TÉCNICO LISBOA

# Uncovering the organizing principles behind signed biological networks

**Gonçalo Archer Franco Frazão**

Thesis to obtain the Master of Science Degree in

## Biomedical Engineering

Supervisor(s): Prof. Yasser Rashid Revez Omar
Doutor Nuno Luís Barbosa Morais

### Examination Committee

Chairperson: Prof. João Pedro Estrela Rodrigues Conde
Supervisor: Prof. Yasser Rashid Revez Omar
Member of the Committee: Prof. Ruy Miguel Sousa Soeiro de Figueiredo Ribeiro

**November 2018**

To my Father, my Mother, and to my younger siblings: D, L, V and T.

# Acknowledgments

A special thanks to my supervisor, Prof. Yasser Omar, for his invaluable wisdom and experienced guiding. He who always managed to prod our ambition, at the same time he was calling me and Bruno back to Earth. I have his laborious example in the highest regard.

A special thanks to, Dr. Nuno Morais, from IMM, for his precious feedback on the biological interpretation of the results.

A special thanks to Bruno Coutinho, the person who interrupted his work innumerable times to attend my technical and existential doubts. Thanks for tutoring my introduction into the marvelous new world of network science, for all the help with *Python* and *C++*, and for the times you had to either curb my impatience or lift my mood.

A special thanks to István Kovács, from the Northeastern University, in Boston, the person who noticed the gap in the literature, of studies on the structure of undirected signed biological networks. Thanks for your continuous feedback on multiple Skype talks over the ocean, and for lending your still unpublished randomization model.

It was great to have the possibility to do my research as part of an international collaboration.

A final special thanks to the Physics of Information and Quantum Technologies Group, at IT, for hosting me during my research project.

# Resumo

Redes são um instrumento útil para descrever sistemas biológicos. Codificam os elementos do sistema em nós e as suas interações em arestas. Redes com sinais codificam informação adicional sobre a interação sob a forma de um sinal em cada aresta.

Apesar dos abundantes estudos sobre padrões locais em estruturas de redes, especialmente redes sociais, que tenhamos conhecimento, estes ainda não foram estudados em redes biológicas com sinais e sem direção. Padrões locais já se provaram úteis na previsão de novas interações em redes construídas a partir de dados incompletos e ruidosos, apoiando previsões funcionais e delimitação de módulos nas redes.

O objetivo deste projeto é a identificação dos padrões típicos em redes de interações genéticas, em particular triângulos e quadrados. A partir de dados disponíveis publicamente, foram construídas duas redes descrevendo as interações entre genes essenciais e não essenciais em levedura.

Para a rede essencial descobrimos que os triângulos com um produto das arestas negativo são favorecidos pela estrutura da rede, enquanto triângulos com um produto positivo são preteridos. Portanto, para definir corretamente as regras estruturais para triângulos em redes genéticas, há que inverter os sinais do Structural Balance, a teoria vigente para *balance* em redes sociais. Nos quadrados verifica-se o contrário, padrões com um produto positivo parecem ser favorecidos, já os com produto negativo são preteridos, de acordo com o espectável em redes sociais.

Estes resultados constituem uma primeira pedra na resolução de um problema importante em Network Science, previsão de ligações, no contexto de redes biológicas com sinal.



**Palavras-chave:** redes com sinais, redes de interações genéticas, Structural Balance, aleatorização de redes, previsão de sinais

# Abstract

Networks are useful tools to describe biological systems, encoding elements of the system and their interactions in nodes and edges, respectively. Signed networks encode further information about the interaction in the form of a signal in each edge.

Despite extensive studies of local patterns in the network structure, especially in social networks, to the best of our knowledge, these have not yet been studied in undirected signed biological networks. Local patterns have proved useful to predict new interactions in networks built from noisy and incomplete datasets, supporting the delimitation of functional modules and even the prediction of single-element function.

This project's goal is to identify the typical patterns in genetic interaction networks, more specifically triangles and squares. Two networks describing the interactions between essential and non-essential genes in yeast were constructed, from data publicly available.

For the essential network, we found that triangles with a negative edge product are favored in the network structure, while triangles with a positive edge product are deprecated. Thus, to correctly define the structural rules in genetic networks, for triangles, one must invert the signals of Structural Balance, a very well-known and long-standing theory for balance in social networks. For squares, the opposite is verified, patterns with positive edge product appear to be favored, while negative product squares appear deprecated, in accordance with what expected for social networks.

These results are a first step to tackle a very important issue in Network Science, link prediction, in the context of signed biological networks.



**Keywords:** signed network, genetic interaction network, Structural Balance, network randomization, sign prediction

# Contents

# Chapter 1

# Introduction

Network theory is a powerful tool to describe and study complex biological systems. In fact, the study of the network structure allows the delimitation of modules in the system [1], the prediction of function for elements of the system [2] and even predicting previously unknown synergies between elements of the system [3]. However, it seems there is a gap in the literature regarding the structure of undirected signed biological networks, and more specifically on which are the preponderant local patterns - motifs. These local patterns have proved useful for the prediction of new interactions in both social (signed) [4] and biological (unsigned) networks [5, 6].

The goal of this project is to perform a thorough statistical analysis of the motifs present in some specific signed biological networks - genetic interaction networks, with the aim of gaining some insight on the principles behind its structural organization and to guide future link-prediction algorithms in signed biological networks.

## 1.1 Network concepts overview

A network consists on a set of vertices, $V$, and a set of edges, $E$, where each edge connects two vertices. Despite its disarming simplicity, it has been successfully used to describe a wide variety of biological systems, from the brain map to the food web [1–3, 7–9] . E.g. to describe the microscopic system of the protein-protein interactions within a cell, we associate each protein with a different vertex (or node), and link each interacting pair with an edge.

When the edges have no direction, but connect the nodes without distinguishing a source and a target node, the network is called undirected.

Furthermore, a network may have both positive and negative links, corresponding to positive and negative interactions between the elements of the system. In this case it is called a signed network, and each edge is represented as $e = (u, v, s)$ where $e \in E;\ u, v \in V$ and $s \in \{+, -\}$.

The structural organization of signed networks has been extensively studied, especially in social sciences. In particular it has been found that the Structural Balance (SB), a theory imported from

psychology [10] to network science [11], holds for social signed networks of highly distinct provenances [12–14]. Although originally formulated in terms of the possible attitudes of a person towards other persons and objects, SB can be put into four elucidative sentences:

1. A friend of a friend will be a friend.
2. An enemy of a friend will be an enemy.
3. A friend of an enemy will be an enemy.
4. An enemy of an enemy will be a friend.

However, because the $4^{th}$ sentence is not verified in many occasions, a weak formulation of SB was admitted, consisting on sentences 1–3. The version including 1–4 was called strong formulation. In summary, the strong formulation of SB states that only the configurations of signed triangles with a positive product of the edges are stable, whereas the weak formulation states that the configuration with one negative signal is unbalanced. As a side note, we draw the reader attention for the fact that SB is verified in undirected or highly reciprocal directed social networks, other theories exist to explain balance in highly directed networks [12]. Since the networks we will study are undirected, we will only test SB.

To check whether SB holds for some network we study the relative frequency of local patterns - motifs - in the network against a null model, figure 1.1 shows what to expect for a structurally balanced network.

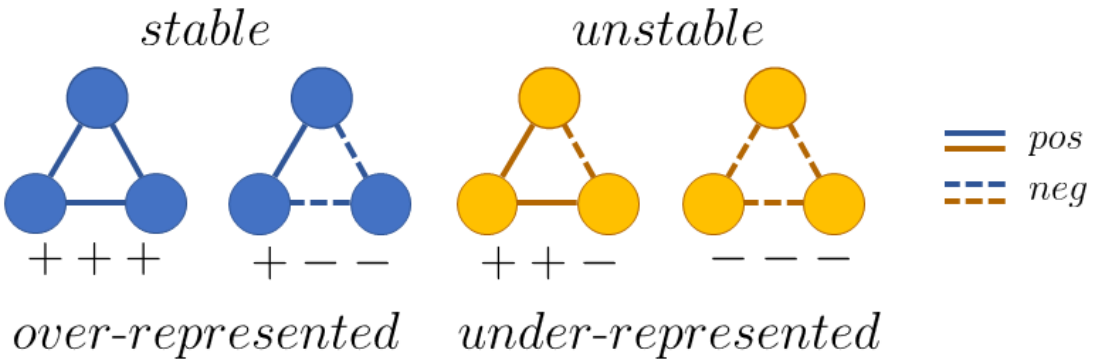



Figure 1.1: Triangles in Structural Balance

## 1.2 Motivation

Signed networks are of great interest in life sciences. In fact, biological function is believed to emerge from both positive and negative correlations [1–3]. In murine embryonic stem cells, it was shown that signed networks provided a better understanding of the regulatory mechanisms than unsigned networks. In particular, the study of a signed weighted gene co-expression network led to the discovery of a pluripotency module and a differentiation module, not detected by an unsigned network analysis [1].

Another recent study [2], mapped the role of genetic interactions in complex cellular function, from a signed genetic interaction network in yeast. Using network analysis tools, it was possible to predict the function of 1 essential and 6 non-essential genes, whose functions were previously unknown, and these predictions were experimentally validated. We will use these yeast genetic interaction datasets in our project.

Network-based strategies for drug repurposing or for finding drug combinations to target multiple parallel pathways are being developed, also exploring signed networks [3]. And the examples could continue...

Although many research communities are perceiving more and more the value of studying biological systems from a Network Science perspective, there is still a gap in the knowledge of the structural principles underlying the networks organization, specially for undirected signed networks!

