The Euler Characteristic and Topological Phase Transitions in Complex Systems

Edgar Amorim

Departamento de Matematica, Universidade Federal Rural de Pernambuco, 52171-900, Recife, PE, Brazil

Rodrigo A. Moreira

Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5B, 02-106 Warsaw, Poland

Fernando A N Santos

Department of Anatomy Neurosciences, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Neuroscience, 1081 HZ, Amsterdam, The Netherlands and Departamento de Matemática, Universidade Federal de Pernambuco, 50670-901 Recife, PE, Brazil *

(Dated: December 10, 2019)

In this work, we use methods and concepts of applied algebraic topology to comprehensively explore topological phase transitions in complex systems. Topological phase transitions are characterized by the zeros of the Euler characteristic (EC) or by singularities of the Euler entropy and also indicate signal changes in the mean node curvature of networks. Here, we provide strong evidence that the zeros of the Euler characteristic can be interpreted as a complex network's intrinsic fingerprint. We theoretically and empirically illustrate this across different biological networks: We first target our investigation to protein-protein interaction networks (PPIN). To do so, we used methods of topological data analysis to compute the Euler characteristic analytically, and the Betti numbers numerically as a function of the attachment probability 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 for gene co-expression networks (GCN) of *Saccharomyces cerevisiae*, also known as baker's yeast, as well as of the nematode *Caenorhabditis elegans*. Supporting our theoretical expectations, we are able to detect topological phase transitions in both networks obtained according to different similarity measures. Later, we theoretically illustrate the emergence of topological phase transitions in three classical network models, namely the Watts-Strogratz model, the Random Geometric Graph, and the Barabasi-Albert model. Given the universality and wide use of those models across disciplines, our results indicate that topological phase transitions may permeate across a wide range of theoretical and empirical networks. Hereby, our paper reinforces the idea of using topological phase transitions to advance the understanding of complex systems more generally.

I. INTRODUCTION

Topology is the branch of mathematics concerned with finding properties of objects that are conserved under continuous deformations. Topological properties of a system are usually global properties that are independent of its coordinate system or a frame of reference. Over the past years, the increasing availability of technology and new experimental methods has led to a massive data production which requires high processing power and the development of new methods and concepts to analyze them. It is therefore necessary to continuously mobilize and develop abstract techniques of algebraic topology to obtain information about the shape of the data. In this context, topological data analysis (TDA) emerges as a promising methodolgical tool [1].

On the one hand, research in stochastic topology, which started with Erdos and Reyni in [2], investigated the problem of tracking the emergence of a giant component in a random graph for a critical probability threshold. The above idea was rigorously extended by Kahle [3] and Linial [4] to simplicial complexes using methods of algebraic topology. In the language of algebraic topology, the generalization of the giant component transition constitutes a major change in the distribution of n-dimensional holes, the so-called Betti numbers of a simplicial complex [3, 4], which is a generalization of a graph.

On the other hand, there is a vast literature in theoretical physics relating phase transitions with topological changes in continuous and discrete systems. Research in continuous systems was made using tools of Morse theory [5–10], while other applications relate to percolation theory [11–14]. In fact, it has been ubiquitously observed that the logarithm density of the Euler characteristic (EC) presents singular behavior around the thermody-

* f.nobregasantos@amsterdamumc.nl

namic phase transition in theoretical models [5, 15], as well as around the zeros of the EC [16–20]. This intuitively led to the introduction of topological phase transitions in simplicial complexes as the loci of the zeros of the EC [21]. In particular, it was observed in Ref. [21] that the zeros of the Euler characteristic also emerges in the vicinity of the generalization of the Giant component transition [4]. Those transitions can be seen as a generalization of percolation transitions to higher dimensional objects and were observed for the random clique complex of the Erdos-Reyni graph and for functional brain networks [21]. Later, the zeros of the Euler characteristic were also investigated independently for the models of site-percolation on lattices, the Poisson-Boolean model, and Gaussian random fields [22]. In addition to the discovery that the zeros of the Euler characteristic are good estimators for the emergence of giant shades, and that it depicts signal changes in the curvature of simplicial complexes [21], it was also observed later that the zeros of the Euler characteristic are also associated with the emergence of giant k-cycles [22].

We can compute the Euler characteristic of a complex network by associating to it a simplicial complex, i.e., a set of simplices (node, edges, triangles, tetrahedrons, etc.). In fact, we can identify the cliques, i.e. the all-to-all connected subgraphs in a network, with its k-dimensional simplices, thus obtaining the clique complex. If we denote Cl_k as the total number of k-cliques in a complex network k-dimensional simplices, the EC is given by

$$\chi = \sum_k (-1)^k \cdot Cl_k. \tag{1}$$

As an exemplary case for the ubiquity of phase transitions in complex systems, this work studies the Euler characteristic and topological phase transitions in Protein-Protein Interaction Networks (PPIN). PPIN are networks in which the nodes are proteins and the edges are the interactions between them. Topological properties of PPIN have been subject of many relevant studies [23–25]. In this work, we study the topological phase transitions in two models of PPIN growth: the totally asymmetric and the heterodimerization duplication-divergence models [26, 27].

To check whether the topological transitions identified are consistent with real PPIN data, we apply the same approach to data of yeast Gene Co-expression Networks (GCN) freely available online from the project Yeastnet v3 [28]. A GCN is a PPIN in which the edges are based on experimental measures of correlation between the activity of genes over different conditions [29]. In fact, we were are to detect topological phase transitions on the data of GCN from Yeastnet [28] and observe that this transition happens in a narrow interval for several Yeast datasets, reinforcing the idea that these topological phase transitions could be biomarkers, i.e., intrinsic properties of those yeast networks. Moreover, for yeast networks with DNA damage, we observe that the phase transition happens at a much lower threshold than the ones detected for other yeast networks, which can be seen as an indicator of network reorganization due to DNA damage [30].

We then study the topological phase transitions in the Heterodimerization model [27], a variant of the Duplication model, and compared our results with the phase transitions displayed in the *C. elegans* database [31].

Based on these empirical analyses, the last section moves towards the emergence of topological phase transitions in three classical network models that are widely used for network analysis across research fields and scientific disciplines: the Watts-Strogratz model, the Random Geometric Graph, and the Barabasi-Albert model. Hereby, we seek to illustrate that topological phase transitions permeate complex systems and might become a critical methodology to characterize the structure and intrinsic properties of complex networks.

