Exploitation of Complex Network Topology for Link Prediction in Biological Interactomes

Thesis by
Gregorio Alanis Lobato

Submitted in Partial Fulfillment of the Requirements for the degree of
Doctor of Philosophy

King Abdullah University of Science and Technology
Computer, Electrical and Mathematical Sciences & Engineering Division
Computer Science Department

Thuwal, Makkah Province, Kingdom of Saudi Arabia
June, 2014
The undersigned approve the thesis of Gregorio Alanis Lobato

Dr. Timothy Ravasi
Associate Professor, Bioscience, King Abdullah University of Science and Technology
Committee Chair (Thesis Supervisor)

Dr. Serafin Batzoglou
Associate Professor, Computer Science, Stanford University
Committee Member

Dr. Victor Solovyev
Professor, Computer Science,
King Abdullah University of Science and Technology
Committee Member

Dr. Mikhail Moshkov
Professor, Applied Mathematics and Computational Science,
King Abdullah University of Science and Technology
Committee Member

Dr. Xin Gao
Assistant Professor, Computer Science,
King Abdullah University of Science and Technology
Committee Member

King Abdullah University of Science and Technology
2014
ABSTRACT

Exploitation of complex network topology for link prediction in biological interactomes

Gregorio Alanis Lobato

The network representation of the interactions between proteins and genes allows for a holistic perspective of the complex machinery underlying the living cell. However, the large number of interacting entities within the cell makes network construction a daunting and arduous task, prone to errors and missing information.

Fortunately, the structure of biological networks is not different from that of other complex systems, such as social networks, the world-wide web or power grids, for which growth models have been proposed to better understand their structure and function. This means that we can design tools based on these models in order to exploit the topology of biological interactomes with the aim to construct more complete and reliable maps of the cell.

In this work, we propose three novel and powerful approaches for the prediction of interactions in biological networks and conclude that it is possible to mine the topology of these complex system representations and produce reliable and biologically meaningful information that enriches the datasets to which we have access today.
ACKNOWLEDGEMENTS

First of all, I would like to thank my PhD advisor Prof. Timothy Ravasi for giving me the opportunity to work with him, for his guidance and for giving me the freedom to pursue my own research interests. I would also like to thank Dr. Carlo Vittorio Cannistraci, my assigned post-doc, for fighting right by my side to get our work published.

An honourable mention goes to my family and friends, especially uncle Mariano and aunt Aurora, my brothers, Mariano and Joel, and my parents, Gregorio and Matilde, because thanks to all of them I was able to get to this stage of my career. Their unconditional support and love have been invaluable for me to never give up.

In addition, I would like to thank my beautiful wife, Mafe, because thanks to her support, understanding and immense love I was able to overcome the challenges that life presented to me during my PhD.

Finally, I would like to thank King Abdullah bin Abdulaziz Al Saud for dreaming of KAUST and KAUST itself for letting me live my dreams. The environment, gorgeous and state-of-the-art facilities and top-class faculty of our University allowed me to take my research to the next level and made of me not only a better scientist but also a better and more enriched individual. At KAUST’s beautiful campus, right by the Red Sea shore, one cannot help it but to discover through inspiration.
# TABLE OF CONTENTS

Copyright 3  
Abstract 4  
Acknowledgements 5  
List of Abbreviations 8  
List of Figures 11  
List of Tables 13  
1 Introduction 14  
2 Background and General Methodology 17  
  2.1 Complex Systems and Systems Biology 17  
    2.1.1 Complex Systems and the Development of Network Science 17  
    2.1.2 Systems Biology and Biological Interactomes 19  
    2.1.3 Protein-protein Interaction Networks 19  
    2.1.4 Genetic Interaction Networks 20  
  2.2 The Link Prediction Problem 20  
    2.2.1 Link Prediction Framework 22  
    2.2.2 Performance Evaluation of Topological Link Predictors 23  
    2.2.3 Important Comments about the Evaluation Process by Gene Ontology 27  
3 From Link Prediction in Brain Connectomes and Protein Interactomes to the  
  Local Community Paradigm in Complex Networks 29  
  3.1 A Local Community Approach to Link Prediction 30  
  3.2 Link Prediction in Brain Connectomes 32  
  3.3 Link Prediction in Social and Ecological Networks 35  
  3.4 Link Prediction in Protein Interactomes 36  
  3.5 The Local Community Paradigm in Real Networks 40  
  3.6 The LCP in Idealised Networks 44  
  3.7 Discussion 48  
4 Minimum Curvilinearity to Enhance Topological Prediction of Protein Interactions by Network Embedding 50  
  4.1 Network Datasets 52
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>Adamic &amp; Adar</td>
</tr>
<tr>
<td>ACDD</td>
<td>Adjusted CDD</td>
</tr>
<tr>
<td>ADN</td>
<td>Autoimmune Disease Network</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under the ROC Curve</td>
</tr>
<tr>
<td>AUP</td>
<td>Area Under the Precision Curve</td>
</tr>
<tr>
<td>BP</td>
<td>Biological Process</td>
</tr>
<tr>
<td>CAA</td>
<td>CAR variant of Adamic &amp; Adar</td>
</tr>
<tr>
<td>CAR</td>
<td>Cannistraci-Alanis-Ravasi</td>
</tr>
<tr>
<td>CC</td>
<td>Cellular Compartment</td>
</tr>
<tr>
<td>CD</td>
<td>Celiac Disease</td>
</tr>
<tr>
<td>CDD</td>
<td>Czekanowski-Dice Dissimilarity</td>
</tr>
<tr>
<td>CJC</td>
<td>CAR variant of Jaccard</td>
</tr>
<tr>
<td>CN</td>
<td>Common Neighbours</td>
</tr>
<tr>
<td>CPA</td>
<td>CAR variant of Preferential Attachment</td>
</tr>
<tr>
<td>CRA</td>
<td>CAR variant of Resource Allocation</td>
</tr>
<tr>
<td>dB</td>
<td>Decibel</td>
</tr>
<tr>
<td>DRYGIN</td>
<td>Data Repository of Yeast Genetic INteractions</td>
</tr>
<tr>
<td>ED</td>
<td>Euclidean Distance</td>
</tr>
<tr>
<td>ER</td>
<td>Erdős and Rényi</td>
</tr>
<tr>
<td>ERPE</td>
<td>External Referee Performance Evaluation</td>
</tr>
<tr>
<td>ERPES</td>
<td>External Referee Performance Evaluation by Sparsification</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>FSW</td>
<td>Functional Similarity Weight</td>
</tr>
<tr>
<td>GI</td>
<td>Genetic Interaction</td>
</tr>
<tr>
<td>GIN</td>
<td>GI Network</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>GWAS</td>
<td>Genome Wide Association Study</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>IG1</td>
<td>Interaction Generality 1</td>
</tr>
<tr>
<td>ISO</td>
<td>Isomap</td>
</tr>
<tr>
<td>IWS</td>
<td>Intense World Syndrome</td>
</tr>
<tr>
<td>JC</td>
<td>Jaccard</td>
</tr>
<tr>
<td>LC</td>
<td>Local Community</td>
</tr>
<tr>
<td>LCL</td>
<td>Local Community Links</td>
</tr>
<tr>
<td>LCP</td>
<td>Local Community Paradigm</td>
</tr>
<tr>
<td>LCP-corr</td>
<td>LCP Correlation</td>
</tr>
<tr>
<td>LCP-DP</td>
<td>LCP Decomposition Plot</td>
</tr>
<tr>
<td>MC</td>
<td>Minimum Curvilinearity</td>
</tr>
<tr>
<td>MCE</td>
<td>Minimum Curvilinear Embedding</td>
</tr>
<tr>
<td>MDS</td>
<td>Multidimensional Scaling</td>
</tr>
<tr>
<td>MF</td>
<td>Molecular Function</td>
</tr>
<tr>
<td>MST</td>
<td>Minimum Spanning Tree</td>
</tr>
<tr>
<td>ncISO</td>
<td>non-centred Isomap</td>
</tr>
<tr>
<td>ncMCE</td>
<td>non-centred MCE</td>
</tr>
<tr>
<td>NeRV</td>
<td>Neighbour Retrieval Visualiser</td>
</tr>
<tr>
<td>PA</td>
<td>Preferential Attachment</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein-protein Interaction</td>
</tr>
<tr>
<td>PPIN</td>
<td>PPI Network</td>
</tr>
<tr>
<td>PRPE</td>
<td>Prune-rediscover Performance Evaluation</td>
</tr>
<tr>
<td>PRPES</td>
<td>Prune-rediscover Performance Evaluation by Sparsification</td>
</tr>
<tr>
<td>RA</td>
<td>Resource Allocation</td>
</tr>
<tr>
<td>Res</td>
<td>Resolution</td>
</tr>
<tr>
<td>RGG</td>
<td>Random Geometric Graphs</td>
</tr>
<tr>
<td>RhA</td>
<td>Rheumatoid Arthritis</td>
</tr>
<tr>
<td>RIN</td>
<td>Residue Interaction Network</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>SNE</td>
<td>Stochastic Neighbourhood Embedding</td>
</tr>
<tr>
<td>SP</td>
<td>Shortest Path</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>TID</td>
<td>Type I Diabetes</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>TPE</td>
<td>Tree Preserving Embedding</td>
</tr>
<tr>
<td>wAA</td>
<td>Weighted Adamic &amp; Adar</td>
</tr>
<tr>
<td>wACDD</td>
<td>Weighted Adjusted Czekanowski-Dice Dissimilarity</td>
</tr>
<tr>
<td>wCAR</td>
<td>Weighted Cannistraci-Alanis-Ravasi</td>
</tr>
<tr>
<td>wCN</td>
<td>Weighted Common Neighbours</td>
</tr>
<tr>
<td>wPA</td>
<td>Weighted Preferential Attachment</td>
</tr>
<tr>
<td>Y2H</td>
<td>Yeast Two-Hybrid</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

2.1 Types of biological interactions investigated in this thesis. .......................... 21
2.2 Link predictors based on network paradigms. ........................................... 21
2.3 Link prediction framework. ................................................................. 22
2.4 Link prediction performance evaluation. ................................................. 25
2.5 Performance evaluation by sparsification. .............................................. 26
2.6 GO biological relevance test. ............................................................... 28

3.1 CAR index. ................................. ................................. ................................. 33
3.2 PRPES in brain connectomes, social and ecological networks. .................. 38
3.3 ERPE in protein interactomes. ............................................................. 39
3.4 Validation of top-ranked PPIs in STRING. ............................................. 40
3.5 LCP-DP, LCP-correlation and ERPES on PPINs. ..................................... 45
3.6 LCP in real networks. ........................................................................... 46
3.7 LCP in idealised networks. ...................................................................... 48