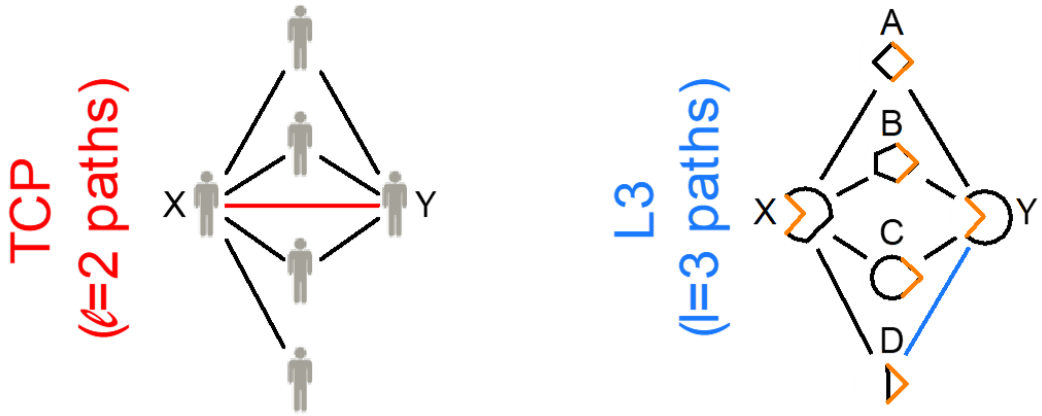
Despite recent systematic mapping efforts, either due to the disproportion between the size of the universe of potential pairwise interactions and the limited experimental resources available, or due to the fundamental technical limitations to probe the interactions, many biological datasets are still far from being complete. A link prediction algorithm could guide the experiments and reduce costs in the first case, and provide accurate estimates in the latter. The biologists from the Boone Lab, University of Toronto, working with genetic interaction networks, are very interested in such a tool.

A successful link prediction algorithm in social networks, based on the assumption that the higher the number of common friends two people have, the higher the probability that they know each other, is the Triadic Closure Principle (TCP) [4]. TCP states that the higher the fraction of common neighbors between two nodes, or alternatively, the higher the number of triangles closed by linking the two nodes, the higher the likelihood of a connection between both, see figure 1.2(a). This principle can be adopted in cell biology, under the rationale that two elements with a great number of common neighbors are likely to take part in the same functional pathways and thus are likely to interact.

Nevertheless, if we consider that an interaction of two proteins is mediated by complementing active sites, a different intuition arises: the likelihood of interaction of two proteins should be proportional to the number of paths of length three connecting both, see figure 1.2(b). Regarding the human protein-protein interaction network (undirected and unsigned), it was shown, by Kovács *et al* [5], that the latter intuition led to a link-prediction algorithm far more accurate than TCP. Muscoloni *et al* [6], revised Kovács' algorithm and showed that link-prediction scores based on the number of paths of length 3 closed by the candidate interaction outperformed TCP and its variations not only on other protein-protein interaction networks, but also on food webs and even world trade networks.

The above mentioned results were observed for undirected unsigned networks. In signed networks, to the prediction of link existence, one must add the sign prediction. So far we know that triangles follow strict rules on social signed networks, and closing squares leads to accurate link-prediction algorithms in biological unsigned networks.

Following the reasoning, we find interesting, not only to study the balance of triangles, as is common practice in social networks, but to include also the loops of length 4 in our analysis. On the one hand




(a) Social networks

(b) Interactome

Figure 1.2: Rationale behind link prediction algorithms in social (a) and biological (b) networks. Figure kindly provided by Kovács [5].

we hope to shine some light on the structural principles behind biological function, while on the other we intend to find out which motifs are important to the sign prediction problem.

## 1.3 Thesis Outline

In the following chapter, we will detail how the biological datasets containing the raw genetic interaction data were experimentally constructed, and explain how the data was processed and filtered in order to generate the genetic interaction networks upon which we are going to work. We also present a brief summary of some topological characteristics of the networks generated by the different filtering protocols.

In chapter 3, we discuss the necessity of a null model and present some possible methods to construct it, from the most naive to the more sophisticated model. We argue that this is a hard problem doing a tour by the different null models in the literature and explaining why they are not applicable to our case. Finally we present the null model we used.

Chapter 4 contains the obtained results. We compare the frequency of each signed triangle against the null model, check if SB is verified, and comment its biological interpretation. Similarly, we compare the frequency of each 4-node signed motif against the null model and check if there is a natural extension of the Structural Balance formulation to 4-node motifs.

We conclude with chapter 5, commenting the importance of the achievements of this thesis and pointing directions for future work.

# Chapter 2

# Biological Datasets

In this chapter we will discuss different methods to construct the biological datasets, then we explain the protocols used to generate the genetic interaction networks from the data, and study the characteristics of the networks obtained.

Biological networks are tools to encode information about biological systems and thus are tightly connected to the way the data is collected.

**Drug interaction networks**

For example, how does one build a drug-drug interaction network? A possible method to build such a network is to search the literature for all the individual pairwise interactions registered. Because this problem is obviously cumbersome, several efforts are being made to assemble large and comprehensive databases, publicly available, to serve the scientific community.

In 2002, a research group from Singapore, assembled the Therapeutic Target Database, merging information "about the known therapeutic protein and nucleic acid targets described in the literature, the targeted disease conditions, the pathway information and the corresponding drugs/ligands directed at each of these targets" [15, 16]. In 2006, a research group from Canada, assembled DrugBank combining "detailed drug (i.e. chemical) data with comprehensive drug target (i.e. protein) information", where "each drugcard entry contains $>$80 data fields" [17, 18]. In 2009, a research group from China, compiled the Drug Combination Database and event including unsuccessful combinations [19, 20]. And there are many other databases. The emergence of these networks is a great facilitator of network approach studies, due to the large amount of data concentrated.

However, these databases are not foolproof. On one hand the information is gathered from very disparate provenances, the drug combinations were assessed by different teams, from different labs in different countries, in different seasons, applying different methods and using different criteria to evaluate the interactions. A second systematic error is the fact that these drug combinations were curated in the literature or from other databases, leading to a great disproportion in the number of beneficial interactions vs adverse interactions, because the first are more prone to be published (negative results often stay in the drawer) and because some databases, such as the Orange Book from the Food and

Drug Administration of the USA [21], only accept beneficial drug combinations. The Orange Book is an information source of all the three databases mentioned above. The approval of the usage of drugs is a lengthy and expensive process and a systematic screening of beneficial and adverse drug interactions does not seem to be in the near future.

**Protein interaction networks**

There are also big protein-protein interaction databases, such as the STRING Database, developed by a research group from Germany [22, 23], in 2000. These databases are very precious, summarizing valuable results spread through the literature in a single platform. Nevertheless, they exhibit similar problems to the ones posed before for drug interaction databases.

However, unlike the previous drug interaction case, it is possible to perform large-scale screening of the protein interactions in a given organism. The yeast two-hybrid assay is a technique to identify protein-protein interactions, and it is scalable. Therefore, we have a method to construct a database in a systematic manner, avoiding many sources of incoherence and bias mentioned above.

Using the two-hybrid screening, Rual *et al.* [24], in 2005, estimate to have covered $\sim$20% of the total search space (complete set of possible pairwise interactions) of the human interactome. In 2008, Yu *et al.* [25], estimate to have screened $\sim$20% of the search space of the yeast interactome. In 2014, Rolland *et al.* [26], the technical advances allowed a screening of $\sim$42% of the entire search space.

These efforts to gradually cover more and more of the interactome, lead to the construction of high quality coherent datasets. A solid foundation for future network approaches to unveil the structural organization of the main agents of both intra and inter-cellular pathways - the proteins.

## 2.1 Synthetic genetic arrays

Synthetic genetic array (SGA) analysis is a high-throughput technique to quantitatively screen genetic interactions. As previously pointed, a systematic approach has the advantage to coherently test and evaluate all the interactions using the same methods and criteria.

Developed in 2001, by Tong (Boone Lab, U. Toronto) *et al* [27], the technique allows the crossing of a query gene deletion mutant with an entire gene deletion mutant array. SGA was first applied in the budding yeast, *Saccharomyces cerevisiae*, unicellular eukaryote organism. Genetic recombination, and a series of mating and selection steps, are used in order to obtain a double mutant array. The steps for SGA construction are automated using robots, and computer programs are used to determine the fitness of each colony, based on their growth.

When two mutations, neither lethal individually, combine to cause an unviable mutant, a synthetic lethal genetic interaction is said to occur. In this first assay, only non-essential genes were screened, i.e. genes for which single deletion mutants are viable. The experiment yielded an unsigned network with 291 genetic interactions, involving 204 non-essential genes.

In 2004, Tong *et al* [28], extend the technique to also probe essential genes, using deletion mutants for non-essential and conditional (partially functional) mutants for essential genes. With the technical

enhancement of the SGA, allowing to cross a query mutant array with a practically genome-wide deletion mutant array, a network of $\sim$4000 pairwise interactions concerning about 1000 yeast genes was constructed.

The technical tools to probe genetic interactions continued to develop. Temperature-sensitive mutant alleles allowed the mutation of essential genes, with the colonies being harvested at semi-permissive temperatures. In 2010, Costanzo *et al* [29], from the same group in U. Toronto, managed to construct a network with $\sim$170 000 interactions, involving 75% of the yeast genome, out of 5.4 million gene pairs examined.

In this paper, interactions were attributed a sign! A positive genetic interaction describes a double mutant exhibiting a fitness that is greater than expected, based on the combination of the two corresponding single mutants. Conversely, a negative interaction is identified when a double mutant displays a fitness defect that is more extreme than expected. The synthetic lethal interactions from the previous studies are now are included in the latter set. We call the reader attention to the fact that these positive/negative interactions were arbitrarily defined, the positive sign has no connotation of genetic cooperation, just as the negative interaction has no meaning of genetic repression whatsoever. Quite on the contrary, as we shall see when summarizing the characteristics of these genetic networks.

Finally, in 2016, after screening 23 million pairwise interactions, Costanzo *et al* [2], constructed a signed genetic network encompassing $\sim$900 000 genetic interactions, covering about 90% of the yeast's genes. We will study networks constructed upon the data generated in this last assay, which is publicly available online [30].