This paper is written as follows: in section II we introduce the protein-protein interaction networks (PPIN) and discuss the duplication-divergence models; in section III we study phase transitions in PPIN through the totally asymmetric model, in section IV we explore topological phase transitions in yeast gene co-expression networks (GCN); and in section V, we mobilize the heterodimerization model and apply it to data on the nematode *C.elegans*. Moving from PPIN to broader theoretical models, in section VI, we study the topological phase transitions in classical complex network models. Finally, section VII offers our concluding remarks: we indicate promising paths for future research mobilizing topological phase transitions for a broad range of biological networks and outline the relevance of using methods and concepts of algebraic topology, in particular the Euler characteristic, to advance the understanding and study of complex systems.

II. FROM PROTEIN INTERACTION NETWORKS TO DUPLICATION-DIVERGENCE

Proteins are macromolecules that participate in a vast range of cell functions, such as DNA replication, responses to stimuli or even the transport of molecules. The interaction between proteins represents a central role in almost every cellular process [32]. For instance, understanding how proteins interact within a network can be crucial to advance the identification of cells' physiology between normal or disease state [33].

The interactions between proteins in a cell can be mathematically represented as a network (or graph), a mathematical structure composed by nodes (or vertices) and edges (or links). In this framework, proteins are the nodes and the interaction between two proteins are represented by an edge. This representation is often referred to as a Protein-Protein Interaction Network (**PPIN**) [23]. In [23] it was first observed that PPIN are scale-free networks, i.e., that there are a few nodes (proteins) participating in the majority of cellular functions. Those nodes

are connected to many other nodes with high degree, the so-called Hubs. In fact, this scale-free property can be observed in a range of other complexes systems, like the world wide web or social networks [34]. The emergence of such systems is characterized by a preferential attachment principle. New nodes that are added to the system have a higher propensity to attach to nodes with higher degree, resulting in a power law degree distribution. In such distribution, 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 [35].

Gene duplication is the process that generates new genetic material during molecular evolution. We can assume that duplicated genes produce indistinguishable proteins which, in the network formalism, implies that new proteins will have the same link structure as the original protein. Every time a gene duplicates, the proteins that are linked with the product of this gene have one extra link on the network. Thus, proteins with more links are more likely to have a neighbor to be duplicated.

The discovery of a scale-free property in PPIN gave rise to many models for generating PPIN based on the gene duplication principle [26, 27, 36–38]. Most of these models rely on divergence, a process in which genes generated by the same ancestral through duplication accumulate independent changes on their genetic profile over time. The so-called Duplication-Divergence (DD) models were first introduced in [36]. The models are based on the observation that PPIN grow after several gene duplications that can occasionally diverge in their functions over time. There are some variants of this model, which include for instance heterodimerization, arbitrary divergence [27] and random mutations [37].

In this paper, we will analyze the simplicial complex associated with two of those models theoretically, via numerical simulations, and then contrast the analysis with empirical data. The first model we explore is the totally asymmetric model of duplication and divergence network growth. It involves a single parameter p, called *retention* probability, proposed by [26]. The second model, introduced in [27], adds a second parameter q to the first model to control the heterodimerization process (defined below). We will also present empirical data from real gene co-expression networks whose behavior is compatible to what is observed in those models.

III. TOPOLOGICAL PHASE TRANSITIONS IN THE TOTALLY ASYMMETRIC MODEL

The totally asymmetric model is defined as 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 [26], is based on the hypothesis that, after duplication, there is a slight chance that the replicate node starts to develop different functions from the original one. In the model, this change is indicated by the deletion of some edges from the replica and is supposed to happen a single time during the growth of the network.

The growth of the network according to this model is defined as follows: Given a number N of proteins and a probability p ($0 \leqslant p \leqslant 1$), the model generates a graph with N nodes starting from a single one following the presented algorithm:

- Duplication: One node, as well as its edges, is randomly selected to duplicate.
- Divergence: Each duplicated edge that emerges from the replica is activated with a retention probability p. The non-activated edges are removed from the network.

Figure 1 illustrates the process.

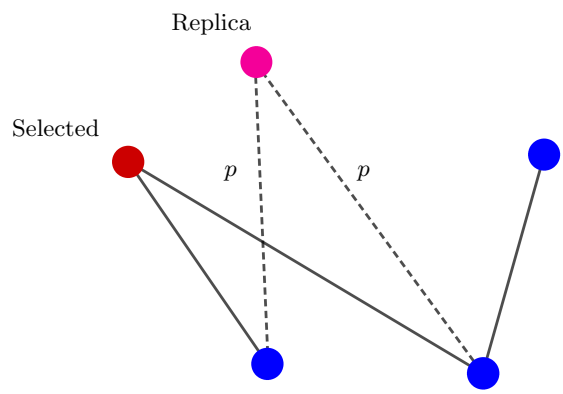


FIG. 1: Illustration of a duplication step in the totally asymmetric model for PPIN. For each duplication step a node is selected to be duplicated (in red) with its edges. Each new duplicated edge (dashed lines) that goes from the replica (pink) can be activated with independent probability p. The non-activated edges are removed.

The duplication step aims to capture 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 [26] consider that any duplication that leads to an isolated node should not be considered. However, since isolated nodes are also considered when computing the Euler characteristic, we kept them to guarantee consistency between theory and experiment. In fact, this approach can be seen as a different perspective 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 isolated duplicated nodes will make the modeling more appropriate to match topological phase transitions with the experiments.

For the computation of the Euler characteristic (or the Euler entropy), observe that the Duplication-Divergence 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 $\chi = N - E$, where N represents the total number of nodes, and E the total number of edges of the network. This simplicity enables us to analytically obtain the expected mean value of the EC in terms of the probability p for a graph with N nodes, by

$$\langle \chi \rangle = N - \langle E \rangle. \tag{2}$$

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

$$\langle E \rangle(p) = \prod_{k=1}^{N-2} \left(\frac{2}{k+1} p + 1 \right) \tag{3}$$

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

$$\langle E \rangle(p) = \prod_{k=1}^{N-2} \left(\frac{2}{k+1} p + 1 \right) = N. \tag{4}$$

A network that has more edges than nodes certainly has cycles [39]. So, the first topological 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 [2].

We define topological phase transitions as the singularities of the Euler entropy or the zeros of the Euler characteristic. This definition is inspired on results obtained for a range of continuous systems [5], and was introduced to discrete data in [21]. In figure 2, we observe the behavior of S_χ as a function of the probability p of the totally asymmetric model, for a network with $1,000$ nodes. One can observe that there is a single singularity when p reaches the critical value of ≈ 0.56 (dashed black line), in the vicinity of the critical probability of the Giant component transition.