4.1 MCE/ncMCE for network embedding. ...................................................... 56
4.2 Scoring approaches for link prediction by network embedding. ................ 57
4.3 Dimension determination criteria. ............................................................ 60
4.4 Fraction of FPs visited by SPs and MC. .................................................... 61
4.5 Embedding of the radar signal dataset. .................................................... 63
4.6 Discrimination between good and bad candidate PPIs. ............................ 64
4.7 PRPES on RGGs. .................................................................................... 66
4.8 MCE vs Isomap. ..................................................................................... 67
4.9 ncMCE vs embedding techniques. ............................................................ 69
4.10 ncMCE vs bio-inspired techniques. ........................................................ 70
4.11 AUC and Resolution criteria for dimension determination. .................... 71
4.12 PPIN sparsification and re-densification. ............................................... 73
4.13 In silico STRING validation. ................................................................. 74

5.1 GIN sparsification. .................................................................................. 81
5.2 ACDD improvement on CDD. ................................................................. 82
5.3 ACDD outperforms the best links predictors. ......................................... 83
5.4 Overlap of top-100 candidate GIs and DRYGIN. ................................... 85

6.1 Construction of the ADN. ....................................................................... 89
6.2 Structural properties of the ADN. ............................................................ 90
6.3 The ADN after community detection. ........................................... 91
6.4 PRPES on the ADN communities. .............................................. 94

A.1 LCP-DP for networks of biological origin. .................................. 114
A.2 LCP-DP for social networks. .................................................... 115
A.3 LCP-DP for networks of different origins. .................................. 116
LIST OF TABLES

3.1 Formulae for the classical, CAR-based and bio-inspired neighbourhood-based techniques ................................................................. 32

6.1 Weighted link predictors ................................................................. 92

A.1 Network datasets ....................................................................... 111
Chapter 1

Introduction

Complex systems, sets of interacting components whose collective behaviour produce emergent properties that are difficult to understand by the study of the individual parts, can be represented as graphs, also known as networks, where nodes represent the system components and links between these nodes represent the relationships between them. Networks not only have facilitated the visualisation and analysis of such systems but have also led to the conclusion that there are structural properties common to several complex system representations, regardless of their domain (Albert and Barabási, 2002).

The information associated with the organisation of network topologies (topology referring to the way in which system components are connected and their properties therein) was generated gradually and in accordance with the growth process of the network, i.e. how components joined the system and how links started to form between them. As a result, this information imprinted in the topology of complex networks can be exploited for different purposes, which range from detecting the most salient system components (Jeong et al., 2001), to determining the shortest route from one to another (Johnson, 1977), to establishing the entities that allow for the system’s controllability (Liu et al., 2011; Cowan et al., 2012; Ruths and Ruths, 2014), to forecasting the future state of the network structure (Liben-Nowell and Kleinberg, 2007). The latter, known as the link prediction problem, is the focus of this thesis. In this document, the extent to which link prediction is possible on the mere basis of network topology is investigated, with a special focus on biological interactomes.

Link prediction is not only important for companies such a Facebook, interested in predicting new acquaintances for commercial purposes, but it is also important for fields such as systems biology, which deals with networks of interactions between proteins or genes (Lesk, 2007). Despite big efforts to chart complete protein and genetic interaction maps, experimental errors, biases and constraints, result in bio-networks that are still incomplete and contain large amounts of spurious interactions.
(Vidal et al., 2011). From this perspective, link prediction is not only very important in biology, in the sense that its application reduces the universe of interactions to test in the lab and might result in more complete and reliable interaction maps, but it is also challenging, in the sense that link predictors must be designed in such a way as to avoid the spurious information stored in the topology under study.

This work’s contributions to network science and systems biology are:

- A novel and accurate topological link predictor that belongs to the category of node-neighbourhood approaches, based on the intuition that if two persons do not know each other but they have a lot of friends in common and, at the same time, these common friends are acquainted, there is a high likelihood that these two persons will meet in the future (Chapter 3, Section 3.1).

- A new paradigm of link formation in dynamic, heterogeneous networks without growth, developed on the basis of the above intuition (Chapter 3, Section 3.5).

- A new quantifiable network structural feature along with a new means to visualise the structure of complex networks, based on the above mentioned paradigm (Chapter 3, Section 3.5).

- A novel and powerful topological link predictor that belongs to the category of network embedding approaches, based on the hypothesis that a network topology lies near or on a low-dimensional manifold embedded in a high-dimensional space. This technique was especially designed to deal with noisy and sparse networks (Chapter 4, Section 4.3).

- Two methods for the automatic determination of the dimension of network embedding that results in the best performance of a network embedding-based link predictor (Chapter 4, Section 4.4).

- A graph dissimilarity index that belongs to the category of node-neighbourhood approaches, based on the combination of the concepts of network clustering and preferential attachment (Chapter 5, Section 5.3).

- A network densification experiment that, combined with the link prediction task, clearly shows that protein and genetic interactomes are not formed by a uniform random process and that they contain important information stored in their topology, which can be exploited for several purposes (Chapter 4, Section 4.5.7 and Chapter 5, Section 5.3).

- A network-based pipeline for the study of Genome-wide Association data along with a new network dataset of genes linked whenever there is evidence that they are both associated with
at least one autoimmune disease (Chapter 6).

- The weighted version of the node-neighbourhood based approaches introduced for link prediction in weighted networks (Chapter 6, Section 6.3.1).

The development and application of the above listed contributions are organised throughout this document as follows. An overview of the history of network science, the introduction of network formation models, their application to different domains such as systems biology, the importance of topological link prediction and the general methodology used throughout the thesis appear in Chapter 2. The new paradigm of link formation in static networks and the novel link predictor, network visualisation tool and structural measure arising from this paradigm are presented in Chapter 3. Network embedding, its application to link prediction and the proposal of a new embedding-based predictor in the context of protein interactomes are discussed in Chapter 4. The importance of epistatic networks and link prediction over their distinctive topologies are considered in Chapter 5. The representation of Genome-wide Association data as a weighted network, the study of the resulting topology with network-based tools and weighted link prediction are described in Chapter 6. Finally, Chapter 7 includes a summary of the entire document, the significance of this work’s contributions and future research opportunities deriving from the results here presented.

It is worth noting that this thesis has three appendices: Appendix A lists all the network datasets used in this work, Appendix B discusses the computational time complexity of different link prediction techniques and Appendix C contains the list of publications arising from this work.
Chapter 2

Background and General Methodology

This chapter contains excerpts from the following publications:


2.1 Complex Systems and Systems Biology

2.1.1 Complex Systems and the Development of Network Science

Systems are said to be complex when the interaction of heterogeneous components leads to collective behaviours that are not directly correlated with the singular component behaviour. Some examples are the cohort of instruments that interact to generate a symphony, or the interplay of the different body organs that physiologically support a human action. To facilitate the study and visualisation of complex systems, they can be represented as graphs/networks of interactions (links/edges) be-
tween their constituents (nodes/vertices). In order to better understand the sophisticated processes that result in their realization, evolution and behaviour, an important number of tools have been developed, borrowing theories and methods from very diverse disciplines that, together, have led to the development of what today is known as network science.

The beginnings of network science date back to 1735, when Leonhard Euler represented land masses from the Prussian city of Königsber as nodes and bridges connecting them as edges with the aim to investigate whether there was a walk through the city that would cross each bridge once and only once, which he proved to be impossible (Euler, 1741). However, the formal study of network properties as a means to define general paradigms of complex systems, started with the studies on random networks of Erdős and Rényi (ER) in the late 1950s (Erdős and Rényi, 1960). Unfortunately, due to the lack of large network datasets, they were unable to test their predictions on real data.

Nearly four decades later, Duncan Watts and Steven Strogatz reconciled empirical data on networks with mathematical representations, describing the small-world paradigm, which untangled how regular networks react if a certain amount of disorder, in the form of a rewiring probability $p$, is introduced in their topological organisation (Watts and Strogatz, 1998). They found that network representations of real systems are highly clustered like regular lattices and present a low average shortest-path length like ER graphs. One year later Albert-László Barabási and Réka Albert explained how the scale-invariance of many real networks originates from a specific growth process, in which a new node tends with higher probability to be linked to those nodes (network hubs) that already have a large number of neighbours, i.e. a high node degree (Barabási and Albert, 1999) (this paradigm, named preferential attachment is consistent with a previous study by de Solla Price on scientific networks (de Solla Price, 1976)). Although the model of network growth proposed by Barabási and Albert generates networks with node degree distributions decaying as a power law, as it is the case for several real networks, they lack the characteristic high clustering present in real network topologies. To account for this missing property, Krioukov and colleagues proposed a model that generates scale-free, highly-clustered networks by distributing a fixed number of nodes $N$ quasi-uniformly at random in the hyperbolic space $\mathbb{H}^2$ represented by the Poincaré disc of radius $R \sim \ln N$ and by connecting each node-pair with a probability that depends on the inter-node hyperbolic distance (Krioukov et al., 2010). Based on the same geometric framework, Papadopoulos and colleagues proposed the Popularity-Similarity model that is also able to generate scale-free, clustered networks but in a growing fashion (Papadopoulos et al., 2012). These last two models allow for the study of network structure and function from a geometric framework, in which power-law degree distributions and strong clustering emerge naturally as a result of the mathematical properties of
hyperbolic spaces.

2.1.2 Systems Biology and Biological Interactomes

The application of network science has had a positive impact in diverse scientific fields. Good examples are social sciences (Wasserman and Faust, 1994) and systems biology (Loscalzo and Barabasi, 2011). The latter, whose goal is to unravel the complexity of biosystems, has gained special attention in the scientific community because the focus of systems biology on the integration of gene and protein activity, could boost the development of personalised medicine (Barabási et al., 2011; Loscalzo and Barabasi, 2011; Vidal et al., 2011) and improve our knowledge of the intricate mechanisms that make life possible (Lesk, 2007).

Systems biology deals with networks of proteins, genes or metabolites and the physical or logical interactions amongst them (Lesk, 2007). Although the network-based techniques presented in this thesis were applied to network representations of systems from different domains (see Appendix A), our main investigations were done on Protein-protein Interaction (PPI) and Genetic Interaction (GI) networks.

2.1.3 Protein-protein Interaction Networks

Most functions within the cell occur thanks to the organised and coordinated interaction between proteins. Failure of control mechanisms in charge of these delicate processes can derive in disease or even death (Lesk, 2007) and that is why the study of PPI Networks (PPINs) is extremely important. In PPINs, nodes represent proteins and undirected edges can represent a physical docking between two proteins due to amino acid attractive forces (see Figure 2.1A) or logical relationships in the sense that the activity of a protein is affected by a change in an external process in which another protein is involved (Vidal et al., 2011; Lesk, 2007).