## 2.2 Filtering Protocols

To access the data, we downloaded the files under the title "Raw genetic interaction datasets: Pairwise interaction format", from [30]. In the .zip file, there is a list of all query and array mutant strains represented in the genetic interaction network along with their corresponding fitness estimates and 3 .txt files with the information for each pairwise interaction, one for the pairs of non-essential genes (NxN), other for the essential double mutants (ExE), and a third with the crossing of the essential vs non-essential genes (ExN). The header and first lines of the ExE file with the results from the SGA analysis are given in figure 2.1.

```
1 Query Strain ID Query allele    Array Strain ID     Array allele    Arraytype
2 YAL001C_tsq508  tfc3-g349e      YBL023C_tsa111      mcm2-1          TSA30
3 YAL001C_tsq508  tfc3-g349e      YBL026W_tsa1065     lsm2-5001       TSA30
4 YAL001C_tsq508  tfc3-g349e      YBL034C_tsa274      stu1-5          TSA30
5 YAL001C_tsq508  tfc3-g349e      YBL034C_tsa454      stu1-8          TSA30
```

```
1 Interaction score   P-value     Query SMF   Array SMF   DMF         DMF std
2 -0.0348             5.042e-03   0.8285      0.9254      0.7319      0.0102
3 -0.3529             3.591e-06   0.8285      0.9408      0.4266      0.0790
4 0.0126              4.625e-01   0.8285      0.8925      0.7520      0.1338
5 0.0043              4.998e-01   0.8285      0.7988      0.6661      0.0831
```

Figure 2.1: Partials of the data file

The data is organized in 11 columns: query strain ID, with the gene name and its position in the query array; query allele name, for some genes, several mutants were constructed, mutating different alleles of the gene (lines 4 and 5 of fig. 2.1), correspond to two alleles of the same array gene); array strain ID, with the gene name and its position in the array; array allele name; array type and temperature, the array can either be a deletion mutant array (for non-essential genes) or a temperature-sensitive array (mostly for essential genes, as in the fig. 2.1 case), the temperature can be 26°C or 30°C; genetic interaction score, $\epsilon$, computed as eq. 2.1; p-value, $P$, measure of the statistical significance of the individual interaction; query single mutant fitness, $f_i$, a quantity proportional to the colony growth; array single mutant fitness, $f_j$; double mutant fitness, $f_{ij}$; and standard deviation of the double mutant fitness.

$$\epsilon = f_{ij} - f_i f_j \tag{2.1}$$

But how does one draw a network out of this information overload? In fact, it is not a trivial process, not all tested gene pairs should give way to a genetic interaction. The data must be filtered first.

Costanzo *et al*, in their supplementary material [2], suggest three possible thresholds: a stringent confidence, which accepts interaction with $P < 0.05$ and $\epsilon > 0.16$ or $\epsilon < -0.12$, yielding a reduced number of false positives at the cost of augmented false negatives; an intermediate confidence accepting interactions with $P < 0.05$ and $|\epsilon| > 0.08$, reaching a compromise between false positives and false negatives; and a lenient threshold which accepts all the interactions for which $P < 0.05$, this results in a great number of false positives with a very few false negatives. Applying this to 2.1, information from lines 2 and 3 would be coded in a network, using lenient cutoff, but intermediate or stringent cutoffs would only accept line 3. In their papers, the group from Toronto always uses the intermediate threshold to filter the data.

We will now detail the steps of our path to construct a genetic interaction network.

**Protocol _31**

The protocol to generate a network from the raw data, result of the SGA analysis, is not completely detailed in the 80-pages supplement. For example, some gene pairs were tested several times, in fig. 2.1 lines 4 and 5 correspond to the same pair of genes, with the array gene mutated in different alleles. Moreover, some pairs of alleles were tested repeatedly for different array types and temperatures, to calibrate the mutant fitness scores. For a few cases, these redundant experiments output both positive and negative interaction scores!

In method _31 we iterate the lines of the data file. If the interaction is not accepted, either due to $\epsilon$ or $P$-value, skip to the next line. In case it is accepted, associate the interaction sign, $sgn(\epsilon)$, with the genes pair, for the genes name take the string before the '_' character in columns 1 and 3. In the end, go through the list of gene pairs and randomly select one of the possible signals for the interaction, obtaining a signed edge list, i.e. a signed network. We note that gene pairs with different possible signs are rare, and thus the variation in the networks produced by the protocol, from the same data file, is really low.

**Protocol _41**

A different variation of the protocol is to map all the alleles to the corresponding genes. Then, for each gene, randomly choose one allele and disregard the information on all the other alleles of the same gene. Finally, iterate the lines of the data file for the pairs of chosen alleles and register the signs of the accepted interactions.

This method obviously reduces the number of edges in the output network, specially in the essential case, where more redundant mutants were constructed.

**Protocol _7**

In September 2018, István Kovács spent some time in Toronto, with the Boone Lab. Among other things, they discussed the details of the filtering protocols. From the fruits of the discussion, we implemented a new protocol.

First, filter the data file, keeping only the accepted interactions. Then, map all the alleles to the respective query and array genes. If an allele appears on both query and array sides, choose that and disregard all the other alleles of the gene. If there is no overlap between the query and array sides, choose one for each side and consider only that allele for the given side. If a pairwise interaction still present entries with opposite sign, ignore the interaction. Convert the remaining interactions to a signed edge list.

## 2.3 Network characteristics

The team from Boone Lab, U. Toronto, has many years of studying these essential and non-essential genetic interaction networks in *S. cerevisiae*, thus we will summarize some of their findings about the networks structures which will reveal useful when interpreting our projects results on the most and less represented motifs.

The genetic interaction networks were examined using SAFE, spatial analysis of functional enrichment, a systematic method to annotate genetic networks and facilitate their functional analysis [31]. SAFE associates each gene with its respective Gene Ontology terms. The Gene Ontology Consortium has the aim to standardize the vocabulary used by the scientific community to describe genes and proteins in eukaryotic organisms [32]. With this purpose, three independent ontology database were constructed: molecular function, classifying genes according to the complexes/pathways where their products actuate (such as $H^+$-transport or Golgi transport complex); biological processes, indexing genes to the large processes involving their products (such as mRNA processing or cytokinesis); and cellular compartments, classifying genes according to the compartments where their products are active (such as nucleus or peroxisome).

To visualize the networks in 2-D, distances between genes were computed, using similarity measures between the sets of genetic interactions of each gene. It was found that, thresholding the genetic similarity network for high correlations (only visualize genes with at least one similarity coefficient greater than

the threshold), yielded several isolated clusters enriched for genes classified in the same complexes or pathways. When thresholding the network for intermediate correlation coefficients, clusters enriched for the same bioprocesses arose. Finally, thresholding the similarity network for low correlations lead to the formation of clusters, enriched for genes actuating on the same cell compartments.

**Negative interactions connect functionally related genes**

The network analysis revealed that negative genetic interactions tend to connect functionally related genes. In fact, 50% of the essential gene pairs whose products physically interact are linked by a negative interaction. Similarly, 63% of gene pairs annotated to the same essential protein complex are also bound by a negative genetic interaction.

Another interesting observation is that "essential genes that fall into a cluster within the set that was enriched for complexes/pathways were connected by a negative interaction with a relatively high density (60 to 90%), but they were rarely connected by a positive interaction" [2]. Which is an indication that negative interactions do not necessarily imply genetic repression.

Finally, there is a direct correlation between the magnitude of the negative interaction and the closeness of the genes' functional role, in both essential and non-essential networks.

**Positive interactions map general regulatory connections**

No correlation between positive interactions and other molecular or functional relationships, including physical interactions, was observed. "Thus, while negative interactions identified clear functional relationships between genes, positive interactions among partial loss-of function alleles of essential genes represent a different type of relationship that is not captured by other large-scale data sets or functional standards" [2].

In contrast with the observed for negative interactions, "the majority of positive interacting gene pairs in both the essential (ExE, 78%) and nonessential (NxN, 75%) genetic interaction networks occurred between distantly connected genes whose products appeared to function in different cell compartments" [2], and thus there seems to be no connection between interaction density and functional role. About 30% of the positive interactions related to the protein-degradation complexes represent genetic suppression, indicating that positive interactions do not necessarily mean genetic synergy.

Although less clear than in the negative interactions case, positive interactions seem associated with defects in cell cycle progression or cellular proteostasis and appear to map general regulatory connections.

**Global network features**

The three protocols described above output different networks. In table 2.1, we present the number of nodes, $N$, the number of positive, $L_+$, and negative edges, $L_-$, the total number of edges, $L$, and the network average degree (number of links in a node), $\langle k \rangle$, for the essential (ExE) and non-essential (NxN) networks, filtered with each of the protocols, with intermediate thresholding.

In protocol _31, for each pair of genes, an edge is formed if at least one of the multiple tested

Table 2.1: Network features for the different protocols

| | $N$ | $L_+$ | $L_-$ | $L$ | $\langle k \rangle$ |
|---|---|---|---|---|---|
| ExE_31 | 855 | 33,238 | 46,569 | 79,807 | 187 |
| ExE_7 | 855 | 21,861 | 33,053 | 54,914 | 128 |
| ExE_41 | 855 | 21,103 | 31,606 | 52,709 | 123 |
| NxN_31 | 4688 | 153,795 | 230,774 | 384,569 | 164 |
| NxN_7 | 4688 | 152,341 | 229,092 | 381,433 | 163 |
| NxN_41 | 4688 | 149,079 | 224,790 | 373,869 | 160 |

combination between mutant alleles of the genetic pair is accepted. In protocol _7, a single mutated allele per gene is considered, but the alleles are chosen from a pool of alleles with, at least two accepted interactions (unless the gene was only present on the query or array side, in that case the allele chosen has at least one accepted interaction). In protocol _41, the first step is to choose a single mutated allele per gene, without any additional information. Thus, the first protocol yields the more densely connected networks, and the others follow in a descending order. Because there are more redundancies in the ExE screening, the differences introduced by the protocols in the network average degree are more disparate.

Similarly, we present the variation of the same global features with the cutoff in the genetic interaction scores in table 2.2. The number of edges obviously decreases with the confidence required to accept the interaction. The magnitude of the positive and negative interactions is not similarly distributed and the number of positive edges suffer a higher decrease, in proportion, each time the confidence is stricted.