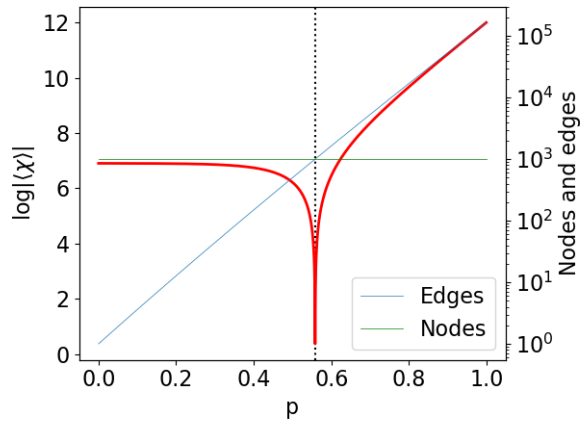


FIG. 2: Euler entropy $S_\chi = \log |\langle \chi \rangle|$ as function of the retention probability p for the Duplication Divergence model with $1,000$ nodes. S_χ was determined by (2). The singularities $S_\chi \to -\infty$, or the zeros of χ locate the topological phase transitions on the duplication divergence model. We put the number of nodes and edges in a logarithmic scale to make visible the change of domain between the number of nodes and the number of edges that occurs at the phase transition's critical probability.

To characterize this phase transition and understand its implications, we analyze the model through another topological parameter, known as Betti numbers. In general, Betti numbers, indicated by β_n, are defined as follows: β_0 is the total number of connected components. β_1 is the number of cycles on the graph. For $n > 1$ the structure becomes more complex, but roughly speaking, β_n indicates the number of n-dimensional "holes" in the network [40].

Figure 3 presents the behavior of the Betti numbers β_0 and β_1 as functions of p compared with the Euler characteristic. Here, we were not able to achieve an analytic expression for the Betti numbers, they were obtained numerically by generating and averaging $1,000$ simulations for each value of p (from 0 to 1 at steps of 10^{-2}), $1,000$ nodes each.

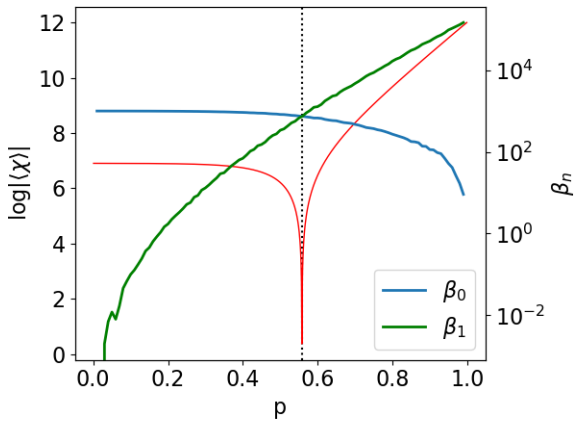


FIG. 3: Betti numbers β_0 and β_1 for the Duplication Divergence model as function of the retention probability p (N = 1,000). At the first phase the network is fragmented, with many components. As p increases, β_1 (i.e, the number of cycles) surpasses β_0 when p crosses the transition point, indicating a major change in 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, $\chi = 0$, thus confirming the results discussed theoretically by Linial and Peled[3]. For $p < p_c$ the network is divided into many components because there is a high chance that duplicated nodes do not connect with any neighbor of the original nodes. As p increases, the number of components decreases and we have an 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.

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

IV. TOPOLOGICAL PHASE TRANSITIONS IN GENE CO-EXPRESSION NETWORKS

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 [41], but also functional interactions. In fact, a group of proteins can perform a common biological function without actually being in direct contact, regulating a biological process or making common use of another molecule [32].

To identify functional interactions, a common method consists in analyzing gene co-expression patterns. This method is based on the assumption that genes with correlated activity produce interacting proteins [42]. Therefore, by mapping the correlation between the activity of genes, we can build the so-called co-expression network (GCN) [32].

A GCN is a PPIN in which the edges are based on experimental measures of correlation between the activity of genes over different conditions. Each node of the network corresponds to a gene. An edge connects a pair of genes if they present similar expression patterns over all the experimental conditions [29]. GCN are often used to identify groups of genes that, over various experimental conditions, displays correlated expression profiles. Based on the "guilt-by-association" principle [43], it is possible to hypothesize that those genes share a common functionality. Therefore, the understanding and comparison of topological features across co-expression networks may provide useful information about the strengths (and weaknesses) of the theoretical models used to infer different co-expression networks [44].

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 available in [28]. Our aim is to test whether we observe similar topological phase transitions on real data of GCN, and if such transitions provide us any useful insights about the structural properties of the network. Here, we analyzed 48 empirical GCN from *Saccharomyces cerevisiae*, also known as baker's yeast, available online from the project YeastNet v3 [28]. The whole data-set covers around 97% of the yeast coding genome ($5,730$ genes). Networks have $\approx 800 - 3,000$ nodes and between $7,000$ and $64,000$ links. They were collected through diverse experimental approaches.

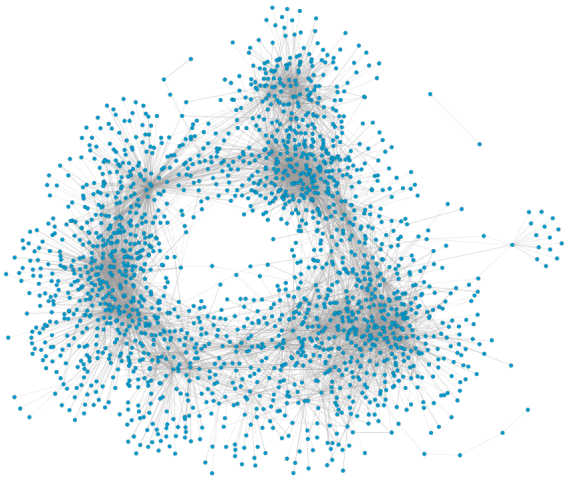


FIG. 4: Network representation of gene co-expression network for the yeast using Cytoscape [45]. The presented yeast network has $1,893$ nodes, $7,955$ links and it was obtained from Yeastnet v3 [28].

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 then. We associate to each edge a weight that corresponds to the absolute value of the correlation between the expression levels of the nodes in the network.

After creating GCN from the co-expression matrices, we proceed with a numerical process called filtration. For a given $\varepsilon \in [0,1]$, this process requires the removal of 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 links with correlation greater than $1-\varepsilon$. We perform this procedure until ε reaches 1 and the network becomes fully connected. In this paper, we used the filtration process to track how the topology of the GCN changes as a function of the correlation between the genes.