A myriad of methods for detecting PPIs in a high-throughput fashion have been proposed but they are usually an improved variation of the Yeast Two-Hybrid (Y2H) screening system or of co-immunoprecipitation followed by mass spectrometry (Lesk, 2007; Vidal et al., 2011). Despite many efforts to improve the accuracy and sensitivity of these mapping techniques (Vidal et al., 2011), the network structures resulting from their application may present high levels of noise, i.e. a number of false positive interactions that can ascend up to 60% and a number of false negatives (representative of the incompleteness of these datasets) that can go up to 40% (Kuchaiev et al., 2009).
2.1.4 Genetic Interaction Networks

Phenomena like incomplete penetrance or the missing heritability problem, have prompted scientists to find alternatives to the common disease-common variant hypothesis (Burga et al., 2011; Visscher et al., 2012; Lehner, 2013). As a result, the interest in understanding how genes contribute collectively to determine phenotype has increased and one of the strategies to address the problem is through the detection of GIs and the construction of GI Networks (GINs) (Costanzo et al., 2011).

A GI or epistatic interaction is detected when the simultaneous mutation of two genes generates a deviation in phenotype that is worse (negative interaction) or better (positive interaction) than that generated by the multiplicative effect of the single mutant phenotypes (Boone et al., 2007) (see Figure 2.1B). GIs reveal functional compensation and buffering under genetic perturbations, thus providing information about the functional associations between biological pathways and protein complexes (Costanzo et al., 2011). If the establishment of the whole set of GIs for every gene were possible, one would have a means to untangle the complex relationships between genotype and phenotype and a detailed functional map of the cell (Costanzo et al., 2011), which has direct impact in our understanding of human diseases and the ability to design personalised treatment plans. Although there exist very well-established methods for the high-throughput detection of epistasis in model organisms (Lehner, 2013) such as the budding yeast Saccharomyces cerevisiae (Costanzo et al., 2010) or the nematode worm Caenorhabditis elegans (Byrne et al., 2007), that result in systematically more accurate and complete GINs compared to PPINs, given the huge amount of gene pairs to test for interaction and the difficulty to monitor phenotypic changes in human there is not a complete nor a reliable human epistatic network. Fortunately, there are works around this issue like the approach used by Goh and colleagues for the construction of a human disease network (Goh et al., 2007), which we exploit in Chapter 6 for the study of interactions between genes associated with autoimmune diseases.

2.2 The Link Prediction Problem

Do the network growth models described in the previous section represent realistic mechanisms of complex system formation? In 2001, Mark Newman decided to carry out an empirical study to test whether the paradigms of clustering and preferential attachment do play an important role in link formation processes (Newman, 2001a). He analysed the network topology of collaborations between scientists who published in the Los Alamos Archive and in Medline at two different time points
(1999 and 2000 for the former and 1998 and 1999 for the latter) and found that the probability of a pair of disconnected scientists working together in the second network snapshot increases with the number of collaborators that they have in common, and that the probability of a scientist attracting new collaborators in the second network snapshot increases with the number of past collaborations in the first one (Newman, 2001a). Newman’s findings opened a new line of research within network science aiming to predict links based only on the structural motifs that form complex networks (see Figure 2.2 for examples of techniques for link prediction that arose thanks to Newman’s results). We refer to this prediction approach as Topological Link Prediction and we deeply investigate its application throughout this thesis.

Figure 2.2: For a pair of non-adjacent nodes $x - y$ (dashed line), the Common Neighbours (CN) index assigns a likelihood score based on the intersection (white nodes in the figure) between the sets of neighbours of $x$ and $y$ ($\Gamma(x)$ and $\Gamma(y)$ respectively). The Preferential Attachment (PA) index scores based on the product of the degrees of $x$ and $y$ (the number of blue edges times the number of green edges).

\[
\text{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)| = 3 \\
\text{PA}(x, y) = |\Gamma(x)| \cdot |\Gamma(y)| = 6 \cdot 5 = 30
\]
2.2.1 Link Prediction Framework

The idea of predicting links on the mere basis of network topology originates from the fact that part of the information associated with the organisation of the topology itself was generated gradually and in accordance with the growth process of the network. Links between nodes exist because two people have similar interests (social networks), two proteins are bound together to perform a specific function (protein interaction networks), or two cities are connected through a direct flight (flight maps); nodes and edges are only abstractions of the dynamics that exist within complex systems, and the information contained within these abstractions can definitely be exploited for the prediction of new links, especially in cases where there is very few, unreliable or null information about the system’s interacting entities (as it is often the case in biological networks like the ones described in the previous section).

How to take advantage of existing network topologies to predict missing edges is still not clear. However, techniques have been developed to do so, both in unspecific networks (here referred to as General Purpose and Maximum Likelihood techniques, see Chapter 3) and in very specific ones, such as biological networks (see Chapters 3, 4 and 5). Taking the network of interest as the input, these prediction techniques proceed as follows (see Figure 2.3):

1. Assign a likelihood score to every pair of nodes that is not directly connected in the current network topology (each of these pairs is called a candidate interaction/link).

2. Sort the list of candidate interactions by likelihood score (for some approaches, the larger the score, the more the interaction is likely to be real, for others the scoring works in the opposite way).

![Figure 2.3](image)

More formally, given the network representation of a system with $N$ nodes and $L$ links, we can
describe this observable network topology with an $N \times N$ symmetric, adjacency matrix $A^O$ with entries $A^O_{ij}$ that take values 0 or 1 if there is an interaction between nodes $i$ and $j$ or if there is not, respectively (this thesis only deals with undirected, connected, unweighted networks, see Chapter 6 for an exception). A link predictor produces a list with $\frac{N(N-1)}{2} - L$ rows, based on the topology of $A^O$, with the best candidate interactions at the top of the list. These candidate links are very likely to be present in a true network $A^T$ with $L + \ell$ links, which we do not have access to because there is still no direct flight between airports $i$ and $j$, if $A^O$ is a flight map; individuals $i$ and $j$ are not friends yet, if it is a social network; or biologists have not been able to test for the interaction between proteins $i$ and $j$ due to experimental constraints, if $A^O$ is a PPIN.

### 2.2.2 Performance Evaluation of Topological Link Predictors

Performance assessment for any given prediction tool can avail of several methods; specific choices depend mainly on the type of network and the available information about the nodes and links. The most common approaches are (see Figure 2.4):

**Situation A** Some systems allow access to their network representations at different timestamps. One example is Facebook; the number of links at time $t_{i+1}$ is (almost certainly) larger than at time $t_i$. Thus, one can apply a prediction technique to the network topology at $t_i$ and verify whether the links at the top of the candidate interaction list appear in the network at $t_{i+1}$. Liben-Nowell and Kleinberg formalised the link prediction problem in this situation by applying the performance evaluation process just described to two snapshots of five coauthorship networks from arXiv (Liben-Nowell and Kleinberg, 2007). However, given the difficulty to have access to the temporal evolution of a network, either because there are few entities that monitor this process or because companies do not provide public access to their datasets’ changes over time, this situation is not covered in this thesis.

**Situation B** When networks at different timestamps are not available, the procedure is to remove $\ell$ links randomly from the available network, obtain a candidate interaction list that is sorted by likelihood using a prediction technique, and take $\ell$ links from the top of this list to verify if they match those removed from the network. This performance evaluation process is usually applied to networks whose topology can be trusted, i.e. their links can be considered true positives, and it is usually called Prune-rediscover Performance Evaluation.

**Situation C** For some systems, it is possible to access reliable node property information that serves as a criterion to decide whether two nodes can interact or not. An example is Gene
Ontology (GO), a vocabulary for processes, functions and cellular localisation of genes and proteins. If, for instance, two proteins perform the same function and are located in the same cellular compartment, they are likely to interact. Thus, the proportion of interactions at the top of the candidate list that fulfils the given criterion constitutes a prediction technique’s performance measure. Note that, in contrast with Situation B, this evaluation process is an assessment of pure link prediction in which no links are removed from the observed network topology. In this case, the GO serves as the external referee that decides whether the high-scored candidate interactions are biologically relevant or not. This GO strategy has been adopted in past studies that deal with link prediction in bio-networks (Saito et al., 2002, 2003; Chen et al., 2005, 2006a,b; You et al., 2010). The proteins or genes involved in the interactions from the candidate list are annotated via GO terms (molecular function or MF, biological process or BP, and cellular compartment or CC). If the terms associated with a protein or gene pair have a high GO semantic similarity (more on how to compute this similarity in the following subsection), the interaction is considered to be biologically relevant (marked with a Yes in the table in Figure 2.4, rightmost panel) and is used to quantify the precision of the predictors. A recursive procedure is applied to create a performance curve. Each time, an increasing fraction of candidate links (the first 100, the first 200 and so on) is taken from the top of the list of ranked candidate interactions for consideration. The fraction of candidate interactions that are relevant to GO generates a point on the performance curve. Conventionally, a number of top-ranked candidate links equivalent to 10% of the links in the observed network, is used to compute the entire performance curve (You et al., 2010) mainly because it has been estimated that the complete biological interactomes should have around 10% additional interactions to the ones that have been detected so far (Sprinzak et al., 2003).

The Area Under the Precision Curve (AUP), normalised with respect to the x-axis so that it ranges from 0 to 1, summarises the performance of the prediction technique for a given network (see Figure 2.4, rightmost panel). The External Referee Performance Evaluation procedure is usually applied to networks whose topology cannot be 100% trusted, i.e. they can be the result of experimental bias and contain a high percentage of spurious interactions.

It is worth noticing that in all the above described situations, the performance measure of choice is Precision = \( \frac{TP}{TP + FP} \) (TP = True Positives, links pruned from the observed network and that a link predictor has to rediscover or candidate links from a PPIN or GIN that are biologically relevant; FP = False Positive, links pushed to the top of the list that do not belong to the list of pruned
Figure 2.4: Depending on the type of network under study and on how reliable it is, there are different approaches to evaluate the performance of a link predictor. In any case, all approaches resort to Precision as a measure of performance, due to the reasons exposed in this subsection.

edges or that are not biologically relevant). Even when other performance statistics, like the Area Under the ROC Curve (AUC), have been used (Clauset et al., 2008; Guimerà and Sales-Pardo, 2009), Precision should be preferred to avoid having an ill-posed performance evaluation process. To compute an AUC, the number of True Negatives (TN, interactions that are definitely not present in the system) and False Negatives (FN, interactions that occur or can occur but have not been detected or realized) are needed, along with the above mentioned TPs and FPs. In any network topology, labelling a non-present interaction as TN or FN is wrong because that interaction has a probability to occur at any given time of the system’s evolution (making it a FN) and, at the same time, it also has a probability to not occur at all (making it a TN). For instance, a GI may not be present in the networks available today because the experimental techniques needed to detect it are not yet available or a direct flight from one airport to another does not exist yet because one of the cities connected by this flight is not a tourist destination yet (think of places like Dubai, UAE). When using Precision, such a compromise is not made because labelling links as TPs and FPs is much safer, especially in biological interactomes where an external referee such as the GO is at hand.