Table 2.2: Network features for the different thresholdings

| | $N$ | $L_+$ | $L_-$ | $L$ | $\langle k \rangle$ |
|---|---|---|---|---|---|
| ExE s | 853 | 3,676 | 23,568 | 27,244 | 64 |
| ExE i | 855 | 21,861 | 33,053 | 54,914 | 128 |
| ExE l | 855 | 48,678 | 50,763 | 99,441 | 233 |
| NxN s | 4673 | 19,549 | 128,456 | 148,005 | 63 |
| NxN i | 4688 | 152,341 | 229,092 | 381,433 | 163 |
| NxN l | 4688 | 689,943 | 726,023 | 1,415,966 | 604 |

# Chapter 3

# Randomization models

In order to predict new links on a network, prior knowledge on the main principles underlying the organization of the biological entities in the network is required. More specifically, in this project, we have the aim to study which structural patterns are statistically more frequent in the networks. However, to make a judgment on the statistical relevance of a motif, we need a null model to which we can compare the network to assess.

The common practice in network science is to generate an ensemble of networks preserving some property of the original one (e.g. keeping the degree distribution), and then compare the characteristic in study (e.g. frequency of a specific motif) in the original network against the ensemble. If the characteristic of the original network is not replicated in the ensemble, then it must be consequence of some aspect other than the preserved structural property.

The Z score, i.e. the number of standard deviations a value is from the mean, is the measure commonly used to compare the property value in the original network and in the random network population. The Z score is computed as in (3.1), where $X$ is the property value, $\mu$ is the mean value of the property in the population and $\sigma$ is its standard deviation.

$$Z = \frac{X - \mu}{\sigma} \tag{3.1}$$

Choosing the appropriate network randomization model, to use as the null model in our statistical study, is not a trivial problem. In the present chapter, in section 3.1, we introduce the naive approach to randomize a network and explain why it does not lead to a meaningful null model. In section 3.2, we expose a commonly used model that preserves some structural properties of the network in study, overcoming the main limitation of the naive approach, but needs to be adapted to signed networks. In the following section 3.3, we adapt the model to create synthetic signed networks preserving the fundamental structural properties required, but still with a flaw that makes it unacceptable for the genetic interaction network case. Next, in section 3.4 we go through the literature, quickly explaining several of the methods used and why they fail to suit our problem. Finally, in section 3.5, we present the model we chose to generate the random ensemble.

## 3.1 Random Network model

The first naive approach to generate a random network, proposed in 1959 by Erdős and Rényi [33], is to fix the number of nodes, $N$, and the number of links, $L$, of the input network and randomly distribute the number of links between the nodes. A second similar approach is to connect each pair of nodes with some constant probability $p$, to produce a synthetic network with the same number of nodes, $N$ and, on average, the same number of links. This approach was proposed by Gilbert in the same year [34].

However, the random networks generated by the aforementioned models fail to mimic the networks describing real systems regarding connectedness, average path length, clustering coefficients and degree distribution [35]. The degree of a node is the number of edges connected to it, and the degree distribution $p_k$ is the probability of a randomly selected node having degree $k$. In the previous models, the degree distributions of the random networks follow a binomial distribution and, for a large number of nodes $N >> \langle k \rangle$ it is well approximated by a Poisson distribution (3.2). Nevertheless, many real networks have their degree distribution following a power law distribution (3.3).

$$p_k = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle} \tag{3.2}$$

$$p_k = Ck^{-\gamma} \tag{3.3}$$

The conformations of networks with these degree distributions are very different, e.g. in a network following a Poisson distribution most nodes lay in the narrow vicinity of $\langle k \rangle$, whereas in a power law network there is a large number of small degree nodes as well as some highly connected nodes.

## 3.2 Random Rewire model

In sequence with the previous section, we present a slightly more elaborated model, the Random Rewire, to generate networks with the same degree distribution as the input network.

The central idea is that performing an *edge swap* changes the network without changing its degree distribution. An edge swap consists on selecting two edges, $(u, v)$ and $(x, y)$, and attempting a swap to $(u, x)$ and $(v, y)$, as pictured in fig. 3.1. The attempts fails if one of the edges $(u, x)$ or $(v, y)$ already exists, or if $u = x$ or $v = y$, to avoid self-loops. Note that, after the *edge swap*, each nodes preserves its degree.

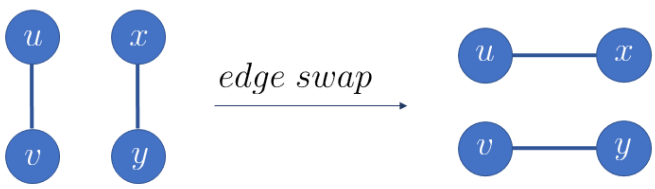



Figure 3.1: An edge swap

To implement the Random Rewire model in this project we used the *generation.random_rewire()* function, from the *Python* library *graph-tool* [36]. The function has an input parameter *n_iter*, specifying the number of sweeps over the edge list, $E$, where each edge is visited once in a random order and an edge swap is attempted.

We now have a model capable of producing a population of networks with the same degree distribution as any given network. This model was constructed thinking on unsigned networks and a naive generalization to signed networks is to swap $e = (u, v, s_1)$ and $f = (x, y, s_2)$ into $e' = (u, x, s_1)$ and $f' = (v, y, s_2)$. Nonetheless, although it preserves the global number of $+$ and $-$ edges in the network, the generalization disrupts the signed-degree distributions (repicture fig. 3.1 with a positive and a negative edges).

The signed-degrees, i.e. the number of positive and negative edges connecting a vertex, are of vital importance to the functional characterization of nodes in signed biological networks [1, 37]. Thus, to generate random networks with significant biological meaning, the signed-degrees are a relevant feature to preserve.

## 3.3 Signed Rewire model

To achieve our goal to randomize a network, maintaining its signed-degree distributions, we can still build upon the previous model. Let us consider $E_+$ and $E_-$, the positive and negative subsets of the network edges set $E$, defined as:

$$\begin{aligned} E_+ &= \{e \in E \colon e = (x, y, +)\} \\ E_- &= \{e \in E \colon e = (x, y, -)\} \end{aligned} \tag{3.4}$$

Now, let us apply the Random Rewire to each subset, obtaining new randomized subsets, $E'_+$ and $E'_-$, with the same positive and negative-degree distributions, respectively, as the original set $E$. We can now merge the two randomized subsets, obtaining a new set of edges, $E' = E'_+ \cup E'_-$, with the same signed-degree distributions as $E$.

In the figure 3.2, we present a simple example of a Signed Rewire randomization. Note that the nodes in the randomized version of the network keep the exact same number of positive and negative edges, as intended. However, it is also possible to understand that the Signed Rewire output edge list may include parallel edges, even if there were none in the input list.

This maybe a problem in the case where parallel edges have no biological meaning, like the genetic interactions case. Nonetheless the method is good for networks allowing parallel edges, as the drug interaction case. We will apply this model, in a near future, to study the motifs in signed drug-drug interactions in humans. However, we cannot accept parallel edges in the null model to which compare our current networks, since parallel edges are not present in the genetic interactions context.

Therefore we looked into the literature searching for a method to synthesize undirected signed networks from given positive and negative-degrees, and with no parallel edges.

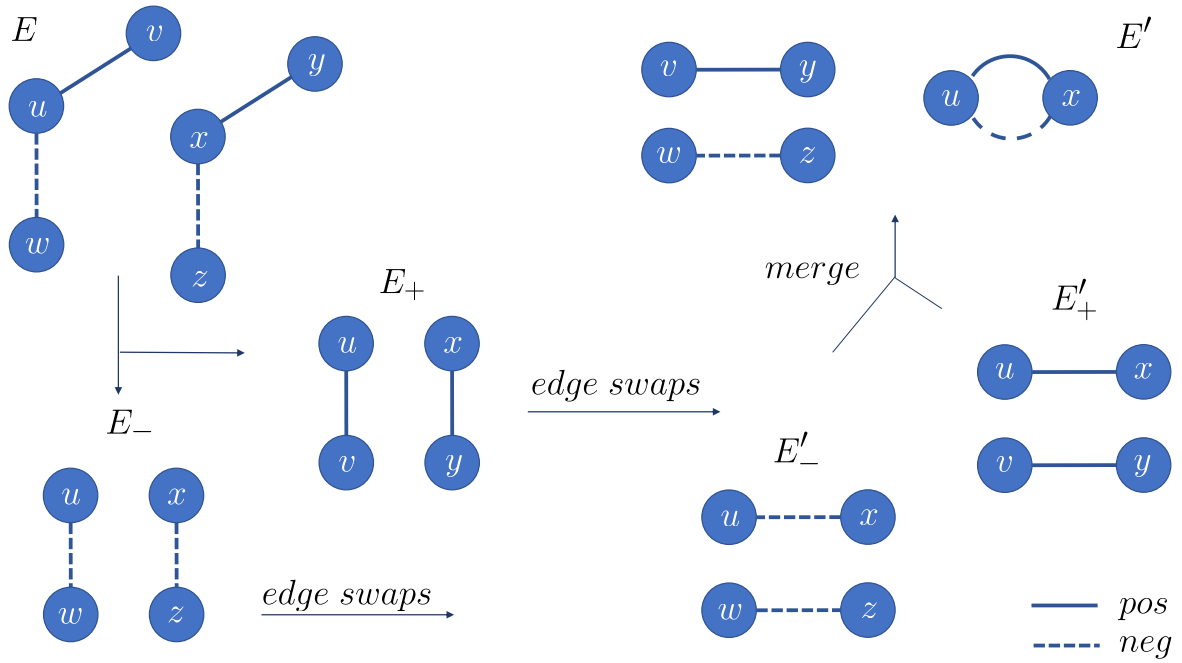



Figure 3.2: Sign Rewire ilustration

## 3.4 Literature Review

There are many papers in the literature discussing efficient and meaningful network randomization models. However, to the best of our knowledge, there is none focusing on undirected signed networks!

**Randomization of unsigned networks**

As studied in section 3.1, an unconstrained model to generate random networks fails to mimic structural properties of real networks, responsible for many topological features of the network. An algorithm to generate networks with an exact degree sequence has been proposed, in 1978, by Bender and Canfield [38]. Their model consist in "cutting" all the edges in the network, obtaining an even number of half-edges, and preserving the number of links for each node, and then randomly pair the half-edges, to obtain a new network with the same degree distribution. However, if one wants to reject parallel edges and self-loops, but assuring the model is homogeneously sampling the search space (i.e. all the possible networks with the given degree distribution are generated with equal probability), the method becomes very slow and inefficient, specially for highly connected networks. This is, in general, the problem with highly constrained methods, they are either slow and inefficient, or constrain too much the degrees of freedom and do not sample all the search space (the output networks are very similar to the input).