In contrast to the results of the previous case, where the EC was written analytically as function of the probabilistic parameter p for the totally asymmetric model, the EC can be obtained numerically as function of the threshold ε by

$$\chi(\varepsilon) = \sum_{k=1}^{N} (-1)^{k+1} Cl_k(\varepsilon), \tag{5}$$

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

As the computation of topological invariants is an NP-complete problem [46], here, we only illustrate data-sets where the computation was feasible, namely, to 20 Yeast networks. Figure 5 shows the Euler entropy as a function of ε for the 20 selected 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. This occurs because, as ε increases, the network becomes denser, and the computation of the cliques become extensive. For this reasons, in some of the datasets, we could not compute the Euler characteristic to a threshold where one could detect a topological phase transition. Nonetheless, 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, in red, where the singularity happened at $\varepsilon \approx 0.573$. This network comes from an experiment whose goal was to evaluate response to DNA damage due to methylmethane sulfonate and ionizing radiation [30]. In fact, It is known that eukaryotic cells respond to DNA damage by rearranging its cycle and modulating gene expression to ensure efficient DNA repair [30]. Therefore, our analysis suggests that the Euler entropy could be sensible to this reorganization process in the damaged yeast network. Further investigating the relation between GCN with DNA damage and percolation transitions will be an important step to establish whether the zeros of the Euler characteristic can be used as a topological bio-marker of such systems, with the potential to detect DNA damage.

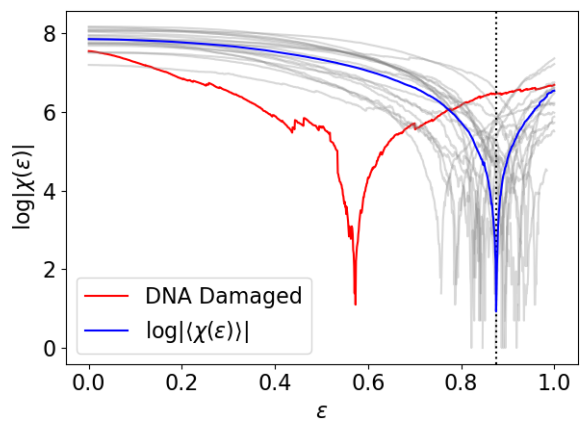


FIG. 5: Euler entropy S_χ as a function of the correlation threshold ε for 20 co-expression networks from Yeastnet database. Each gray line corresponds to a unique co-expression network, while the blue line is the average of the gray ones. Most of the topological phase transitions were detected on those networks for ε in the average critical value $\varepsilon \approx 0.876$, except for the yeast network with DNA damage, where the transition happens for $\varepsilon \approx 0.573$ (red curve). This particular experiment was designed to measure the response to DNA damage in the network [30], and the shift in the critical threshold of the topological transition is apparently sensitive to that response.

In order to provide a deeper characterization of the topological phase transition of GCN, we also calculated the Betti curves β_0 and β_1 for the same yeast networks-From a theoretical perspective, a topological phase tran-

sition should occur at vicinity of the zeros of EC and when $\beta_0 \approx \beta_1$. In Figure 6, we illustrate the averaged Betti curves for the above mentioned dataset.

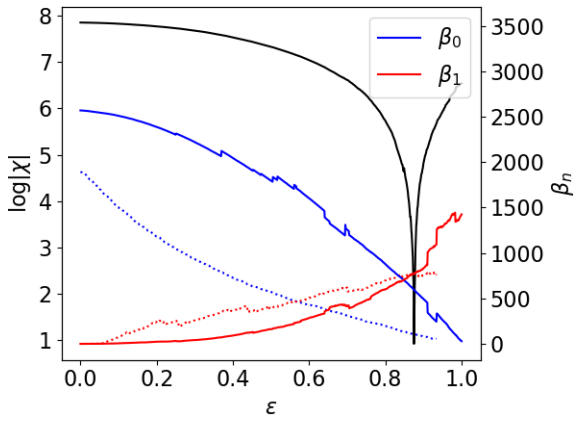


FIG. 6: Average Betti curves (β_0 and β_1 as a function of the correlation threshold ε). Observe that, as with the totally asymmetric model analyzed previously, the domain of β_0 and β_1 shifts at the vicinity of the singularity of the Euler entropy (black line), i.e., at the zeros of the Euler characteristic. Dotted lines correspond to the Betti numbers of the Yeast network with DNA damage.

We can observe agreement between the threshold transition of S_χ and the threshold where $\beta_1 \approx \beta_0$, in analogy with the topological transitions reported for theoretical models of random simplicial complexes [4, 47] and for functional brain networks [21].

For values of the correlation threshold below the transition, the network is fragmented in many components, since β_0 is greater than β_1. As ε gets closer to the transition, more edges are added to the network, lowering the number of connected components, and changing the network structure to a denser one and with more cycles, i.e., higher β_1. Once the number of connected components approaches 1, i.e., $\beta_0 \to 1$, almost every new edge has high probability to create more loops, and the topological transition happens precisely when the number of loops (β_1) is greater than the number of connected components (β_0).

These observations about the behavior of Betti numbers are compatible with the results observed in [48], and provide some possible interpretation for this shift at the critical point of the topological phase transition. In [48] a linear correlation of $R = -0.55$ was observed 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 in the network. The higher β_1, the higher is the complexity, which implies lower chances of survival [48]. This study provides 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 yeast gene co-expression networks, we observed that, for one of the networks, the transition indicated by the Euler characteristic 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 [4]. Thus, during the filtration process, the network corresponding to yeast with DNA damage reaches the topological phase transition earlier than the other data (where the transitions happened near $\varepsilon \approx 0.876$), indicating a rapid appearance of cycles in the network, i.e., the increase of network complexity. Further experimental studies are desired to evaluate the statistical significance of the shift in topological phase transitions for yeast PPIN under DNA damage. Our results, and recent results in functional brain networks [21] together with arguments discussed in [48] reinforce the hypothesis that topological phase transitions have the potential to be used as intrinsic biomarkers for protein interaction networks more generally.

V. TOPOLOGICAL PHASE TRANSITIONS IN THE HETERODIMERIZATION MODEL

We now move towards a model where protein interactions display more complex dynamics and could potentially capture different empirical data for PPIN. In fact, this model is a variation of the totally asymmetric model discussed in the previous sections. Applied to networks, the model is described as follows: Given a number of nodes N and two probabilistic parameters p and q, the model generates a graph from two nodes and a single edge following the steps below:

- Duplication: One node, as well as its edges, is randomly selected to duplicate.
- Divergence: Each duplicated edge that emerged from the replica is activated with retention probability p.
- Heterodimerization: The replicated and the original nodes can 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 [49]. It is also known that this model produces cliques with similar size and quantity than those observed in some real PPIN [27], contrasting with the totally asymmetric model where we could only observe cliques with small sizes.