It is also important to point out that, whenever the performance of a link predictor is assessed, this performance is compared with that of a random link predictor that, instead of generating a candidate list with scores based on network structural features, it simply outputs a random permutation of every node-pair not directly connected in the observable network topology under study.
In order to test the true robustness of a link predictor to lack of topological information, a strategy for evaluating its performance at different levels of random sparsification of the observed network can be employed. The idea is to remove a fixed portion of links $\ell = 0.1L$ (10% of the links) uniformly at random from the original topology (sparsification process). Then, the same fixed amount of random links, $\ell$, is pruned from the network sparsified in the previous step (a total of 20% of links removed). This process is repeated several times up to the point where network connectivity is about to be lost. The performance of the link prediction technique is computed at each sparsification step to generate a sparsification curve that depicts the link predictor’s power and tolerance to link removal (see a graphic representation of the entire process in Figure 2.5). To gain statistical power, several sparsification curves are constructed, averaged and reported along with the average performance curve of the above described random predictor.

![Figure 2.5](image)

*Figure 2.5:* In the performance evaluation process by sparsification, a fixed amount of edges is removed from the observed network topology and a link predictor is applied to the pruned network. Then the ability of the link predictor to push the pruned edges to the top of the list of candidate links is assessed and reported as a performance point on a sparsification curve. The process is repeated removing more and more edges up to the point where the network connectivity is lost. Note that when an external referee is at hand, a pure prediction process is possible and a performance point without link removal can be reported.
The above performance evaluation methodologies are repeatedly used throughout this document. For simplicity, we refer to them as follows:

- For networks that fall into *Situation B*, pure Precision is measured through the Prune-rediscover Performance Evaluation (PRPE).
- For networks that fall into *Situation C*, AUP is measured through the External Referee Performance Evaluation (ERPE).
- When network sparsification is employed for networks that fall into *Situation B*, each point of the sparsification curve corresponds to average Precision and the evaluation process is called Prune-rediscover Performance Evaluation by Sparsification (PRPES).
- When network sparsification is employed for networks that fall into *Situation C*, each point of the sparsification curve corresponds to average AUP and the evaluation process is called External Referee Performance Evaluation by Sparsification (ERPES).

### 2.2.3 Important Comments about the Evaluation Process by Gene Ontology

The use of all the GO terms in the ERPE, Molecular Function (MF), Biological Process (BP) and Cellular Compartment (CC), has been motivated by the guilt-by-association principle, which states that interacting proteins and genes are likely to be involved in similar biological functions or be located in the same cellular compartment (Oliver, 2000): 63% of interacting proteins have at least one common function, ~76% of them share a cellular compartment in yeast (Saito et al., 2002, 2003) and genes involved in similar bio-processes are enriched for GIs (Costanzo et al., 2010), furthermore there are several studies that have employed the GO to reduce the amount of false positive PPIs resulting from computational predictions (Mahdavi and Lin, 2007; Zeng et al., 2008) and to detect functional interaction patterns (Turanalp and Can, 2008). In particular, Zeng and colleagues showed that GO semantic similarity is a very strong tool to discriminate between true and spurious PPIs. In their experiments they were able to detect 99.61% FPs and 83.03% TPs by means of the GO only. Yet another study showed that an important amount of human embryonic stem cell PPIs share GO terms (Zuo et al., 2009), which is another confirmation of the validity of the ERPE.

In summary, protein and gene pairs that are involved in the same BP, perform similar MF or are located in the same CC are very likely to occur (Saito et al., 2002, 2003; Chen et al., 2005, 2006a,b; You et al., 2010). All proteins or genes from a PPIN or GIN are annotated with their
GO terms and the similarity between them is measured using the R package GOSemSim (Yu et al., 2010) and the Wang GO Semantic Similarity method (Wang et al., 2007). The GOSemSim function used in this document takes as input the list of proteins or genes that form the PPIN or GIN, annotates them, computes the Wang GO semantic similarity between proteins or genes and outputs a matrix whose entries are the GO similarities for every protein or gene pair. There are several GO semantic similarity indices (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999) that were originally developed for natural language taxonomies and it is not known if they are 100% suitable for GO. Wang’s measure was created from the ground up, especially for the GO and its values (with a range between 0 if there is not information in the GO for one or both proteins/genes or if they do not have a similar MF, BP or CC and 1 if the proteins/genes share one or more identical GO terms) are more consistent with the human perspective and the manual gene clustering into GO terms (Wang et al., 2007). Whenever the Wang similarity is in the high end of the range, the proteins or genes being analysed can be considered analogous in their MF, BP or CC (Wang et al., 2007). Thus, as suggested in previous studies (Chen et al., 2005, 2006a,b; You et al., 2010), only those pairs with Wang similarity above 0.5 were considered to be biologically relevant (see Figure 2.6).

Figure 2.6: The list of proteins/genes that form the biological network under study is input to the R package GOSemSim and Wang’s GO Semantic Similarity is used to generate three matrices, each one measuring the pairwise similarity between genes in the three GO categories: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). The maximum between these three matrices is taken to generate a final matrix in which the relevance test is carried out: only those gene pairs with similarities ≥ 0.5 pass the test (similarities in red are examples of the opposite situation). Notice that the MAX operator is equivalent to an OR gate, in the sense that if one ontology has a very low value for a pair of genes, it can be balanced by another ontology.
Chapter 3

From Link Prediction in Brain Connectomes and Protein Interactomes to the Local Community Paradigm in Complex Networks

This chapter was published as:


Abstract

Growth and remodelling impact the network topology of complex systems, yet a general theory explaining how new links arise between existing nodes has been lacking, and little is known about the topological properties that facilitate link-prediction. Here we investigate the extent to which the connectivity evolution of a network might be predicted by mere topological features. We show how a link/community-based strategy triggers substantial prediction improvements because it accounts for the singular topology of several real networks organised in multiple local communities - a tendency here named Local Community Paradigm (LCP). We observe that LCP networks are mainly formed by weak interactions
and characterise heterogeneous and dynamic systems that use self-organisation as a major adaptation strategy. These systems seem designed for global delivery of information and processing via multiple local modules. Conversely, non-LCP networks have steady architectures formed by strong interactions, and seem designed for systems in which information/energy storage is crucial.

Digging into the properties of complex networks is fundamental for the definition of general paradigms of complex systems, in which complexity is distinctively generated by the topological integration of many interacting parts. Although several network formation models have been proposed (see Chapter 2), topological evolution when exclusively new links are added to the network has yet to be congruously formalised, and it connects with a practical and front-line issue, namely the link prediction problem (Liben-Nowell and Kleinberg, 2007). Many applications have to predict new links in large and sparse complex networks merely with the knowledge of network topology, and new solutions could impact both science and engineering positively. Meanwhile, link prediction reflects the extent to which the evolution of a network might be modelled on the basis of topological features intrinsic to the network itself (Liben-Nowell and Kleinberg, 2007; Wang et al., 2012). In this chapter three central questions that stem naturally from the state of the art are inspected:

1. How can the prediction of new links on the exclusive basis of network topology (topological link prediction) be improved?

2. Is it possible to define a network paradigm that enables the prediction of new links in many real networks, on the assumption that their topologies are shaped in accordance with a single, general link-growing process (as postulated in the paradigm)?

3. Is it possible to conjecture any connection between a topological network paradigm and a class of physical systems?

### 3.1 A Local Community Approach to Link Prediction

A myriad of complicated techniques for topological link prediction with two or more parameters to tune have been proposed (Liben-Nowell and Kleinberg, 2007; Lü and Zhou, 2011; Getoor and Diehl, 2005), even inspired by concepts that originate from statistical mechanics and theory of disordered systems. Nevertheless, such elegant techniques are at the moment mere proof of concepts rather than concrete methods to apply on real problems. Apart from the problem to tune the parameters in an unsupervised framework, the greatest obstacle is their prohibitive computational time, which
in practice relegates their application to toy networks of very small dimensions (few hundreds of
nodes), in comparison to the giant networks used in real problems. For these reasons, in order
to answer the first question, we preferred to focus our attention on efficient and parameter-free
Node-Neighbourhood-based approaches (Liben-Nowell and Kleinberg, 2007), which are commonly
employed in both research and application and whose design is inspired by the main patterns char-
Node-Neighbourhood-based approaches assign a likelihood score to each pair of non-connected nodes
(candidate links), and then produce a ranked list in decreasing order to advocate candidate interac-
tions. The Common Neighbours (CN) index (Newman, 2001a) is the progenitor of these methods
and follows the natural intuition that the likelihood that two nodes \( x \) and \( y \) interact increases if
their sets of first, node neighbours \( \Gamma(x) \) and \( \Gamma(y) \) overlap substantially:
\[
CN(x, y) = |\Gamma(x) \cap \Gamma(y)|.
\]
The other indices are often a variation or generalisation of CN (Liben-Nowell and Kleinberg, 2007):
Jaccard (JC) is a normalisation of CN (Jaccard, 1912), Adamic & Adar (AA) (Adamic and Adar,
2003) and Resource Allocation (RA) (Zhou et al., 2009) give more importance to CNs with low
degree, while Preferential Attachment (PA) (Newman, 2001a) is the degree product of nodes \( x \) and
\( y \) (for their formulae see Table 3.1). In contrast to the existing methodologies, which are focused on
groups of common nodes and their node neighbours, a strategic shift from nodes to links is embraced
here that represents a new way to treat complex networks (Ahn et al., 2010). In particular, Ahn et
al. reconceived communities as groups of links rather than as mere groups of nodes, and proposed a
link-based approach for graph partition that outperforms node-based techniques (Ahn et al., 2010).
However, the potential of the link/community viewpoint is still largely unexplored, and we sensed
that this strategy might also be successful for the design of novel link prediction indices; our interest
is to introduce a new philosophy in the formulation of parameter-free/neighbourhood-based indices,
advocating a shift in perspective from nodes to links, and in particular from nodes to community
links. The Cannistraci-Alanis-Ravasi (CAR) index (see Figure 3.1 for definition and examples and
Table 3.1) stems from the fusion of the old node-based and new link-based perspectives. CAR sug-
gests that two nodes are more likely to link together if their common, first neighbours are members
of a strongly inner-linked cohort, named a Local Community (LC), and the LC’s internal links are
called Local Community Links (LCL) (see Figure 3.1). A consequence of this formulation is that, in
respect to CN, CAR offers more discriminative resolution between candidate links characterised by
the same number of common, first neighbours (Figure 3.1), and this boosting in resolution is clearly
derived by the use of the link/community perspective, which is introduced adopting LCL in CAR’s
formula (see Figure 3.1 and Table 3.1). To demonstrate that the formulation of CAR is not a banal
trick, but the precise introduction of a link/community strategy in designing neighbourhood-based indices, we propose for each of the above mentioned classical methods a CAR-based variant. If LCL is seen as an index enhancer, it can be plugged into PA, AA, RA and JC indices so that these techniques also shift to the link/community perspective (thus defining the CAR variant of Preferential Attachment (CPA), the CAR variant of Adamic & Adar (CAA), the CAR variant of Resource Allocation (CRA) and the CAR variant of Jaccard (CJC)) and in the rest of this chapter we will extensively prove the value of this idea. Table 3.1 provides the formulae of the considered classical indices and their respective CAR-based variations.