In 2002, to overcome the excessive constraints hindering the design of a fast and unbiased model, Chung and Lu [39, 40], developed an algorithm to generate random networks lifting the restriction of an exact degree distribution. Their model generates a network with a given expected degree sequence $(k_1, k_2..., k_n)$ by independently assigning an edge to each pair of vertices $(i, j)$ with probability $p_{ij}$ proportional to the product of the nodes' degrees $k_i k_j$. The naive approach by Gilbert [34], explained in section 3.1, is a particular case of this model, where the expected degree sequence is constant. The

Chung and Lu model is a popular model for the randomization of unsigned networks, however, to the best of our knowledge, there is no extension to signed networks avoiding parallel edges. Other methods have been proposed to generate networks from a given expected degree distribution [41, 42], using different normalizations or combinatorials to compute the edge probability $p_{ij}$, but again, there seems to be no trivial extension to none of them to serve our purposes for the randomization of signed networks.

**Random models in social signed networks**

So, in our literature review, we moved from articles on network randomization to the null models used in papers studying the Structural Balance and link prediction in signed social networks. Below we list and comment some of those randomization models.

In 2010, Szell *et al* [13], found that SB was verified in friendship-enmity networks derived from online games. In this article they "define a null model by keeping the topology fixed and by randomly assigning the $L_+$ plus-signs and $L_-$ minus-signs on the existing links, where $L_+(L_-)$ are the original numbers of friendship (enmity) links". With this method, the network framework is preserved, as is the fraction of positive and negative signs. However, the signed-degrees are disrupted and thus this null model does not suit our problem.

Also in 2010, Leskovec *et al* [12], studied the Balance in three large social networks, such as the one formed by up and downvotes for Wikipedia admin candidates. When treating their data without direction, they "shuffle the signs of all edges in the graph, keeping the fraction of positive edges the same". Which is equal to Szell's method.

In 2011, Facchetti *et al* [14], adopted an algorithm for the calculation of ground-state in large-scale Ising spin glasses to compute global levels of Balance in the same three social networks. For the null model, the framework of the network was preserved and each edge signal was considered an independent random variable with Bernoulli distribution with probability $p = L_+/(L_+ + L_-)$ of being positive. This method generates an ensemble of networks with, on average, the same number of positive and negatives as the original network and avoids some biases present in the methods used in the aforementioned articles. Nevertheless, and once more, our goal remains unachieved.

When testing their link-prediction algorithm on social directed networks, Leskovec *et al* [43], used an extra null model, where they "generate a random network where each node maintains the same number of incoming and outgoing positive and negative edges". This latter model preserves the signed-degree distribution and moreover the incoming and outgoing degrees. However and astonishingly, there is not a trivial extension of the algorithm for undirected networks.

**Randomization of directed networks**

As Iorio *et al* [37] put it, in their article on the efficient randomization of biological networks, "the problem of randomizing an undirected and unweighted network while preserving the degree of its nodes (...) unfortunately presents itself with analytical and numerical challenges. With the additional constrain that the network to rewire is bipartite (...) this problem reduces to randomizing a binary matrix", a problem to which the solution is known [44].

Although, at a first look, it may seem that undirected networks should be easier to randomize since they do not distinguish source and target nodes in an edge, and thus are less constrained. It is precisely the fact that directed networks can be partitioned in two sub-sets (source nodes and target nodes), such that there are no edges linking vertices within the same set, that makes the signed-degree preserving rewire feasible!

## 3.5 Alpha Randomization

During our quest to find the appropriate randomization model, we kept in dialogue with István Kovács, from the Northeastern University. At some point he noticed that it was possible to adapt an ongoing project of his, on the randomization of subgraphs, to fit our purpose and kindly provided us with his code.

His method, still unpublished, treats the negative signs as a subgraph of all existing links. Then, a maximum entropy model of this subgraph is built, using the known subgraph degrees and taking all non-existing links as hard constraints. The entropy model is used to compute some hidden parameters for each node, in an iterative manner, until reasonable convergence is observed. Finally, the parameters are used to calculate a probability of existence for each unconstrained link and generate a random subgraph. All the edges in the subgraph are assigned a negative sign and the remaining links, from the original edge list, are assigned a positive sign.

In the end, we have an ensemble of randomized networks, with the same structural framework as the input network, just as the models for undirected signed networks visited in the literature review [12–14]. Thus, the absolute number of triangles and squares in the randomized network is constant, because the framework is fixed. However, István's new method has the plus to preserve, on average, the signed-degree distributions.

To empirically validate the model, we compared the signed-degree distributions of the original networks with the average distributions of the 100 network ensemble generated and verified the efficacy of the method. In fig. 3.3 we plot the log-log (left) and linear (right) histograms obtained using ExE_31 as input.

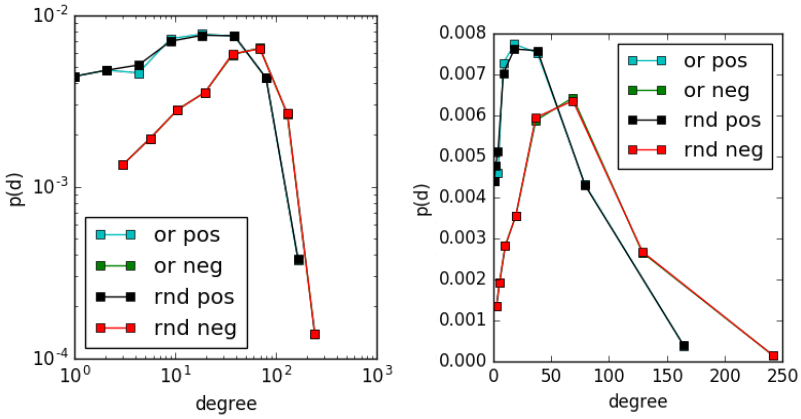



Figure 3.3: Positive and negative degree distributions of ExE

# Chapter 4

# Results

In section 1.2 we have motivated the statistical analysis of the 3 and 4-node motifs present in the genetic interaction networks, to further our understanding of the principles underlying its structural organization and to drive new sign prediction algorithms.

In chapter 2, we have discussed the different protocols to filter the raw data in order to obtain meaningful interaction networks. In the present chapter we will study how the protocols influence the final results.

In chapter 3, we explained the need for a random population and narrated our course in pursuit of *the* randomization model. All the results presented in this chapter used the randomization model detailed in 3.5 to generate the random populations with 100 synthetic networks. The measure used to evaluate each signed motif is the Z score (eq. 3.1). If the frequency of a motif in the input network is more than 2 standard deviations above the population average, we consider that motif to be over-represented in the network, and highlight its Z score in green. Similarly, if the frequency of a motif in the input network is more than 2 standard deviations below the population average, we consider that motif to be under-represented in the network, and highlight its Z score in red. Z scores between -2 and 2 are considered as *no signal*, and highlighted in gray. These motifs, neither over- nor under-represented, are said to be balanced.

When counting 3-node patterns, we considered the 4 standard signed triangles (fig. 4.1, top row). For the 4-node patterns we distinguished 6 different motifs: besides considering each possible combination of 4 signals, we also distinguish the square with two positive and two negative signs in the case where there is a node connected to two '+' edges, and the case where each node has a '+' and a '$-$' edge (fig. 4.1, bottom row).

## 4.1 Essential genetic interactions network

From the experimental data available in [30] we extracted the essential genetic interaction network (ExE_7), using the protocol _7 described in section 2.2, with intermediate filtering.

Using ExE_7 as the input network for the Alpha Randomization algorithm (section 3.5), we produced

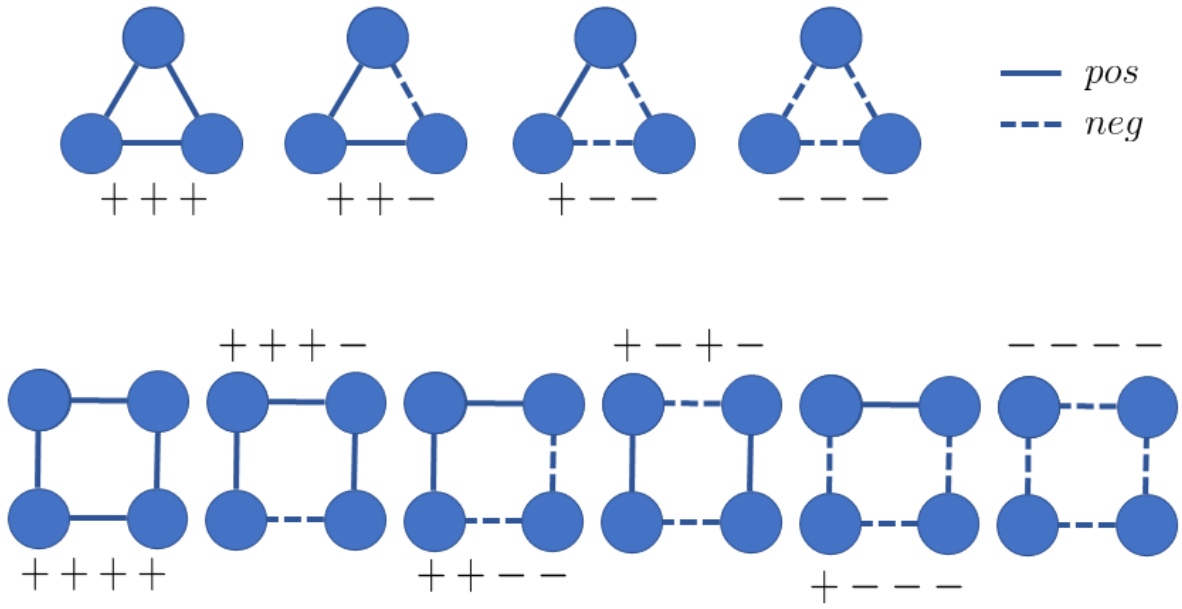



Figure 4.1: 3 and 4-node patterns

a random population of 100 networks with the same framework as ExE_7 and preserving, on average, the same signed-degrees distribution. Then we computed the frequencies of each signed 3 and 4-node motifs in the population and calculated their average, $\mu$, and standard deviation, $\sigma$. Finally we compared the frequencies of the motifs in the essential genetic interaction network, obtaining tables 4.1 and 4.2.