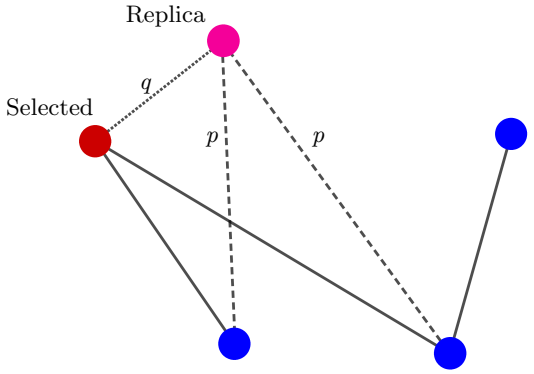


FIG. 7: Illustration of a duplication step in the heterodimerization model for PPIN. For each duplication step a node is selected to be duplicated (red) with its edges. Each new duplicated edge (dashed lines) that goes from the replica (pink) can be activated with independent probability p. Also, an 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. Given the complexity of the distribution of cliques in these models [27], we were only able to generate these networks numerically, and we further compared with experimental data. In Figure 8a, we observe the average of the Euler entropy S_χ as a function of the retention probability p for a representative network with $1,000$ nodes and different values of q.We see that the heterodimerization parameter q induces the emergence of further topological phase transitions. These additional topological phase transitions emerge because the heterodimerization step makes the appearance of the highest possible order. The higher the value of q, the more probable it becomes for cliques of larger sizes to appear and, therefore, higher is the likehood of further topological phase transitions. In this context it is important to note that [50] reports the abundance of large cliques in real PPIN data, which makes the heterodimerization model more compatible with empirical data.

Given the bigger size, higher number and complex distribution of the high dimensional cliques in this model, 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 did not compute the Betti numbers as we did for the totally asymmetric model. Nevertheless, to reinforce the significance of our analysis, we compare our numerical simulations to experimental data for PPIN which, in fact, presents a similar phase transitions profile.

In Figure 8b (top) we illustrate the Euler entropy of a PPIN for the *C. elegans*. The data for this nematode was obtained from the Wormnet v.3 database [31]. 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 yeast GCN 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 were able to achieve a similar profile of the Euler entropy by simply setting the representative parameters, as can be seen at the bottom of Figure 8b, which 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 at $2,219$ so that we can compare it with the *C. elegans* empirical data. We also set the value of the heterodimerization probability q to 0.05. Even though this value was only set for comparison purposes, not intending to perform an optimal fit to the empirical data, it is within the same order of magnitude of the value of $q = 0.03$ used in [27], where the authors managed to achieve a good approximation for the clique distribution across the PPIN considered.

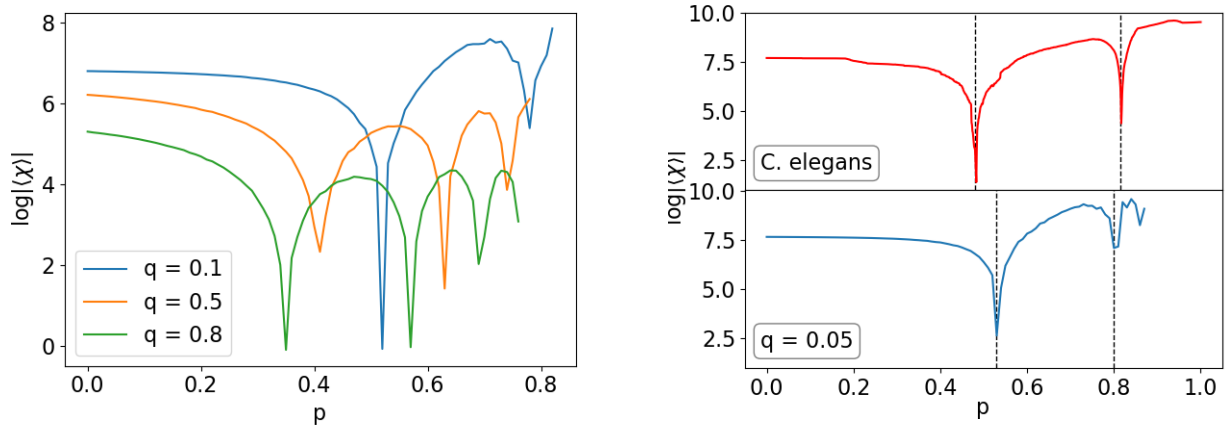
This analysis suggests the presence of 'dimmers' during the growth of the *C. elegans* network. Therefore, the topological phase transitions displayed by the heterodimerization model are more appropriate for this empirical data when compared with the Duplication-Divergence model.

Our theoretical and empirical analysis of complex biological networks reinforces the strong evidence that the zeros of the Euler characteristic can be interpreted as an intrinsic fingerprint of a given network [21]. This suggests that topological phase transitions are a critical analytical tool to characterize complex systems more generally.

VI. TOPOLOGICAL PHASE TRANSITIONS IN COMPLEX SYSTEMS

In the remainder of this paper, we illustrate how commonly used complex networks models can be studied using the zeros of Eq. (1), the Euler characteristic χ. In the previous sections, we showed that topological phase transitions were associated with the $\log|\langle\chi\rangle|$ of the protein-protein interaction network using the Duplication-Divergence and the Heterodimerization models [36], which were subsequently compared against empirical data.

Moreover, given the recent use of the zeros of the Euler characteristic as a potential topological biomarker in functional brain networks and as markers for percolations in random field models, it becomes highly interesting to investigate the presence of topological phase transitions in a wider range of systems and network models, namely the Watts-Strogatz model, the Random Geometric model, and the Barabasi-Albert model. Together, these models capture a wide range of phenomena in network science across numerous fields, such as small-world property, percolation transitions, preferential attachment, to name a few.The observation of topological



(a) (b)

FIG. 8: Euler entropy of the Duplication-Divergence model with Heterodimerization. (a) Average of the Euler entropy as function of the retention probability p for different values of q. Note that, in contrast to the previous model, there are more transitions as we increase the value of q. This is because the heterodimerization step makes it possible for cliques of different sizes to appear, as observed in real PPIN [27]. (b) (top) Euler entropy of PPIN from *C. elegans* in contrast with an illustrative simulation for the heterodimerization model with $q = 0.05$ (bottom). Both have two singularities.

phase transitions in such models is therefore a critical step towards the use of such methodology in complex systems more generally.