Table 3.1: Formulae for the classical, CAR-based and bio-inspired neighbourhood-based techniques

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical</td>
<td>CN</td>
<td>$CN(x, y) =</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>$PA(x, y) =</td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>$AA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{</td>
</tr>
<tr>
<td></td>
<td>RA</td>
<td>$RA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{</td>
</tr>
<tr>
<td></td>
<td>JC</td>
<td>$JC(x, y) = \frac{</td>
</tr>
<tr>
<td>CAR-based</td>
<td>CAR</td>
<td>$CAR(x, y) = CN(x, y) \cdot LCL(x, y) = CN(x, y) \cdot \sum_{z \in \Gamma(x) \cap \Gamma(y)} \gamma(z)$</td>
</tr>
<tr>
<td></td>
<td>CPA</td>
<td>$CPA(x, y) = e_x e_y + e_x CAR(x, y) + e_y CAR(x, y) + CAR(x, y)^2$</td>
</tr>
<tr>
<td></td>
<td>CAA</td>
<td>$CAA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\gamma(z)}{</td>
</tr>
<tr>
<td></td>
<td>CRA</td>
<td>$CRA(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{\gamma(z)}{</td>
</tr>
<tr>
<td></td>
<td>CJC</td>
<td>$CJC(x, y) = \frac{CAR(x, y)}{</td>
</tr>
<tr>
<td>Bio-inspired</td>
<td>IG1</td>
<td>$IG1(x, y) = 1 + \text{the number of nodes that interact with } x \text{ and } y \text{ but nothing else in the network.}$</td>
</tr>
<tr>
<td></td>
<td>CDD</td>
<td>$CDD(x, y) = \frac{</td>
</tr>
<tr>
<td></td>
<td>FSW</td>
<td>$FSW(x, y) = \frac{2</td>
</tr>
<tr>
<td></td>
<td>ACDD</td>
<td>$ACDD(x, y) = \frac{</td>
</tr>
</tbody>
</table>

$x$ and $y$ are network nodes; $s$ is a neighbour common to nodes $x$ and $y$; $|\Gamma(x)|$ refers to the set of neighbours of $x$; $|\Gamma(x)|$ refers to the cardinality of set $\Gamma(x)$, which is equivalent to the degree of $x$; $\gamma(s)$ refers to the subset of neighbours of $s$ that are also common neighbours of $x$ and $y$, thus $\gamma(s)$ is the local community degree of $s$; $e_s$ refers to the external degree of $x$, computed considering the neighbours of $x$ that are not common neighbours of $x$ and $y$; $i_x$ refers to the cardinality of set $\Gamma(x)$, which is equivalent to the number of common neighbours shared by $x$ and $y$; $\Gamma(x)$ is the set of neighbours of $x$, including $x$; operation $\Delta$ corresponds to the symmetric difference, i.e. $\Gamma(x) \Delta \Gamma(y) = (\Gamma(x) \cup \Gamma(y)) - (\Gamma(x) \cap \Gamma(y))$; $\lambda_{xy} = \max(0, n_{avg} - |\Gamma(x)|)$ where $n_{avg}$ is the average node degree in the network. Note that the networks under study cannot contain isolated nodes in order to avoid empty unions in, for example, JC or CJC.

### 3.2 Link Prediction in Brain Connectomes

To test the performance of CAR-based indices against the classical indices, we propose an innovative benchmark problem that originates from the neurosciences. The notion of connectome (Sporns et al., 2005) suggests that the brain circuitry can be outlined as a network of neurons connected by links,
Figure 3.1: When a link or a node directly interacts with a seed node (black nodes), it belongs to the first-level neighbourhood and conveys first-level topological information. Conversely, a link or a node that interacts with the first-level neighbourhood conveys second-level information. Second-level information is valuable and its use can significantly improve topological link prediction, but unfortunately it is also very noisy, and for this reason difficult to integrate with the first-level information. CAR is designed to capture and filter meaningful second-level information by exploiting common, first neighbours. The topological information conveyed by the internal links between common, first neighbours is valuable second-level information. Indeed, the more the common, first neighbours reciprocally interact, the more they represent a local community, which in turn encompasses the two seed nodes and thus increases their interaction likelihood. Here we introduce the idea that the likelihood of an interaction is a function of both the number of common, first neighbours (as in the CN index) and of the number of links between them (local community links), as expressed in the formula of CAR.

which are synapses. Several neuroscientific studies have demonstrated that certain forms of learning consist of synaptic modifications (Corti et al., 2008), while the number of neurons remains basically unaltered (Ziv and Ahissar, 2009). A first model of this process was proposed in 1949 by Hebb and subsequently used in Hopfield’s model of associative memory (Ziv and Ahissar, 2009). The Hebbian learning theory assumes that different engrams (memory traces) are memorised by the differing neuronal cohorts that are co-activated within a given synaptic network. It is also termed cell-assembly hypothesis (Ziv and Ahissar, 2009), because these neuron-assemblies are shaped during engram formation by a re-tuning of the strengths of all the adjacent synapses extant in the network. Recent studies (Ziv and Ahissar, 2009), however, demonstrated that learning new motor or sensory tasks is associated with the development of new and non-overlapping sets of persistent synapses. Ziv et al. commented on these findings by suggesting that synapse-assemblies, rather than cell-assemblies, might be viewed as the elementary entities of stored memories (Ziv and Ahissar, 2009), which in turn amounts to a link/community reinterpretation of learning and memory processes in neuroplasticity. We thought to test this synapse-assembly hypothesis in a computational framework, where the formation of new synapses during learning might also be influenced by the local synapse communities extant in the network. Consequently, we proposed to use the topological prediction of new links in a brain-connectome to model the part of the growth and remodelling process that is conditioned by the connectome topology during synaptic formation. Figures 3.2A and 3.2B show
the prediction power of CAR-based indices versus that of the classical node-neighbourhood-based indices. We considered the first and (to date) only available in-vivo single neuron connectome (obtained by means of in vivo two-photon calcium imaging in combination with large-scale electron microscopy) that reports mouse primary visual cortex (layers 1, 2/3 and upper 4, Network 17 in Table A.1 in Appendix A) synaptic connections between neurons (Bock et al., 2011). The evaluation was performed using PRPES (see Section 2.2.2 in Chapter 2). CAR-based indices proved to be the best methods and performed significantly better (p-value = 0.05) than the classical node-based models such as CN in the destruction and re-prediction of up to 50% of the original connectome synapses (Figure 3.2A,B). Interestingly, the 50% deletion level seems to be a critical value. Synaptic deletion beyond this limit induced a strong reduction in performance in all indices; nevertheless CPA and PA were notably the only indices whose prediction remained stable in the given condition - with CPA performing also as the best index overall. This finding suggests that for synaptic deletion greater than 50%, a significant quantity of connectome nodes is also deleted causing network disconnection into different components and network clustering coefficient reduction. The consequence is a general drop in performance for all the non-PA based indices (Feng et al., 2012) that, as a result, assign a score of zero to candidate links whose seed nodes are located in different components. On the contrary PA’s formulation is based on node degree products only, thus its performance is less affected. By definition, this result also corroborates the reliability of the analysed connectome and of the experimental approach used to detect it. Additionally, the result confirms that the design of our computational experiment was correct, and illustrates, as expected, that initial link growth is dominated by a self-organising process of node preferential attachment (Barabási and Albert, 1999).

Furthermore, we considered two different in-silico connectomes, the Macaque cortical connectome and the C. elegans frontal ganglia connectome (Networks 16 and 13 respectively in Table A.1 in Appendix A), which previous studies had assembled in order to merge partial information obtained from disparate literature and database sources (Kaiser and Hilgetag, 2006). Here, in particular, we concentrated our attention on performances when random deletion of only 10% of links is applied, to ensure that the community structure (if present) is still preserved, and in order to objectively compare the proposed community-based indices versus the classical ones. Although CAR-based indices performed largely better than random predictors in these two simulations too (see Figure 3.2C), their superiority over the classical indices was diminished but still statistically significant in the Macaque connectome, and comparable to the classical methods in the C. elegans connectome. This finding may be due to: i) the unreliability and paucity of the two different connectomes, which do not directly derive from in-vivo investigations; ii) the absence of sufficient topological information that is
generated by the neuroplasticity process of learning. In particular, the Macaque cortical connectome (aka corticocortical network) is incomplete and differs in anatomical scale: the links are long-range projections between cortical areas and subareas within one hemisphere of the primate brain (Kaiser and Hilgetag, 2006). Although the C. elegans frontal ganglia connectome is reconstructed at the synaptic level, in-silico merging causes it to be similarly unreliable. Significantly, it has recently been emphasised how published wiring diagrams for C. elegans are neither accurate nor complete and self-consistent (Varshney et al., 2011). Nevertheless, the process of synaptic formation (or, to be precise, at least that part of the process that depends on connectome topology) between existing neurons seems to be more accurately predicted by our local community based models than by PA and the other classical neighbourhood-based approaches, and this is the first substantial result we have obtained for link prediction in computational network biology. A straightforward implication of this finding might open up new avenues in computational learning models, in which the Hebbian theory might be complemented by an epitopological learning theory, whereby engram formation stems from the growth of additional synapses within local communities of pre-existing synapses. A second important implication is that our local community based interpretation of synaptic learning could also be a computational evidence to support the recently proposed unifying hypothesis on the autistic brain called Intense World Syndrome (IWS), which sustains that the core pathology of the autistic brain is hyper-reactivity and hyper-plasticity of local neuronal circuits (Markram et al., 2007). Markram et al. advocate that such excessive neuronal processing and plasticity in circumscribed circuits lead to excessive neuronal learning: hyper-perception, hyper-attention, and hyper-memory, which may lie at the heart of most autistic symptoms (Markram et al., 2007). Thus, we speculate that our local community based interpretation of synaptic learning might be also adopted as part of a computational model to simulate some of the basic mechanisms that are assumed by the IWS to explain how the autistic person is an individual with far above average capabilities (due to enhanced perception, attention and memory) (Markram et al., 2007).