To better interpret the Z score values, we may recall that 60 to 90% of the essential genes involved in the same complex or pathway were connected by a negative interaction, while 78% of the positive interactions connected genes whose products acted in different cell compartments. In fewer words, negative interactions tend to connect functionally related genes, while positive interactions reflect general regulatory mechanisms [2].

### 4.1.1 3-node motif analysis

Table 4.1: 3-node motifs in ExE network

| | + + + | + + − | + − − | − − − |
|---|---|---|---|---|
| ExE_7 | 57,410 | 249,638 | 286,371 | 191,000 |
| Alpha $\mu$ | 63,949 | 236,999 | 322,325 | 161,146 |
| Alpha $\sigma$ | 1,157 | 1,703 | 982 | 1,972 |
| Z score | -5.6 | 7.4 | -37 | 15 |

In a first glance at table 4.1, the low Z-score of '$+ - -$' stands out. The number of these triangles in the ExE_7 network is more than 30 standard deviations below the random population average, indicating that the motif is highly unstable in this network. In fact, if we associate a negative interaction with functional relatedness and a positive interaction with different cellular compartments, and consider a '$+ - -$' motif, our previous associations will conflict: the 3 genes either participate in closely related functions, or act in distinct cellular compartments (figure 4.2, right).

There is a prominent presence of '$- - -$' triangles in the genetic network, 15 standard deviations

above the random population with the same negative-degree distribution. If this fact is not consequence of the network characteristics (degree distribution), then it might mean that there is some structural advantage associated with the motif, and that it is beneficial to have a high density of negative interactions between essential genes associated with the same bioprocesses.

The high Z score of the '+ + −' also indicates this as a structurally preferred motif and reflects a previous finding: members of the same essential protein complex often show an identical positive interaction profile, that is, genes playing a role in the same pathways tend to be regulated by a similar set of genes (figure 4.2, left).

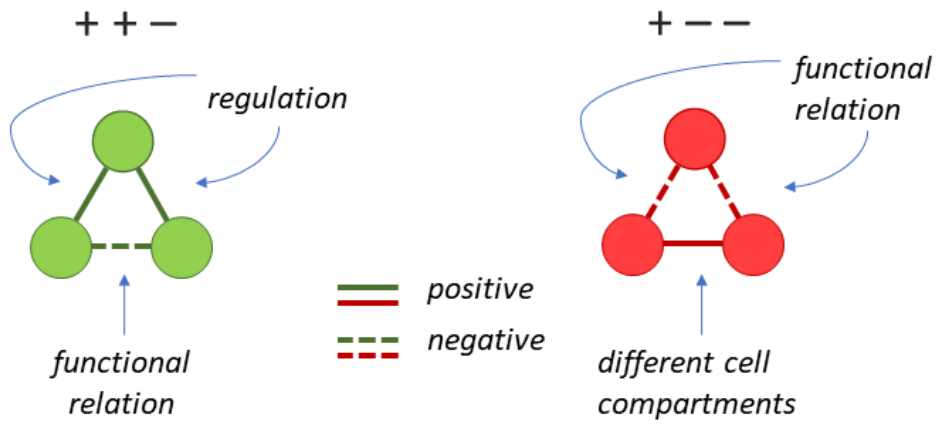



Figure 4.2: Biological interpretation of the over-represented '+ + −' and under-represented '+ − −' motifs.

### 4.1.2 4-node motif analysis

Table 4.2: 4-node motifs in ExE network

| | + + + + | + + + − | + + − − | + − + − | + − − − | − − − − |
|---|---|---|---|---|---|---|
| ExE_7 | 3,499,656 | 16,321,927 | 24,146,243 | 11,023,454 | 30,242,237 | 12,528,317 |
| Alpha $\mu$ | 3,496,635 | 17,194,465 | 23,342,812 | 10,496,253 | 31,488,630 | 11,743,039 |
| Alpha $\sigma$ | 82,843 | 215,056 | 71,037 | 51,187 | 207,553 | 182,670 |
| Z score | 0.04 | -4.1 | 11 | 10 | -6.0 | 4.3 |

The sub-cellular meaning of the genetic interactions is not totally clear, and the number of interactions per pattern has just grown, increasing the uncertainty on the interpretation of each motif. To interpret the Z scores obtained for the different signed squares, we will henceforth abstract from the biological analysis of the patterns and adopt a network science perspective.

Looking at table 4.2, it is interesting to note that motifs with an odd number of '+' and '−' signs are under-represented in the network, supporting the idea that motifs with a negative product of its edges are not stable in real networks.

There are three motifs that appear to be stable and preferred in the structural organization of the genetic network: '+ + − −', '+ − + −' and '− − − −'. Once again, they all have in common the positive product of its edges, making way for an extension of the Structural Balance formulation for squares.

The '+ + + +' case its curious: we know from Costanzo [2] that the positive essential genetic interaction network has a very low density, and we have also seen that the configuration '+ + +' is penalized by the network organization, nevertheless the '+ + + +' square, with a positive edge product, is not under-represented in the network but has a count in the line of the random population.

### 4.1.3 Reversing the signals on Structural Balance

So far it appears that the distribution of the motifs of 3 and 4 nodes in the essential genetic interaction network obeys some rules. However, when we compare the Z scores obtained for the 3-node motifs, with what is expected by the strong formulation of SB, we see that these are inverted (table 4.3). In fact, the motifs predicted to be over-represented by SB ('+ + +' and '+ − −') are under-represented in the network, whereas the unstable motifs according to SB ('+ + −' and '− − −') are the more prevalent! This suggests that the characteristics we attribute to the '+' and '−' signs in social networks might be reversely applied to the genetic '+' and '−' interactions.

In social networks, "positive ties are more likely to be clumped together, while negative ties tend to act more like bridges between islands of positive ties" [12]. In the genetic networks, however, genes related to the same pathway form clusters with a very high density of negative interactions, whereas the majority of positive interactions connects genes from different clusters. In short, both friendship and negative genetic interaction are mainly an intra-cluster bond, while enmity and positive genetic interaction tend to be a connection inter-clusters.

Hence, to correctly define the rules of balance in genetic networks, for triangles, one must invert the strong formulation of Structural Balance. This is, to the best of our knowledge, a completely new result for both the Biology and Network Science communities!

Since we only have one essential genetic interaction network, we thought of obtaining more networks from the same raw data. To achieve this we used the different proposed cutoffs for the genetic interaction score (see section 2.2). The network obtained with a stringent threshold was named 'ExE_7 s', the network obtained using a lenient threshold was named 'ExE_7 l', and the one obtained using the intermediate threshold was renamed from 'ExE_7' to 'ExE_7 i'. In table 4.3, we present the Z scores obtained for the different thresholds and the expected outcome from the strong formulation of SB.

Table 4.3: Variation of triangle Balance with the thresholding

| Z score | + + + | + + − | + − − | − − − |
|---|---|---|---|---|
| ExE_7 s | -6.3 | 0.1 | -19 | 13 |
| ExE_7 i | -5.6 | 7.4 | -37 | 15 |
| ExE_7 l | -4.8 | 12 | -20 | 9.7 |
| SB strong | | | | |

Again we see that SB must be reversely formulated when applied to genetic networks. We note that the absence of signal in the Z score of '+ + −' is not contradicting the previously stated rule, for the stringent threshold the motif is not over- neither under-represented. We hypothesize that this might be due to the abrupt reduction of positive edges, from the intermediate to the stringent cutoff (decrease to

1/7 of the positive edges, while negatives were reduced to 70%, see table 2.2).

We see that the Balance in the essential genetic interaction network seems independent of the false positive and false negative rates. An important result since many biological networks arise from experiments where these rates are never null.

Table 4.4: Variation of triangle Balance with the filtering protocols

| Z score | + + + | + + − | + − − | − − − |
|---|---|---|---|---|
| ExE_31 | -5.1 | 6.7 | -19 | 8.6 |
| ExE_7 | -5.6 | 7.4 | -37 | 15 |
| ExE_41 | -5.7 | 7.8 | -32 | 16 |
| SB strong | | | | |

Finally, we also wanted to see if the newly found Balance results were consistent among the different filtering protocols discussed in section 2.1. With that intent, we randomized the different networks and computed the Z scores (table 4.4). We were eager to verify that the Balance is independent from the nuances of the filtering protocols.

### 4.1.4 Extending Structural Balance to squares

As mentioned in section 1.2, closing paths of length three is the state of the art for link prediction in biological networks [5, 6]. Therefore, it is of the most interest to find if there is some rule for the Balance in squares, so that we can add a sign to the links being predicted. With this goal, we computed the Z scores for the 4-node motifs in the networks filtered with the different thresholds for the genetic interaction scores, table 4.5.

Table 4.5: Variation of square Balance with the thresholding

| Z score | + + + + | + + + − | + + − − | + − + − | + − − − | − − − − |
|---|---|---|---|---|---|---|
| ExE_7 s | -0.4 | -5.0 | 0.1 | 4.6 | -12 | 5.4 |
| ExE_7 i | 0.04 | -4.1 | 11 | 10 | -6.0 | 4.3 |
| ExE_7 l | -0.1 | -2.5 | 106 | 13 | -3.7 | 1.3 |

Again, there seems to be some consistency on the signed squared motifs distribution along the different cutoffs. For the 4-node motifs case, due to the even number of edges in each loop, the product of the edges of the squares would not change and thus, there is no argument to reversely formulate the Balance rules.

It is clear that the patterns with a negative edge product, '+ + + −' and '+ − − −', are completely deprecated in the structural organization of the essential genes. An important fact to be considered when predicting new links using paths of length three in biological signed networks.

The overall tendency for 4-node motifs appears to be: patterns with a negative edge product are not stable, and thus under-represented; patterns with a positive edge product, apart from '+ + + +', are favored by the network structural organization and thus over-represented; there is nothing we can conclude about the '+ + + +' square.

We remind that the absence of signal in the Z scores for '+ + − −' in the stringent thresholding and for '− − − −' in the lenient, is not considered to be contradicting the empirical rule stated above. The

abnormally high Z value for '+ + − −' is due to an abnormally low value in the standard deviation of the motif, about 20 times smaller than the rest of the standard deviations.