To bridge the empirical analysis of protein structures and the discussion of theoretical network models, we first explore the classical Watts-Strogatz network, a small-world model [51–53] that can also be used to study PPIN.

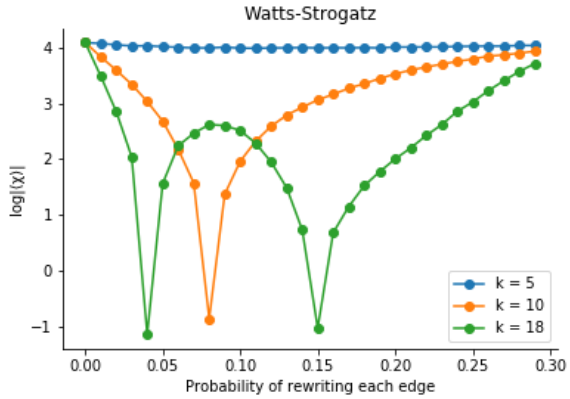


FIG. 9: Topological phase transition for the Watts-Strogatz small-world model, for three values of first neighbors connections k, namely $k = 5$ (blue), $k = 10$ (orange), and $k = 18$ (green).

The small-world model is constructed from a ring with n nodes, and each node is connected with its k first neighbors or $k-1$ neighbors if k is odd. After that, each edge is randomly rewired with probability $p \in]0.01, 0.30[$ with stepsizes of 0.01. Once built, we can study the topological phase transitions displayed by this network as a function of the rewiring probability. In fact, we show in Figure 9 that a small-world network of $n = 60$ nodes also has the characteristic transition observed in the previous section, besides the reduced number of vertices used. This is a strong indication or the versatility of the methodology used in this article to localize non-trivial topologies.

We can proceed with a more detailed inspection of the Watts-Strogatz model by changing the number k of first neighbors. For $k = 5$, Figure 9 shows that there is an absence of transitions. Increasing the number of first neighbors to $k = 10$, we can see one transition at $p = 0.085$, and for $k = 18$ we have two transitions, the first at $p = 0.045$ and the second at $p = 0.145$. Observing these topological transitions requires the use of an ensemble with thousands of replicas for each set of p and k parameters that are subsequently averaged, which restricted our analysis to $n = 60$ nodes. However, the use of a limited number of nodes and the possibility of differentiating among different topologies is a clear strength of the present methodology, since real-world data sometimes has few nodes to be analyzed.

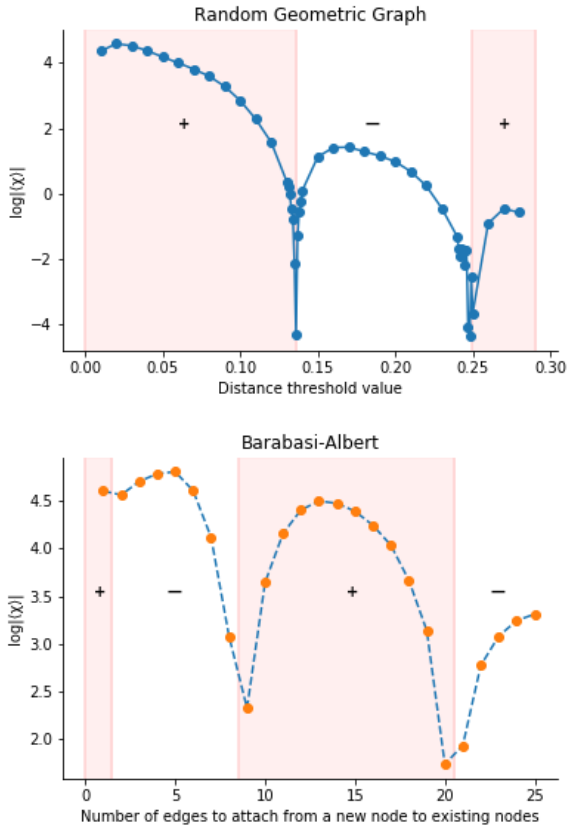


FIG. 10: Topological phase transitions for two classical complex networks models. Top: Random Geometric graph. Bottom: Barabasi-Albert model. In both plots, the curves connecting the points are guides to the eyes.

Once studied through a small-world model, we can go further and, in a second step, investigate the topological phase transitions of the Random Geometric Graph model [54]. This one is generated by using the number of nodes n and a distance r as inputs. First, one chooses n nodes uniformly at random in the unit cube, which are subsequently connected by an edge if the distance between them is lower than or equal to r. We can then study the topological phase transitions in this network as a function of the distance r.

We show the Euler entropy and its topological phase transitions for the Random Geometric model in Figure 10 for $n = 100$ nodes and $0 < r \leq 0.28$ with a step size of 0.01, averaged from an ensemble of 6,000 replicas. Observe that the phase transitions also correspond to changes in the signs of χ, which can be understood as changes in the sign of the mean curvature of the nodes [21] as described by the Gauss-Bonnet theorem for networks [55]. We illustrate the sign changes by shading Figure 10 according to the sign of the Euler characteristic. This interpretation can be important when we want to investigate topological changes in models whose filtration parameter is not necessarily continuous, as we will see below.

Lastly, we illustrate the topological changes in the Barabasi-Albert network model [56]. This model is known for introducing random, scale-free networks using a preferential attachment mechanism. The network starts with few nodes and no edges and a number of edges attached to the new node with probability proportional to the node's degree. In contrast to the two previous network models studied in this section, the Barabasi-Albert model has two integers parameters, namely, the maximum number of nodes n and the number of edges m to attach to a new node to existing nodes.

We can analyse the topological phase transitions in the Barabasi-Albert network for a fixed n as a function of m. However, given the discrete nature of the control parameter m of this model, we were only able to infer the topological phase transitions via changes in the signal of the Euler characteristic, which tracks the changes in the curvature of the nodes via the Gauss-Bonnet theorem for networks [55]. Figure 10 shows three topological phase transitions as a function of the number m of edges attached using an ensemble of 2000 replicas. Similarly to the Random Geometric Graph, the transitions are revealed by a sign change of χ, as indicated by the shaded areas for positive Euler characteristic. Our inferred results for the first transition, $m = 2$, is in agreement with the percolation transition for the Barabasi-Albert model [57].However, little is known regarding curvature changes in the Barabasi Albert model for higher values of m. For this particular model, one interesting analogy could be made regarding the relationship between Lyapunov exponents, curvature and phase transitions known in statistical mechanics [58, 59]. In fact, Lyapunov stability theory was adapted to the context of networks [60], and particularly to the Barabasi-Albert model [61].The authors studied for which values of m anomalous events occur in this network. We noticed that some of the critical invervals for m in Ref. [61] are analogous to the changes in the curvature as a function of m observed in our work. Therefore, the connection between then deserves further investigations.