3.3 Link Prediction in Social and Ecological Networks

CAR-based indices were in general significantly better than classical ones also when applied to diverse kinds of complex networks such as the human/animal social networks (Figure 3.2C, Networks 26, 20 and 30 in Table A.1 in Appendix A). Only the food webs represented a separate case (Networks 19 and 31 in Table A.1 in Appendix A). For the food web of Tuesday Lake (excluding the performance of PA) the CAR-based indices were the only ones that performed better than random, and in general
significantly better than the classical indices. For the food web of Grassland species (excluding the performance of JC) we registered a paradoxical behaviour, since it is the only network where the performance of classical methods is significantly better than CAR-based indices (the topological motivation of this result will be further investigated in the next section). There is, nevertheless, an explanation for these opposite behaviours based on the structure, characteristics and members of these two ecosystems. While the Tuesday Lake is regarded as one of the most complete and organised webs in the field of ecology (Cohen et al., 2009; Yiqi, 2005), the Grassland web is considered highly unsaturated (Tscharntke et al., 2001). Moreover, the Grassland web has a linear structure that in ecology is known as the cascade model: a serial trophic organisation, where the bigger organism eats the smaller. In contrast, the Tuesday Lake web has a pyramidal organisation that differs from the cascade model in the distribution of links between basal, intermediate, and top communities of species (Yiqi, 2005). Thus, the Grassland web has a highly sparse and almost linear connectivity, which makes it hard for local communities to emerge; on the contrary, the Tuesday Lake web has a more densely connected and organised structure, which is actually related with the presence of local communities at different levels of the pyramidal organisation. As a further investigation, we compared CAR-based indices with two sophisticated statistical inference techniques: the Hierarchical Random Graph (Clauset et al., 2008) and the Stochastic Block Model (Guimerà and Sales-Pardo, 2009), known as maximum likelihood approaches. The comparison followed the same procedure shown in Figure 3.2C, and we did not detect any remarkable increase in performance for these advanced algorithms in comparison to the family of CAR-based indices.

3.4 Link Prediction in Protein Interactomes

A second aspect of the link prediction problem regards the inference of missing interactions from an observed network (Liben-Nowell and Kleinberg, 2007). In varying disciplines, a network is constructed on the basis of experiments, and for at least two reasons some links might not be observable: i) by nature, the links are not directly detectable; ii) the experiments and/or the execution time are very expensive. The problem of observability affects systems biology, a discipline in which the topological prediction of novel interactions in protein networks (interactomes) is particularly useful, especially and specifically when other types of information, such as biological prior knowledge, are not available (You et al., 2010). This problem differs from the original formulation (link prediction in network evolution over time) discussed above, and the class of link prediction indices (named bio-inspired indices, see Table 3.1) currently employed in systems biology derives from methods
invented to infer similar attributes between adjacent protein nodes (You et al., 2010): the Interaction Generality 1 (IG1) index (Saito et al., 2002) originates from phenomenological evidences in experimental protein interaction detection; the Czekanowski-Dice Dissimilarity (CDD), its adjusted version Adjusted CDD (ACDD) (Brun et al., 2003; Liu et al., 2009; Alanis-Lobato et al., 2013) and the Functional Similarity Weight (FSW) (Chua et al., 2006) stem from methods for protein functional prediction in interactomes; and Isomap (ISO) (Kuchaiev et al., 2009; You et al., 2010) is based on high-dimensional properties and embedding of protein networks. The strategy used for evaluation differs too (You et al., 2010), in this case the ERPE is used (see Section 2.2.2 in Chapter 2). Yeast networks are the preferred benchmark, because of the large amount of information available for yeast in terms of both detected interactions and GO associations (You et al., 2010; Ahn et al., 2010). We accordingly re-analysed three different, independently produced, yeast networks that had been used in previous studies for performance testing (Networks 1, 2 and 3 in Table A.1 in Appendix A). CAR and CAR-based indices not only substantially outperformed the other methods in the three networks (they are the only indices to attain AUP always equal or higher than 0.85, Figure 3.3B-D), but also proved to be the most efficient, in that they simultaneously provided consistent robustness (Figure 3.3E-F) and the lowest maximum-computational time.

We besides investigated the extent to which the topological link prediction could be practically useful. To this end, we proposed an in-silico validation and tested the 100 best ranked interactions, both for CAR and for FSW (reference method for the bio-inspired indices), with the STRING database (Szklarczyk et al., 2011). Considered to be the most complete and reliable database of PPIs, STRING integrates multiple sources of information and provides a confidence score for each interaction (Szklarczyk et al., 2011). Once again, CAR confirmed the benefit of using a link/community strategy and performed impressively during this validation test, both in each single network evaluation (Figure 3.4) and in the general evaluation of robustness (Figure 3.3F). Summarising, we observe that in general using the CAR-trick to convert the classical methods to the new link-community perspective, generates a significant increase in link prediction performance in several complex networks, suggesting that a new family of community-based link predictors is here proposed. In addition, we notice that whenever the community structure of the network (if present) is preserved, it can be exploited to achieve more accurate predictions using the new CAR-based family of link predictors.
Figure 3.2: A In-vivo single neuron connectome: mouse visual cortex neuro-synaptic connections. PRPES was used for this analysis with 1000 different random sets of links removed at each sparsification level (ranging from 10% to 90% of removed synapses), in practice mean precision and standard error were considered at each stage. To assess deviation from the mean random predictor performance, the Prediction Power was computed in dB: 10·log_{10} \frac{\text{Precision}}{\text{Random predictor}}. Thus, considering the different levels of sparsification, a Prediction Power Curve was generated and the area under this curve (AUPPC) was adopted as a comprehensive measure of performance. Notice that the progressive removal of links from the original topology made the average LCP-corr drop down to the point where it is almost 0. Since deletion of more than 50% synapses caused node isolation and the disappearance of local communities, the AUPPC for the plots in panel A was computed considering only half of the experimental range (10% to 50%). CAR-based indices provided an average AUPPC of more than 50, and the performance of the best approach (CPA), represented a 186% improvement over the lowest performing technique (PA). JC was not reported.

B The performance of CAR-based and classical predictors was also studied over 8 different networks when 10% of the links were removed 1000 times uniformly at random (which ensures community preservation and a fair benchmarking, see average LCP-corr values, \(L_i\), below each plot). The difference in performance between the CAR-based (in red) and the classical techniques (in black) was statistically significant every time the studied network exhibited a LCP structure (see p-values below each plot computed as a permutation test with 1000 sampling realizations in which maximum likelihood techniques, in blue, were not considered).
Figure 3.3: A Precision curve of each technique by ascertaining whether the top-ranked candidate links had meaningful association in GO in Network 1. For clarity, only CAR-based, classical and bio-inspired prediction techniques are shown. B-D The AUP summarises the general performance for each technique in the three analysed networks: Network 1, 2 and 3 in Table A.1. E GO performance robustness (PR) is computed as the minimum AUP amongst the networks and used for comparison between CAR and FSW. F In-silico STRING validation: the first 100 best interactions for each method are tested in STRING. The minimum number of verified-interactions amongst the three networks represents the minimum precision, which is a measure of PR for comparison - with regard to STRING - between CAR and FSW.
Figure 3.4: The dashed-red lines depict CAR’s GO precision curve for the first 100 candidate PPIs. The top-100 candidate interactions proposed by CAR were intersected with the STRING database. This database provides each interacting pair with a confidence value. The sub-networks formed by these candidate links are displayed, and the protein pairs found in STRING are shown in red. Mean STRING confidence and GO semantic similarity are also reported along with their standard deviation.

3.5 The Local Community Paradigm in Real Networks

A necessary condition for the application of CAR’s indices is that, during its growth, the network in question evolves in accordance with a general process, one of whose distinctive features is the development of diverse, overlapping and hierarchically organised local communities (Ahn et al., 2010). We defined this form of topological self-organisation as a LCP and we therefore propose the LCP Decomposition Plot (LCP-DP) to visualise and investigate the effect of LCP on network topology. The LCP-DP is a form of network decomposition because each real link in the network is plotted in a bi-dimensional space according both to its number of CNs (reported on the x-axis) and to the respective number of LCLs (reported on the y-axis). More specifically, given that the number of LCLs is a squared function of the CNs, we found it more convenient to report the square root of LCL on the y-axis, so as to linearise the visualisation. The result of this decomposition is a plot that offers a link-based visualisation of the analysed network and provides information on the presence and size of its local communities.

Figure 3.5A shows the LCP-DP for the three previously employed yeast PPINs. Surprisingly, we discovered that Network 3 differs from the other two, whose LCP-DP patterns resemble each other (Figure 3.5). In particular, the maximal local community size for Network 3 is around 20 CNs, while for the other two networks the maximal local community size is about 60 CNs. Considering that all the networks have a comparable number of nodes and links, and that the maximal size of
local communities in Network 3 is $1/3$ of that in the other networks, we can conclude that Network 3 is more strongly characterised by small local communities. The fact that multiple small local communities in a large and sparse network most probably do not overlap with each other explains why all the indices (CAR-based, classical and bio-inspired) performed in a comparable manner and with better results in Network 3 (Figure 3.3D) than in the other two networks (Figure 3.3B-C), where instead the CAR-based performed significantly better than the others. Following a similar rationale, we observe that the *C. elegans* connectome (Figure 3.5B, left panel), although larger (more nodes) and sparser (less links) than the Macaque connectome (Figure 3.5B, right panel), is more strongly characterised by small local communities, and this explains why the performances of the CAR-based and classical indices were not significantly different and were overall higher in this connectome than in the others (almost all the indices were between 12.5 and 13.5 dB, Figure 3.2C).