To further test the consistency of the Balance observed, we asked how the scores were affected by the changes introduced in the network by the different filtering protocols, and repeated the procedure at intermediate filtering threshold obtaining table 4.6.

Table 4.6: Variation of square Balance with the filtering

| Z score | + + + + | + + + − | + + − − | + − + − | + − − − | − − − − |
|---|---|---|---|---|---|---|
| ExE_31 | -1.2 | -3.0 | 8.2 | 11 | -2.2 | 1.8 |
| ExE_7 | 0.04 | -4.1 | 11 | 10 | -6.0 | 4.3 |
| ExE_41 | -1.5 | -1.8 | -1.9 | -1.6 | 1.7 | 2.0 |

Before taking any inference about the rules of Balance, it is worth commenting the strange case of network ExE_41. None of the motifs in the network stands out from the values of the random ensemble. One could be lead to think that either the network was not successfully randomized or that the distribution of the motifs is explained by the signed-degree distribution. Against the first hypothesis arises the fact that, in table 4.4, we can see some properties in the structure of ExE_41 standing out from the average of the same ensemble, and consequently, the synthetic networks generated must have a significantly different structure than the original network. From the second hypothesis a question emerges: "If the signed-degree distributions explain the frequencies of the 4-node motifs in the network, why don't they also explain the 3-node values?" or, reversely, "If the signed triangle frequencies are not explained by the degree distributions, why are the squares explained?", so far we have not come up with a satisfactory answer... Could this be a point in favor of the usage of 3-node motifs in a future sign prediction algorithm, to the detriment of squares?

Although we present the result obtained we shall not take the last line of table 4.6 into account when empirically deriving an extension of SB to 4-node motifs in genetic networks. The first five motifs support the trend already verified before. However, the two extra voids of signal regarding the motif '− − − −' shake the idea that this pattern is favored by the network structural organization. Additional tests on other genetic interaction networks may shine more light on the balance of this particular motif.

## 4.2 Non-essential genetic interaction networks

After exploring the structural organization of the essential genetic interaction network, we also extracted a network with the interactions between pairs of non-essential genes of the yeast, using once more method _7, for each of the stringent, intermediate and lenient cutoffs on the genetic interaction scores.

### 4.2.1 3-node motif analysis

The procedure from section 4.1.1 was repeated, and the results for the intermediate threshold, considered to yield the network with higher biological meaning, are presented in table 4.7.

The 3-node motif distribution indicates that, in the non-essential genetic interaction network (NxN_7), the Balance may follow the reverse of the weak formulation of SB. Contrary to what is observed in ExE_7, now the '+ + +' triangle is not under-represented, but has no signal. As for the remaining motifs, '+ − −' is still the most unstable triangle, and the negative edge product triangles continue to appear as over-expressed in the original network.

Table 4.7: 3-node motifs in NxN network

| | + + + | + + − | + − − | − − − |
|---|---|---|---|---|
| NxN_7 | 248,809 | 1,034,004 | 1,535,057 | 895,411 |
| Alpha $\mu$ | 245,210 | 1,024,586 | 1,567,758 | 875,727 |
| Alpha $\sigma$ | 2,229 | 4,362 | 1,782 | 5,326 |
| Z score | 1.6 | 2.2 | -18 | 3.7 |

The (reversed) weak formulation of SB is consistently observed among the different filtering protocols, at intermediate threshold levels (table 4.9). However, the same does not apply when varying the cutoff value (table 4.8), the distribution of the triangles is completely different depending on the thresholding applied.

With respect to the triangles, it appears that there are some faint principles underlying the organization of the non-essential genetic interaction network. However, this structural organization vanishes with fluctuations of the false positive and false negative rates. Therefore, we consider that there is no evidence to support an empirical formulation for the 3-node Balance in the non-essential genetic interaction network.

Table 4.8: Variation of triangle Balance with the thresholding

| Z score | + + + | + + − | + − − | − − − |
|---|---|---|---|---|
| NxN_7 s | -0.8 | -4.2 | 2.8 | -0.1 |
| NxN_7 i | 1.6 | 2.2 | -18 | 3.7 |
| NxN_7 l | -2.7 | 4.7 | -4.5 | 1.3 |
| SB weak | | | | |

Table 4.9: Variation of triangle Balance with the filtering

| Z score | + + + | + + − | + − − | − − − |
|---|---|---|---|---|
| NxN_31 | 1.6 | 2.4 | -18 | 3.9 |
| NxN_7 | 1.6 | 2.2 | -18 | 3.7 |
| NxN_41 | 1.5 | 2.0 | -19 | 5.2 |
| SB weak | | | | |

### 4.2.2 4-node motif analysis

As done previously with the essential networks, after having studied the 3-node motifs, we looked into the squares expecting to find some type of Balance for the 4-node motifs.

Our fist impression, from the comparison of table 4.10 with the results previously obtained for the Balance of squares in ExE (section 4.1.4), is that the structure of the non-essential genetic interaction network (NxN) may be ruled by some principles other than the ones underlying the organization of the essential network.

Table 4.10: 4-node motifs in NxN

| | + + + + | + + + − | + + − − | + − + − | + − − − | − − − − |
|---|---|---|---|---|---|---|
| NxN_7 | 22,686,972 | 118,935,361 | 186,300,835 | 85,676,108 | 268,487,582 | 115,162,487 |
| Alpha $\mu$ | 21,561,989 | 119,289,124 | 181,574,119 | 82,703,089 | 276,574,788 | 115,546,236 |
| Alpha $\sigma$ | 256,790 | 838,778 | 398,302 | 227,644 | 773,710 | 908,546 |
| Z score | 4.4 | -0.4 | 12 | 13 | -10 | -0.4 |

But before comparing the differences and similarities of NxN and ExE, it is useful to look first at the variation of the motifs distribution with the cutoff on the interaction scores and with the different filtering protocols (tables 4.11 and 4.12).

Table 4.11: Variation of the square Balance with the thresholding

| Z score | + + + + | + + + − | + + − − | + − + − | + − − − | − − − − |
|---|---|---|---|---|---|---|
| NxN_7 s | 11 | -6.9 | 2.9 | 10 | -5.4 | 0.8 |
| NxN_7 i | 4.4 | -0.4 | 12 | 13 | -10 | -0.4 |

We must keep in mind that NxN networks are considerably larger than the ExE networks, and NxN_0, with a higher rate of false positives, is quite dense ($\sim$1.5 million edges). To count the frequencies of the square motifs in 100 versions of NxN_0 would take approximately one month of computer time, thus these were not computed due to lack of time.

The only qualitative difference between the Z scores of the lenient and intermediately thresholded networks is in the motif '+ + + −'. According to the extension from section 4.1.4, motifs with a negative edge product should be under-expressed, here we see a first case in which the motif is balanced with the random population.

Table 4.12: Variation of the square Balance with the filtering

| Z score | + + + + | + + + − | + + − − | + − + − | + − − − | − − − − |
|---|---|---|---|---|---|---|
| NxN_31 | 4.9 | -0.3 | 13 | 13 | -12 | -0.5 |
| NxN_7 | 4.4 | -0.4 | 12 | 13 | -10 | -0.4 |
| NxN_41 | -3.5 | -4.2 | -6.3 | -1.7 | 4.1 | 4.9 |

The results of the ensemble produced from the network filtered with protocol _41 are again strange. Similarly to what happened for ExE, the triangle distribution in the random population goes in line with the ones from the other protocols, but the squares show puzzling distributions. Although now there is signal from the NxN_41 network, we will disregard it again and attribute those nonsensical results to some problem in the protocol itself.

For the non-essential network, the '+ + + −' motif is no longer deprecated, apart from the stringent thresholding, this motif appears to be as common as in the random case. Although, the rule "squares with a negative edge product are deprecated", empirically verified for the essential interactions, is not contradicted, it is not totally clear if this particular motif is unstable or balanced.

A second difference resides in the positive sub-network. In the essential case, '+ + +' was under-represented while '+ + + +' was balanced, when compared to its random counterpart. Whereas in the non-essential case, '+ + +' is balanced and '+ + + +' is over-represented.

While in ExE it was doubtful if '− − − −' should be considered over-represented or balanced, in NxN

it appears clearly balanced. Besides this, the positive Z scores for '$- - -$' showed a stronger signal in ExE than in NxN.

It was already known that the essential and non-essential networks showed some differences, such as network density and many other functional divergences. We now know that the essential and non-essential networks seem to favor and deprecate some 4-node motifs in a disparate way, '$+ + + +$', '$+ + + -$' and '$- - - -$'.

## 4.3 Discussion

In conclusion, we sum up the motif distributions observed to propose a Genetic Structural Balance. Regarding the signed triangles, we verified that the essential genetic networks followed the reverse of the strong formulation of social Structural Balance (SB social), whereas the distributions in the non-essential networks, thresholded with intermediate confidence, followed the reverse of the weak formulation. However, since the latter results were not consistent for the different confidence thresholds, we will not propose a Structural Balance for the non-essential network (SB N) on the 3-node motifs, but only for the essential network (SB E). The signed triangle results are summarized in table 4.13.

Table 4.13: Genetic Structural Balance: triangles

| | + + + | + + − | + − − | − − − |
|---|---|---|---|---|
| SB social | | | | |
| SB E | | | | |
| SB N | xxx | xxx | xxx | xxx |

We emphasize one more time that the need to reverse the signs of the classic SB to match the genetic Balance is, to the best of our knowledge, a completely new result.

As for the squares, the empirically verified distributions for the genetic networks can be found in table 4.14. We also included the SB social, considering positive edge product squares as over-represented and negative ones as under-represented. Now there is consistency between the genetic and social SB's, without having to reverse signs, due to the even number of edges in the squares.

Table 4.14: Genetic Structural Balance: squares

| | + + + + | + + + − | + + − − | + − + − | + − − − | − − − − |
|---|---|---|---|---|---|---|
| SB social | | | | | | |
| SB E | | | | | | |
| SB N | | | | | | |

Despite the differences between the essential and non-essential empirical rules, it is important to notice that there is never direct contradiction. There is no motif simultaneously over-represented in one formulation and under-represented in the other. Thus, in the future, it may be possible to merge both definitions of the genetic SB, with broader studies including genetic networks of organisms other than the yeast.