Ultimately, the aim of this section was to illustrate that topological phase transitions, evidenced by the zeros of the Euler characteristic, are present in three classical network models that are widely used in network analysis across a variety of fields. This reinforces the ubiquity of such topological phase transitions in complex systems more generally, as suggested in [21].

VII. CONCLUSIONS

In this work, we investigated the concept of topological phase transitions in a wide range of theoretical and empirical networks. Those transitions are characterized by the zeros of the Euler characteristic at a critical probability or threshold and represent a generalization of the giant component percolation transition introduced by Erdos [2–4] and recently expanded to data driven complex

networks [21]. The zeros of the Euler characteristic also indicate signal changes in the mean curvature [55] and the emergence of a giant k-cycle in a simplicial complex. We first focused our attention on protein-protein interaction networks, particularly to networks generated by the totally asymmetric duplication-divergence model. We verified that such transitions correspond to the emergence of a giant component in the network, as observed in several yeast and gene co-expression networks datasets [28]. Our results give strong support to the hypothesis that percolation transitions are topological biomarkers of a network [21]. In fact, by analyzing a network for yeast under DNA damage [30], we found that the transition point shifted to a much lower value compared to the interval in which the topological phase transitions of the other yeast network datasets took place. Although this result deserves further investigation, we provide evidence that the zeros of the Euler characteristic can be seen as suitable topological invariants to distinguish macroscopic properties of the yeast networks.

In a second step, we explored topological phase transitions in the *C. elegans* PPIN. In contrast to the yeast networks, the same approach gave rise to two topological phase transitions, indicating that the totally asymmetric model is insufficient to capture the topological aspects of the *C. elegans* PPIN data properly. This suggests that there is another network process at play that leads to the PPIN growth. 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 model and observed that, for an illustrative set of parameters, we could obtain an Euler entropy profile similar to the one observed for the *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.

It is important to emphasize that other models suggest different processes for PPIN growth, implying that our analysis did not exhaust the possibilities of models considered for PPIN. In [27] for example, the authors present an arbitrary divergence model in which they could replicate with high confidence the clique distribution observed for PPIN data. Other models simulate different aspects of the growth of a PPIN [27, 37, 38] and the analysis of such models through the lenses of topological phase transitions and the Euler characteristic would give us precious information about PPIN.

More generally, in order to further explore topological phase transitions in complex systems, we extended our analysis to three classical network models, namely: the Watts-Strogatz model [51], the Random Geometric model [54], and the Barabasi-Albert model [56]. We showed that topological phase transitions permeate these models which are widely used across disciplines.

In short, this work contributes to a better understanding of the topological properties of PPIN and opens up the perspective of using topological phase transitions in complex systems and network science more generally.

VIII. ACKNOWLEDGEMENTS

This work is funded by CAPES, CNPq, and FACEPE (Brazilian Agencies). F. A. N. Santos would like to express his gratitude for the useful discussions with Linda Douw, Mauricio D. Coutinho-Filho, Ernesto P. Raposo, Mauro Coppeli, James Charles Phillips, Stefan Boettcher, Bertrand Berche, Katharina Natter and Fernando Moraes during the development of this work. E. Amorim would like to thank V. A. Caldas and T. Shimizu for the learnfull visiting period at AMOLF.

We would like to acknowledge FACEPE Pronex APQ-0602-105/14 and FACEPE-Fulbright ARC-0103-1.01/17.

[1] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 1 2009.

[2] P Erdös and A Rényi. On Random Graphs I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

[3] Matthew Kahle. Topology of random simplicial complexes: a survey. 1 2013.

[4] Nathan Linial and Yuval Peled. On the phase transition in random simplicial complexes. *Annals of Mathematics*, 184(3):745–773, 11 2016.

[5] Marco Pettini. *Geometry and topology in Hamiltonian dynamics and statistical mechanics*. Springer-Verlag New York, New York, 2007.

[6] Michael Kastner. Phase transitions and configuration space topology. *Reviews of Modern Physics*, 80:167–187, 2008.

[7] F A N Santos, L C B da Silva, and M D Coutinho-Filho. Topological approach to microcanonical thermodynamics and phase transition of interacting classical spins. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(1):013202, 1 2017.

[8] Mark Buchanan. It's just a phase... *Nature Physics*, 4(1):5–5, 1 2008.

[9] Matteo Gori, Roberto Franzosi, and Marco Pettini. Topological origin of phase transitions in the absence of critical points of the energy landscape. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(9), 9 2018.

[10] Michael Kastner and Oliver Schnetz. Phase transitions induced by saddle points of vanishing curvature. *Physical Review Letters*, 100(16), 4 2008.

[11] Leo Speidel, Heather A Harrington, S Jonathan Chapman, and Mason A Porter. Topological data analysis of continuum percolation with disks. 2018.

[12] B. L. Okun. Euler characteristic in percolation theory. *Journal of Statistical Physics*, 59(1-2):523–527, 4 1990.

[13] K. R. Mecke and H. Wagner. Euler characteristic and

related measures for random geometric sets. *Journal of Statistical Physics*, 64(3-4):843–850, 8 1991.

[14] Richard A Neher, Klaus Mecke, and Herbert Wagner. Topological estimation of percolation thresholds. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(01):P01011, 1 2008.

[15] F. A.N. Santos and M. D. Coutinho-Filho. Topology, symmetry, phase transitions, and noncollinear spin structures. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 2009.

[16] D. Gandolfo. About topological invariants and statistical mechanics. *Physica A: Statistical Mechanics and its Applications*, 358(1 SPEC. ISS.):22–29, 12 2005.

[17] Philippe Blanchard, Christophe Dobrovolny, Daniel Gandolfo, and Jean Ruiz. On the mean Euler characteristic and mean Betti's numbers of the Ising model with arbitrary spin. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(03):P03011–P03011, 3 2006.

[18] J. A. Rehn, F. A.N. Santos, and M. D. Coutinho-Filho. Combinatorial and Topological Analysis of the Ising Chain in a Field. *Brazilian Journal of Physics*, 42(5-6):410–421, 12 2012.

[19] Philippe Blanchard, Santo Fortunato, and Daniel Gandolfo. Euler–Poincaré characteristic and phase transition in the Potts model on Z2. *Nuclear Physics B*, 644(3):495–508, 11 2002.