From the topological similarity between Network 3 and the *C. elegans* connectome, evidenced by using the LCP-DP, we can now infer that a very sparse and clustered network topology, characterised by the presence of multiple small local communities that do not overlap with each other, improves link prediction in general and minimise the difference between CAR-based and classical indices. Conversely, in the extreme and opposite case of very densely connected networks, the occurrence of large and indistinct communities - which most likely overlap reciprocally and encapsulate the small local communities - erases the community distinction that is fundamental for the efficiency of CAR and of prediction methods in general. This is easily demonstrable because the more a network tends to the ideal case of a fully-connected network, the more it converges towards a unique and large single community. On the other hand, in Figure 3.5C we clearly demonstrate that when we randomly sparsify the networks in question, using the ERPES procedure (see Section 2.2.2 in Chapter 2), CAR, whenever the network community structure is preserved, is more efficient and robust than the other indices in prediction of candidate protein interactions.

A second discovery, which emerges from the LCP-DP, is the strong correlation between the two variables CN and LCL, which we call the LCP-corr, and which might be interpreted as a typical feature of LCP networks (Figure 3.5A-B). More specifically, the LCP-corr is defined as the Pearson correlation coefficient between the variables CN and LCL as plotted in the LCP-DP. Looking at the curve of average LCP-corr values - computed for the Mouse connectome configurations, in the randomly destroyed synapses experiment of Figure 3.2A - we observe that the LCP-corr decreases significantly between 10% ($\text{LCP-corr} = 0.60$) and 20% ($\text{LCP-corr} = 0.31$) of removed synapses, and this is a confirmation that the choice to focus our attention on the comparison of the predictors when only 10% of the links are randomly removed, was a correct procedure to preserve the community
structure. In fact the average LCP-corr values of the 8 partially destroyed complex networks adopted in Figure 3.2C are always higher (except for the Grassland web) than LCP-corr = 0.5. On the other hand, also the examples provided in Figure 3.5A (Network 1 has LCP-corr = 0.92; Network 2 has LCP-corr = 0.95; Network 3 has LCP-corr = 0.90) might lead one to suspect that large LCP-corr coefficients are invariably associated with the occurrence of LCP, and small LCP-corr coefficients with LCP’s non-occurrence. Therefore, to investigate the extent to which this hypothesis is generally valid, we considered a total of 48 networks from differing fields (18 biological, 2 ecological, 8 social, 10 atomic, 1 power grid and 9 roadmaps - see Table A.1 in Appendix A).

As evidenced in Figure 3.6, where we juxtapose a few paradigmatic examples, the scenery is more intriguing than we had expected: we found that the region of LCP-corrs between 0.8 and 0.4 represents a threshold (a sort of intermediate region) that distinguishes networks that are characterised by LCP from those that are not. We want to acknowledge that also Lancichinetti et al. (Lancichinetti et al., 2010) went close to the formalisation of the LCP paradigm and the observation of the dichotomy between LCP and non-LCP networks, i.e. networks with very high and very low LCP-corrs respectively. In fact, they provided a systematic empirical analysis of the statistical properties of communities in diverse types of large complex networks (Lancichinetti et al., 2010) and evidenced that the mesoscopic organisation of networks of the same category is extraordinarily similar, concluding that: although the community size distributions are always wide, certain categories of networks consist mainly of tree-like communities, while others have denser modules.

If we investigate the networks at the extremity of the intermediate region, we find that the power grid network is a borderline case, one that can be considered neither as LCP nor as non-LCP, and this finding has strong significance. It has been proved that power grid networks, which are human-designed and propagate electricity rather than information, are neither scale-free nor clustered (Boguñá et al., 2009), that they have homogenous topology and that they are easy to control (Liu et al., 2011). The other interesting borderline case is the Grassland species food web that with LCP-corr of 0.42 is closer to non-LCP networks, and this can motivate its paradoxical behaviour detected in Figure 3.2C (after 10% of random link removal, LCP-corr = 0.41) and discussed above.

On the other hand, the upper-bound case is represented by the air transportation network, which shows a LCP-corr = 0.99 (Figure 3.6 left side, American/Canadian flight map). We mined the literature to find an explanation for the contradiction between the result obtained for the air transportation and for the road transportation networks (Figure 3.6 right side, LCP-corr ranges from 0 to 0.16). Guimerà et al. provided a possible answer (Guimerà et al., 2005), which we shall
We find that the worldwide air transportation network is a scale-free small-world network. In contrast to the prediction of scale-free network models, however, we find that the most connected cities are not necessarily the most central, resulting in anomalous values of the centrality. We demonstrate that these anomalies arise because of the multi-community structure of the network. We identify the communities in the air transportation network and show that the community structure cannot be explained solely based on geographical constraints and that geopolitical considerations have to be taken into account.

Although Guimerà et al. did not formalise any paradigm, we recognise an ante litteram discovery of the LCP in their intuition of the need for a new explanation (besides the scale-free and the small-world paradigms) to characterise the topology of air transportation networks more precisely. Meanwhile, the social (geopolitical) interpretation of the flight networks clarifies the mismatch with the road networks, which are shaped more by geographic constraints, and further justifies our choice to allocate the flight map amongst the social networks (Figure 3.6 and Table A.1).

The conclusions of Guimerà et al. - and the need to formalise them in a paradigm such as the LCP - appear clearer if interpreted in the light shed by two theories recently posited, one by Boguñá et al. on the navigability of complex networks (Boguñá et al., 2009), and the other by Liu et al. on the controllability of the same (Liu et al., 2011). Boguñá et al. illustrated the greedy-routing of information in a network through an example of passenger air-travel. They showed how greedy behaviour takes a passenger from a small airport to larger hub airports, which significantly reduces the distance to the destination (zoom-out, coarse-grained search), and how hubs are thus crucial for global delivery. Once a hub near the destination is reached, hubs are not needed anymore because a less-connected neighbouring airport can take the passenger to the desired city (zoom-in, fined-grained search). Therefore, some particular non-hub nodes centred in local modules are fundamental for the local processing of the general function implemented on the network. On the other hand, Liu et al. found that low-degree nodes (and, counter-intuitively, not hubs) play the most important part in the full controllability of complex networks, and this finding fits with Boguñá’s, because hub nodes may be viewed as collectors and distributors of information: if one of the near destination hubs is unavailable, another one can compensate. They act as intermediary relaxing points that avoid bottlenecks and direct data to more important nodes (driver or processor nodes) that accomplish a dedicated function within local modules, where they naturally assume a position of centrality. Similar conclusions are discussed by Lancichinetti et al. as well (Lancichinetti et al., 2010).
We also introduced the analysis of atomic-level networks. On the right side of Figure 3.6, we show that many organic molecules (their crystals, reticula and lattices) and secondary biological structures have a network topology that is non-LCP. In a similar vein, we investigated the network topology in the tertiary biological structures of proteins: said structures are generated by various classes of non-covalent interactions that occur between protein residues and determine the typical protein folding conformation. Residue Interaction Networks (RINs) have recently been used to describe the protein three-dimensional structure as a graph, where nodes represent residues and edges physico-chemical interactions, e.g. hydrogen bonds or van der Waals contacts (Martin et al., 2011). Various topological properties have been calculated over RINs and have been correlated with differing aspects of protein structure and function (Martin et al., 2011). We found that both the hydrogen bond network of human glutathione peroxidase 4 (GPX4, LCP-corr = 0.90) and the van der Waals contact network of human triosephosphateisomerase (TIM barrel, LCP-corr = 0.88) - which were the only RINs available to use (Martin et al., 2011) - are LCP (Figure 3.6 right side and Table A.1). Consequently, we envisage that this finding might be extended beyond the examples here quoted, to confirm that the LCP state detected within 3D protein conformations could be a generic property of some tertiary-biological structures.

3.6 The LCP in Idealised Networks

LCP accounts for community-based structure in the topology of complex networks and extends present knowledge to a degree that, along with the small-world and preferential-attachment paradigms, may enhance our understanding of systems of interacting units, their evolution and self-organisation. The need for, and value of, such a novel paradigm are further investigated in the following examples, which deal with referential idealised models. The networks in Figure 3.7A are two diverse random regular graphs (random graphs for which each node has the same degree) with the same number of nodes. The Barabási-Albert paradigm (which is based on power-law node-degree distributions) does not detect any difference between these two networks, since the degree distribution of a random regular graph corresponds to a single value that is also the fixed node degree. Even the Watts-Strogatz paradigm (small-world) does not detect any difference, since the two networks have identical clustering coefficients and characteristic path lengths. This is further proof that, as clarified in the original article (Watts and Strogatz, 1998), the small-world paradigm is inapplicable to the detection of topological changes that emerge exclusively at a local structural level in the network. Boguñá et al. showed, firstly, that behind each network there is a hidden metric space that is closely related
Figure 3.5: A Using the LCP-DP for investigation of the network topology and for visualisation of the LCP-corr in protein interactomes. B Using the LCP-DP for investigation of the network topology and for visualisation of the LCP-corr in brain connectomes. C Testing the prediction robustness of CAR and the other indices during PPIN sparsification by random link deletion. ERPES is used, with 10 different sets of random edges removed from the network at each sparsification level. Average AUP with standard-error is reported for each link predictor applied. The Area Under the Sparsification Curve was used to compare the performance of the different prediction techniques. This area is an advanced performance measure in the sense that it accounts both for random variations in the original network topology and for differing levels of sparsification. ISO, which is the only embedding-based method, is not plotted for clarity, but its performance is reported.

to the network topology and, secondly, that global mapping may be inferred from local distances (Boguñá et al., 2009). In a way, the same principle is exploited by Isomap, a landmark algorithm for embedding that was mainly designed to visualise the hidden structure of a dataset or of a network
Figure 3.6: Examples of LCP networks (left panel) and non-LCP networks (right panel). The LCP networks we found range from LCP-corr = 0.84 (C. elegans rostral ganglia neuro-synaptic connectome) to 0.99 (American/Canadian flight map). The networks in the center (Power Grid and Grassland species food web) are borderline cases.

in a bi-dimensional space (Tenenbaum et al., 2000). Surprisingly, the embedding in two dimensions by Isomap suggests substantial divergence between the hidden metric spaces of the two random regular networks, as well as a likewise substantial difference in the local topology (Figure 3.7A). The paradigm we propose is the only one that clearly identifies this topological difference within the hidden metric space of the two networks (Figure 3.7A). The LCP-corr is 0.37 for the network on the left, which consists of only two clusters, while for the second network, which consists of three clusters, the LCP-corr is, as expected, higher (LCP-corr = 0.71, Figure 3.7A on the right side).

