; PORTER, M. A. Topological data analysis of continuum percolation with disks. 2018. Disponível em: <https://journals.aps.org/pre/pdf/10.1103/PhysRevE.98.012318>.

42 NEHER, R. A.; MECKE, K.; WAGNER, H. Topological estimation of percolation thresholds. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2008, n. 01, p. P01011, 1 2008. ISSN 1742-5468. Disponível em: <http://stacks.iop.org/1742-5468/2008/i=01/a=P01011?key=crossref.49687f21268502da4a5a8e6780e62649>.

43 STAUFFER, D.; AHARONY, A. *Introduction to percolation theory.* [S.l.: s.n.]. 181 p. ISBN 0748402535.

44 SINGH, G.; MÉMOLI, F.; CARLSSON, G. *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.* [S.l.], 2007. Disponível em: <https://research.math.osu.edu/tgda/mapperPBG.pdf>.

45 Edelsbrunner; Letscher; Zomorodian. Topological Persistence and Simplification. *Discrete & Computational Geometry*, Springer-Verlag, v. 28, n. 4, p. 511–533, 11 2002. ISSN 0179-5376. Disponível em: <http://link.springer.com/10.1007/s00454-002-2885-2>.

46 PITTEL, B. A Random Graph With a Subcritical Number of Edges. *Transactions of the American Mathematical Society*, v. 309, p. 51–75, 1988. Disponível em: <https://www.jstor.org/stable/pdf/2001158.pdf>.

47 LUCZAK, T.; PITTEL, B.; WTERMAN, J. C. *THE STRUCTURE OF A RANDOM GRAPH AT THE POINT OF THE PHASE TRANSITION.* [S.l.], 1994. v. 341, n. 2, 721–748 p. Disponível em: <http://www.ams.org/journal-terms-of-use>.

48 HOFFMAN, C.; KAHLE, M.; PAQUETTE, E. The Threshold for Integer Homology in Random d-Complexes. *Discrete & Computational Geometry*, Springer-Verlag New York, Inc., v. 57, n. 4, p. 810–823, 6 2017. ISSN 0179-5376. Disponível em: <http://link.springer.com/10.1007/s00454-017-9863-1>.

49 MESHULAM, R.; WALLACH, N. Homological connectivity of random k-dimensional complexes. *Random Structures and Algorithms*, John Wiley & Sons, Inc., v. 34, n. 3, p. 408–417, 2009. ISSN 10429832. Disponível em: <https://dl.acm.org/citation.cfm?id=1526617>.

50 KNILL, O. On the Dimension and Euler characteristic of random graphs. 12 2011. Disponível em: <http://arxiv.org/abs/1112.5749>.

51 GHRIST, R. W. *Elementary applied topology.* [S.l.: s.n.], 2014. 269 p. ISBN 9781502880857.

52 OKUN, B. L. Euler characteristic in percolation theory. *Journal of Statistical Physics*, Kluwer Academic Publishers-Plenum Publishers, v. 59, n. 1-2, p. 523–527, 4 1990. ISSN 0022-4715. Disponível em: <http://link.springer.com/10.1007/BF01015581>.

53 PETEREIT, J.; HARRIS, F. C.; SCHLAUCH, K. petal: A novel co-expression network modeling system. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* IEEE, 2015. p. 234–241. ISBN 978-1-4673-6799-8. Disponível em: <http://ieeexplore.ieee.org/document/7359686/>.

54 SMET, R. D.; MARCHAL, K. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, Nature Publishing Group, v. 8, n. 10, p. 717–729, 10 2010. ISSN 1740-1526. Disponível em: <http://www.nature.com/articles/nrmicro2419>.

55 XULVI-BRUNET, R.; LI, H. Co-expression networks: graph properties and topological comparisons. v. 26, n. 2, p. 205–214, 2010. Disponível em: <https://academic.oup.com/bioinformatics/article-abstract/26/2/205/209253>.

56 VIDAL, M.; CUSICK, M. E.; BARABÁSI, A.-L. Interactome networks and human disease. *Cell*, NIH Public Access, v. 144, n. 6, p. 986–98, 3 2011. ISSN 1097-4172. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21414488http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3102045>.

57 BARABÁSI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, Nature Publishing Group, v. 5, n. 2, p. 101–113, 2 2004. ISSN 1471-0056. Disponível em: <http://www.nature.com/articles/nrg1272>.

58 OHNO, S. *Evolution by Gene Duplication.* Berlin, Heidelberg: Springer Berlin Heidelberg, 1970. ISBN 978-3-642-86661-6. Disponível em: <http://link.springer.com/10.1007/978-3-642-86659-3>.

59 VÁZQUEZ, A.; FLAMMINI, A.; MARITAN, A.; VESPIGNANI, A. Modeling of Protein Interaction Networks. *Complexus*, Karger Publishers, v. 1, n. 1, p. 38–44, 2003. ISSN 1424-8492. Disponível em: <https://www.karger.com/Article/FullText/67642>.

60 PASTOR-SATORRAS, R.; SMITH, E.; SOLÉ, R. V. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, Academic Press, v. 222, n. 2, p. 199–210, 5 2003. ISSN 0022-5193. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022519303000286>.

61 ISPOLATOV, I.; KRAPIVSKY, P. L.; YURYEV, A. Duplication-divergence model of protein interaction network. *Physical review. E, Statistical, nonlinear, and soft matter physics*, NIH Public Access, v. 71, n. 6 Pt 1, p. 061911, 6 2005. ISSN 1539-3755. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/16089769http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2092385>.

62 FARID, N.; CHRISTENSEN, K. Evolving networks through deletion and duplication. *New Journal of Physics*, IOP Publishing, v. 8, n. 9, p. 212–212, 9 2006. ISSN 1367-2630. Disponível em: <http://stacks.iop.org/1367-2630/8/i=9/a=212?key=crossref.d976187ae5c371558c9abecd75b27f7c>.

63 BOLLOBÁS, B. *Modern graph theory.* [S.l.]: Springer, 1998. 394 p.

64 RIVAS, J. D. L.; FONTANILLO, C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, Public Library of Science, v. 6, n. 6, p. e1000807, 6 2010. ISSN 1553-7358. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20589078http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2891586>.

65 GRIGORIEV, A. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Research*, Oxford University Press, v. 29, n. 17, p. 3513–3519, 9 2001. ISSN 13624962. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/29.17.3513>.

66 BENZEKRY, S.; TUSZYNSKI, J. A.; RIETMAN, E. A.; KLEMENT, G. L. Design principles for cancer therapy guided by changes in complexity of protein-protein interaction networks. *Biology Direct*, BioMed Central, v. 10, n. 1, p. 32, 12 2015. ISSN 1745-6150. Disponível em: <http://www.biologydirect.com/content/10/1/32>.

67 ISPOLATOV, I.; YURYEV, A.; MAZO, I.; MASLOV, S. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Research*, Oxford University Press, v. 33, n. 11, p. 3629–3635, 6 2005. ISSN 0305-1048. Disponível em: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki678>.

68 FELL, D. A.; JEONG, H.; TOMBOR, B.; ALBERT, R.; OLTVAI, Z. N.; BARABASI, A. L. *Protein complexes and functional modules in molecular networks.* [S.l.], 2001. v. 268, 12123–12128 p. Disponível em: <www.pnas.org>.