[20] Ph Blanchard, D Gandolfo, J Ruiz, and S Shlosman. On the Euler-Poincaré Characteristic of the Random Cluster Model. pages 523–545, 2003.

[21] Fernando A. N. Santos, Ernesto P. Raposo, Maurício D. Coutinho-Filho, Mauro Copelli, Cornelis J. Stam, and Linda Douw. Topological phase transitions in functional brain networks. *Physical Review E*, 100(3):032414, 9 2019.

[22] Omer Bobrowski and Primoz Skraba. Homological Percolation and the Euler Characteristic. 10 2019.

[23] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 5 2001.

[24] Soon-Hyung Yook, Zoltán N. Oltvai, and Albert-László Barabási. Functional and topological characterization of protein interaction networks. *PROTEOMICS*, 4(4):928–942, 4 2004.

[25] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science (New York, N.Y.)*, 296(5569):910–3, 5 2002.

[26] I Ispolatov, P L Krapivsky, and A Yuryev. Duplication-divergence model of protein interaction network. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 71(6 Pt 1):061911, 6 2005.

[27] I Ispolatov, P L Krapivsky, I Mazo, and A Yuryev. Cliques and duplication–divergence network growth. *New Journal of Physics*, 7(1):145–145, 6 2005.

[28] Hanhae Kim, Junha Shin, Eiru Kim, Hyojin Kim, Sohyun Hwang, Jung Eun Shim, and Insuk Lee. YeastNet v3: A public database of data-specific and integrated functional gene networks for Saccharomyces cerevisiae. *Nucleic Acids Research*, 2014.

[29] Juli Petereit, Frederick C. Harris, and Karen Schlauch. petal: A novel co-expression network modeling system. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 234–241. IEEE, 11 2015.

[30] Audrey P. Gasch, Mingxia Huang, Sandra Metzner, David Botstein, Stephen J. Elledge, and Patrick O. Brown. Genomic Expression Responses to DNA-damaging Agents and the Regulatory Role of the Yeast ATR Homolog Mec1p. *Molecular Biology of the Cell*, 12(10):2987–3003, 10 2001.

[31] Ara Cho, Junha Shin, Sohyun Hwang, Chanyoung Kim, Hongseok Shim, Hyojin Kim, Hanhae Kim, and Insuk Lee. WormNet v3: a network-assisted hypothesis-generating server for Caenorhabditis elegans. *Nucleic Acids Research*, 42(W1):W76–W82, 7 2014.

[32] Danila Vella, Italo Zoppis, Giancarlo Mauri, Pierluigi Mauri, and Dario Di Silvestre. 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*, 2017:6, 2017.

[33] Marc Vidal, Michael E Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–98, 3 2011.

[34] Albert-László Barabasi and Márton Pósfai. *Network Science by Albert-László Barabási*. Cambridge University Press, Cambridge, 1 edition, 2016.

[35] Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2 2004.

[36] Alexei Vázquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Modeling of Protein Interaction Networks. *Complexus*, 1(1):38–44, 2003.

[37] Romualdo Pastor-Satorras, Eric Smith, and Ricard V. Solé. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, 222(2):199–210, 5 2003.

[38] Nadia Farid and Kim Christensen. Evolving networks through deletion and duplication. *New Journal of Physics*, 8(9):212–212, 9 2006.

[39] Béla Bollobás. *Modern graph theory*. Springer, 1998.

[40] Herbert. Edelsbrunner and J. (John) Harer. *Computational topology : an introduction*. American Mathematical Society, 2010.

[41] Javier De Las Rivas and Celia Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, 6(6):e1000807, 6 2010.

[42] A. Grigoriev. 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*, 29(17):3513–3519, 9 2001.

[43] Riet De Smet and Kathleen Marchal. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*, 8(10):717–729, 10 2010.

[44] Ramon Xulvi-Brunet and Hongzhe Li. Co-expression networks: graph properties and topological comparisons. 26(2):205–214, 2010.

[45] P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, N Amin, B Schwikowski, and T Ideker. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, 13(11):2498–2504, 2003.

[46] D.-Z Du, P M Pardalos, Immanuel M Bomze, Marco Budinich, Panos M Pardalos, and Marcello Pelillo. Handbook of Combinatorial Optimization The Maximum Clique Problem. Technical report.

[47] Csaba D. Toth, Joseph O'Rourke, and Jacob E. Good-

man. *Handbook of Discrete and Computational Geometry, Third Edition*. Chapman and Hall/CRC, 3 edition, 2017.

[48] Sebastian Benzekry, Jack A. Tuszynski, Edward A. Rietman, and Giannoula Lakka Klement. Design principles for cancer therapy guided by changes in complexity of protein-protein interaction networks. *Biology Direct*, 10(1):32, 12 2015.

[49] I. Ispolatov, Anton Yuryev, Ilya Mazo, and Sergei Maslov. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Research*, 33(11):3629–3635, 6 2005.

[50] D A Fell, H Jeong, B Tombor, R Albert, Z N Oltvai, and A L Barabasi. Protein complexes and functional modules in molecular networks. Technical report, 2001.

[51] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (393):440–442, 6 1998.

[52] A. del Sol, H. Fujihashi, and P. O'Meara. Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, 21(8):1311–1315, 4 2005.

[53] Neil R. Taylor. Small world network strategies for studying protein structures and binding. *Computational and Structural Biotechnology Journal*, 5(6):e201302006, 2013.

[54] Mathew Penrose. *Random Geometric Graphs Oxford Studies in Probability, 5*. Oxford University Press, USA, 2003.

[55] Oliver Knill. A graph theoretical Gauss-Bonnet-Chern Theorem. 11 2011.

[56] Albert László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 10 1999.

[57] Wolfgang Pietsch. Derivation of the percolation threshold for the network model of Barabási and Albert. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 73(6), 2006.

[58] Lando Caiani, Lapo Casetti, Cecilia Clementi, and Marco Pettini. Geometry of dynamics, lyapunov exponents, and phase transitions. *Physical Review Letters*, 79(22):4361–4364, 1997.

[59] F. Argoul and A. Arneodo. Lyapunov exponents and phase transitions in dynamical systems. pages 338–360. 1986.

[60] Diego Ruiz and Jorge Finke. Lyapunov-Based Anomaly Detection in Highly-Clustered Networks. *Journal of Statistical Physics*, 172(4):1127–1146, 8 2018.

[61] Diego Ruiz and Jorge Finke. Lyapunov–based Anomaly Detection in Preferential Attachment Networks. *International Journal of Applied Mathematics and Computer Science*, 29(2):363–373, 2019.