# Chapter 5

# Conclusions

After a thorough statistical analysis on the 3 and 4-node motifs present in the yeast genetic interaction network, we reached the conclusion that some motifs are favored in relation to others.

The networks formed by the interactions of essential genes (the yeast does not survive their deletion) and non-essential genes (yeast still viable after their deletion), are known to present some topological differences. Nonetheless, the verified distribution of the 4-node motifs in both networks is congruent.

## 5.1 Results

The main result of the present work is the empirical derivation of rules of Balance for the genetic interaction networks of the budding yeast. We found that, for most networks, triangles with a positive edge product are stable and thus over-represented in the network structure, when compared to a null model, while triangles with a negative edge product are deprecated. The only exception was the non-essential network, when imposed a strict false positive rate on the interactions.

This means that, to correctly define the rules of balance in genetic networks, for triangles, one must invert the signals in the strong formulation of Structural Balance, a very well known and long-standing theory for balance in social networks. This is, to the best of our knowledge, a novel result for both the Biology and Network Science communities.

As for the squares, 4-node motifs with a negative edge product were never over-represented, when compared to the null model, and thus are probably unstable and deprecated by the network structure. Squares with a negative edge product were never found to be under-represented, and thus appear to be stable or balanced. These results match the ones expected in social networks.

One important motivation of this study was to find out which type of motif, triangles or squares, would reveal more informative for future sign prediction algorithms.

On one hand we have triangles, the 3-node signed motifs in the essential network presented not only the more disparate Z scores between patterns but also the greater consistency between different thresholding confidences and different filtering protocols. The latter are important arguments since

biological data is frequently noisy, and biological networks have a big variety of false negative and false positive rates. However, in the non-essential network, the Z scores of the triangles yielded contradictory results, depending on the thresholding confidence. Thus, we find 4-node motifs a more useful tool for sign prediction.

Squares have the initial advantage of being the chosen pattern in the state-of-the-art link-prediction algorithms in biology [5, 6], since sign prediction will most likely be associated with link-prediction in signed networks. Although there is some variation on the intensity of the Z scores signals, which may difficult the choice of each motifs' weight in a putative sign prediction algorithm, there is a great consistency in the qualitative results (over-represented, balanced or under-represented motif). Even though, as also for the triangles, some structural differences between the essential and non-essential genetic networks have been detected, these were never contradictory. Finally, a third argument is that, due to its even parity, there is no need to worry about how the edge signals were defined. A useful characteristic when applying the algorithm to a network of which we have little *a priori* information.

As motivated in the beginning of the project, biologists are interested in link-prediction algorithms to either guide their experimental assays and reduce costs, or to make accurate predictions in the search space currently inaccessible due to fundamental technical limitations. Our result is a first step in the way of designing a new successful link-prediction algorithm for genetic interaction networks.

## 5.2 Future Work

A natural follow-up of this work, now that we have a secure filtering protocol to build the network from the raw experimental data and a suited randomization algorithm averagely preserving the network signed degrees, is to build the global genetic interaction network for the yeast (ExE + ExN + NxN) and repeat the procedure. We expect to do this in the next few months.

Other interesting research is to check to which extent this genetic Structural Balance is verified in other organism other than *Saccharomyces cerevisiae*, starting with similar species as *Schizosaccharomyces pombe*, then moving to simple multicellular organisms as *Caenorhabditis elegans*, both *S. pombe* and *C. elegans* have genetic interaction data already available. And even, in a slightly more distant future, mammals.

The number of nodes and specially edges must be taken into account. Our programs, written in *C++*, can count squares in one hundred networks with $\sim$400 000 edges in a couple of days, but cannot handle the same number of networks with $\sim$1 400 000 edges without parallel computation, on networks this size we can still count triangles in a couple of hours. The computational speed is still a bottleneck and thus size is to be considered when selecting a network.

Another branch of future research is to apply the procedure to other signed biological networks, such as protein-protein interactions, or drug-drug interaction, in which a link-prediction algorithm is also welcomed by the experimentalists working in those fields. Alongside with this project, we have already started playing with some human drug-drug interaction networks. We remember that the randomization

model must be adapted to each network specificities, for drug-drug interactions we use the Signed Rewire model, introduced in section 3.3.

In truth, the key to unveil the connection between structure and function in the cell may reside in the merger of different biological networks, such as genetic and protein-protein interactions, two complementary and largely non-overlapping networks, enabling a broader view of the functional architecture of a cell [45].

We will continue to develop this project, in partnership with István Kovács, from the Northeastern University, in Boston, and in dialogue with the Boone Lab, from the University of Toronto, with the aim of publishing an article on the structural organization of biological networks and its correlation to function.

# Bibliography


[1] M. J. Mason et al. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics*, 10:327–51, July 2009.

[2] M. Costanzo et al. A global genetic interaction network maps a wiring diagram of cellular function. *Science*, 353(6306):31381, Sept. 2016.

[3] F. Li et al. Network-based computational drug combination prediction. *bioRxiv*, Apr. 2016.

[4] D. Liben-Nowelland and J. Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Tec.*, 58(7):1019–31, May 2007.

[5] I. A. Kovács et al. Network-based prediction of protein interactions. *bioRxiv*, Mar. 2018.

[6] A. Muscoloni, I. Abdelhamid, and C. Cannistraci. Local-community network automata modelling based on length-three-paths for prediction of complex network structures in protein interactomes, food webs and more. *bioRxiv*, June 2018.

[7] J. G. White et al. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Philosophical Transactions of the Royal Society B*, 314(1165), Nov. 1986.

[8] S. W. Oh et al. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–14, Apr. 2014.

[9] J. A. Dunne, R. J. Williams, and N. D. Martinez. Food-web structure and network theory: The role of connectance and size. *Proc. Natl. Acad. Sci. U.S.A*, 99(20):12917–22, Oct. 2002.

[10] F. Heider. Attitudes and cognitive organization. *The Journal of Psychology*, 21(1):107–112, Oct. 1946.

[11] D. Cartwright and F. Harary. Structural balance: a generalization of Heider's theory. *Psychological Review*, 63(5):277–93, Sept. 1956.

[12] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–70, Apr. 2010.

[13] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci. U.S.A*, 107(31):13636–41, Aug. 2010.

[14] G. Facchetti, G. Iacono, and C. Altafini. Computing global structural balance in large-scale signed social networks. *Proc. Natl. Acad. Sci. U.S.A*, 108(52):20953–58, Dec. 2011.

[15] X. Chen, Z. L. Ji, and Y. Z. Chen. TTD: Therapeutic Target Database. *Nucleic Acids Research*, 30 (1):412–15, Jan. 2002.

[16] Bioinformatics and Drug Design group, N.U. Singapore. Therapeutic Target Database. `http://bidd.nus.edu.sg/group/cjttd/`, Sept. 2017. last accessed 2018-10-13.

[17] D. S. Wishart et al. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Research*, 34(suppl_1):D668–72, Jan. 2006.

[18] D. Wishart et al. DrugBank. `https://www.drugbank.ca/`, Nov. 2017. last accessed 2018-10-13.

[19] Y. Liu et al. DCDB: Drug combination database. *Bioinformatics*, 26(4):587–88, Feb. 2010.

[20] Y. Liu et al. Drug Combination Database. `http://www.cls.zju.edu.cn/dcdb/`, Dec. 2014. last accessed 2018-10-13.

[21] U.S. Food and Drug Administration. Orange Book: approved drug products with therapeutic equivalence evaluations. `https://www.accessdata.fda.gov/scripts/cder/ob/`, Sept. 2018. last accessed 2018-10-13.

[22] C. von Mering et al. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–61, Jan. 2003.

[23] STRING Consortium 2018. STRING database. `https://string-db.org/`, May 2017. last accessed 2018-10-13.

[24] J.-F. Rual et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–78, Oct. 2005.

[25] H. Yu et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–10, Oct. 2008.

[26] T. Rolland et al. A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–26, Nov. 2014.

[27] A. H. Tong et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–68, Dec. 2001.

[28] A. H. Tong et al. Global mapping of the yeast genetic interaction network. *Science*, 303(5659): 811–13, Feb. 2004.

[29] M. Costanzo et al. The genetic landscape of a cell. *Science*, 327(5964):423–31, Jan. 2010.

[30] Donnelly Center, U. Toronto. Global Yeast Genetic Interaction Dataset. `http://thecellmap.org/costanzo2016/`, May 2016. last accessed 2018-09-5.

[31] A. Baryshnikova. Systematic functional annotation and visualization of biological networks. *Cell Systems*, 2(6):412–421, June 2016.

[32] M. Ashburner et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1): 25–29, May 2000.

[33] P. Erdős and A. Rényi. On random graphs I. *Publicationes Mathematicae (Debrecen)*, 6:290–97, 1959.

[34] E. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–44, 1959.

[35] A.-L. Barabási. *Network Science*, chapter 3: Random Networks. Cambridge University Press, Cambridge, UK, 2016.

[36] T. P. Peixoto. The graph-tool python library. `http://figshare.com/articles/graph_tool/1164194`, Sept. 2014. last accessed 2017-05-13.

[37] F. Iorio et al. Efficient randomization of biological networks while preserving functional characterization of individual nodes. *BMC Bioinformatics*, 74:542–55, Dec. 2016.

[38] E. Bender and E. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, May 1978.

[39] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci. U.S.A*, 99(25):15879–82, Dec. 2002.

[40] F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–45, Nov. 2002.

[41] J. Miller and A. Hagberg. Eficient generation of networks with given expected degrees. In *Algorithms and Models for the Web Graph (WAW2011)*, pages 115–126. Springer, 2011.

[42] H. Sayama. Combinatorial Miller-Hagberg algorithm for randomization of dense networks. *arXiv*, Oct. 2017.

[43] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *ACM WWW International Conference on World Wide Web*, pages 641–50, Apr. 2010.

[44] G. Strona et al. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature Communications*, 5(1), June 2014.

[45] B. VanderSluis et al. Integrating genetic and protein-protein interaction networks maps a functional wiring diagram of a cell. *Current Opinion in Microbiology*, 45:170–179, July 2018.