Idealised networks are very useful to test the generality of a hypothesis in different configurations, which in turn are easily generated artificially through a known model and controlled by certain parameters (Watts and Strogatz, 1998). Such networks are particularly crucial to investigations into the behaviour of a given measurement around a critical region that hosts a transition between differing macroscopical states of the system. We simulated differing ER networks by varying the two parameters used in the model $G(n,p)$, namely network size (number of nodes $n$) and edge probability (each edge is included in the graph with probability $p$ independent from every other edge). We performed 100 realizations for each combination of model parameters and plotted in Figure 3.7B the mean (displayed in the centre) and the standard deviation (displayed on the right) of the LCP-corr values computed for the 100 network realizations. In their original 1960 paper
(Erdős and Rényi, 1960), Erdős and Rényi mathematically described the behaviour of $G(n, p)$ at various values of $p$. One of their main conclusions was that $\frac{\ln(n)}{n}$ represents a critical threshold for the connectedness of $G(n, p)$. In particular, they proved that: i) if $p < \frac{(1-\varepsilon)\ln(n)}{n}$ then a graph in $G(n, p)$ will be disconnected because it certainly contains several isolated vertices (which implies the loss of local network communities); ii) if $p > \frac{(1-\varepsilon)\ln(n)}{n}$ then a graph in $G(n, p)$ will be most likely connected (which implies the preservation of local network communities). Interestingly, we discovered that the theoretical transition region (Figure 3.7B, plot on the left) is accurately detected and visualised by the LCP-corr (Figure 3.7B, in the centre and on the right) in our experiments.

From the mean of the LCP-corr values, as plotted in the centre of Figure 3.7B, we discovered that the majority of all possible generable ER models strongly follow the LCP (black area above the critical region). The necessarily small percentage of non-LCP models appear in the white area under the critical region. This result confirms that under the critical region, several isolated vertices appear in the graph, while above the critical region a topology characterised by several community structures begins to emerge. Meanwhile, prompted by standard deviation analysis (Figure 3.7B on the right) we learn that the generation of graphs with intermediate values for LCP-corr occurs very rarely and is only possible when it coincides with the instability expressed in the critical transition region (dark region in the plot); in the other zones (white part of the plot), standard deviation is close to zero and the LCP-corr value is very stable.

In conclusion, this simulation suggests that the generation of: i) ER networks with LCP-corr higher than 0.80 is very frequent; ii) ER networks with LCP-corr close to 0 are not frequent; iii) ER networks with intermediate LCP-corr are very infrequent. These findings are in line with the ones we obtained for real networks, where the identification of LCP networks was very easy (Figure 3.6, left side), both at diverse physical levels and in differing domains (biological, social, atomic), while the identification of non-LCP networks was difficult and their occurrence was almost exclusively detected at the atomic level (Figure 3.6, right side). Interestingly, except for the power grid (which showed LCP-corr = 0.78, a borderline value) and few others, real networks with intermediate levels of LCP-corr were not identified.

Taken together, these discoveries are important steps towards the answer to the second question formalised in the introduction.
4.8

Figure 3.7: A Isomap embedding of two random regular graphs (Nodes = 10, Degree = 5) with equal clustering coefficients (Cp) and characteristic path lengths (Lp) but different LCP-correlation. Neither the Barabási-Albert nor the Watts-Strogatz paradigms were able to explain the topological difference between these two random regular graphs, which show the same node numbers and the same node degree. The paradigm we propose is the only one that, consistent with Isomap embedding, clearly highlights the community organisation diversity present in the hidden metric spaces of the two networks. B Testing the LCP-corr for differing ER random graphs \( G(n, p) \) by varying the number of nodes \( n \) (from 10 to 500, step 10) and the edge probability \( p \) (from 0.01 to 0.99, step 0.02). For each combination \( (n, p) \) the LCP-corr was evaluated 100 times for calculation of mean and standard deviation. The theoretical critical region computed by Erdős and Rényi in their model (plot on the left) is strikingly detectable both by the mean (plot in the centre) and by the standard-deviation (plot on the right) LCP-corr.

3.7 Discussion

The third question posed in the introduction conjectures a sort of generalised systemic parallelism between the occurrence of certain topological properties (as formalised in a paradigm) and the relevance of some physical properties. In effect, LCP networks (Figure 3.6, left side) are related to dynamic and heterogeneous systems that are characterised by weak interactions (Csermely, 2006) (relatively expensive or relatively strong) that in turn facilitate network evolution and remodelling; these are typical features of social and biological systems as well as, at the atomic level, of tertiary protein structures. According to Liu et al., it should be more difficult to achieve full control of these systems by manipulating a few network driver-nodes (Liu et al., 2011). In contrast, non-LCP networks (Figure 3.6, right side) characterise steady and homogeneous systems that are assembled through strong (often quite expensive) interactions, difficult to erase. Given their homogeneous structure, such systems should be easier to control (Liu et al., 2011). This argument is particularly valid for the power grid, which is not densely connected (network density is proven to increase...
controllability), but whose topology is homogeneous enough to be easily controllable (Liu et al., 2011) - a required specification in human-engineered networks.

The LCP architecture facilitates not only the rapid delivery of information across the various network modules, but also the local processing. On the other hand, the non-LCP architecture is more useful for processes where: i) the storage of information (or energy) is at least as important as its delivery; ii) the cost of creating new interactions is excessive; iii) the creation of a redundant and densely connected system is strategically inadvisable. An emblematic example is the road networks, for which the money and time costs of creating additional roads are very high, and in which a community of strongly connected and crowded links resembles an impractical labyrinth.

While the small-world paradigm treated the main effect of re-modelling in real networks, and the scale-free paradigm offered an innovative view of network growth in terms of node preferential attachment, the LCP is a first attempt to advance a link/community-based interpretation of the \textit{epitopological} learning component that appears in many cognitive, social and evolutionary processes.
Chapter 4

Minimum Curvilinearity to Enhance Topological Prediction of Protein Interactions by Network Embedding

This chapter was published as:


Abstract

Motivation: Most functions within the cell emerge thanks to PPIs, yet experimental determination of PPIs is both expensive and time-consuming. PPINs present significant levels of noise and incompleteness. Predicting interactions using only PPIN topology (topological prediction) is difficult but essential when prior biological knowledge is absent or unreliable.

Methods: Network embedding emphasises the relations between network proteins embedded in a low-dimensional space, in which protein pairs that are closer to each other represent good candidate interactions. To achieve network denoising, which boosts prediction performance, we first applied Minimum Curvilinear Embedding (MCE), and then adopted Shortest Path (SP) in the reduced space to assign likelihood scores to candidate interactions. Furthermore, we introduce (i) a new valid variation of MCE, named non-centred
MCE (ncMCE); (ii) two automatic strategies for selecting the appropriate embedding dimension; and (iii) two new randomised procedures for evaluating predictions.

Results: We compared our method against several unsupervised and supervisedly tuned embedding approaches and node neighbourhood techniques. Despite its computational simplicity, ncMCE-SP was the overall leader, outperforming the current methods in topological link prediction.

Conclusion: Minimum curvilinearity is a valuable non-linear framework that we successfully applied to the embedding of protein networks for the unsupervised prediction of novel PPIs. The rationale for our approach is that biological and evolutionary information is imprinted in the non-linear patterns hidden behind the protein network topology, and can be exploited for predicting new protein links. The predicted PPIs represent good candidates for testing in high-throughput experiments or for exploitation in systems biology tools such as those used for network-based inference and prediction of disease-related functional modules.

Detection of new interactions between proteins is central to modern biology. Its application in protein function prediction, drug delivery control and disease diagnosis has developed alongside a deeper understanding of the processes that occur within the cell. One key task in systems biology is the experimental detection of new PPIs. However, such experiments are time consuming and expensive. Because of this, researchers have developed computational approaches for predicting novel interactions (You et al., 2010), intended also to guide wet lab experiments. The topological prediction of new interactions is a novel and useful option based exclusively on the structural information provided by the PPIN topology. This option for prediction is particularly convenient when the available biological information on the proteins being tested for interaction (seed proteins) is incomplete or unreliable. One of the most efficient approaches is the FSW (Chua et al., 2006). Such method belongs to the large and well-established family of predictors that are referred to as node neighbourhood techniques (see Section 3.1 in Chapter 3), because to assign a likelihood score to any candidate interaction (i.e. a pair of non-connected proteins in the observed PPIN), they rely on the topological properties of the seed proteins’ neighbours. The main problem with these techniques is that their performance is poor when applied to sparse and noisy networks (You et al., 2010).

In 2009, Kuchaiev et al. proposed a method for geometric denoising of PPINs. The algorithm is based on the use of Multidimensional Scaling (MDS) to preserve the SPs between nodes in a low dimensional space. The predicted interactions are scored according to their Euclidean Distance (ED) in the low dimensional space, following the principle that the closer two proteins are, the higher the likelihood that they interact (Kuchaiev et al., 2009). Although it is not explicitly mentioned in the article, the embedding method adopted by Kuchaiev et al. is equivalent to Isomap (Tenenbaum et al.,
In an independent study, You et al. proposed a hybrid strategy based on network embedding to assign likelihood scores to candidate interactions. They exploited the notion that a PPIN - or theoretically, any network - lies on a low dimensional manifold shaped in a high-dimensional space. The shape of the manifold and the associated topology are determined by the constraints imposed on the protein interactions through biological evolution. You et al. used a renowned manifold-embedding algorithm, Isomap, to embed the PPIN in a space of reduced dimensionality (You et al., 2010). Then, they applied FSW to the embedded network (pruned according to a cut-off on the EDs) to assign likelihood scores to the candidate interactions. In general, the embedding strategy offers two advantages: (i) the topological prediction performance is improved even when networks are sparse and noisy; and (ii) the computational time is reduced because the time required for the network embedding is much lower than that required by node neighbourhood techniques for computing the topological properties of each candidate interaction. A disadvantage is that if the network is not a unique connected component, only the largest connected component can be considered for embedding (Kuchaiev et al., 2009).

Here, we introduce several variations of these approaches that, altogether, offer a new solution for topological link prediction by network embedding. The first variation uses MCE (Cannistraci et al., 2010) and its non-centred variant, ncMCE (which is introduced for the first time), to project the network on the reduced dimensionality space. MCE is a parameter-free algorithm designed for the unsupervised exploration of high-dimensional datasets by non-linear dimension reduction (Cannistraci et al., 2010). Recently, MCE ranked first amongst 12 different approaches (evaluated on 10 diverse datasets) in a study on the stage prediction of embryonic stem cell differentiation from genome-wide expression data (Zagar et al., 2011). This proof of power and robustness motivated us to test its performance in the context of PPI prediction by network embedding. In the second variation, we use the SP distances (instead of the ED, as in Kuchaiev et al. and You et al.) over the network embedded in the reduced space to assign the likelihood scores to the candidate interactions. The method proposed here undoubtedly presents a novel combination of steps. We prove that the combination of ncMCE/MCE and SP achieves excellent results, boosting the separation between good and bad candidate links.

### 4.1 Network Datasets

The main datasets analysed in this work comprise four yeast PPINs (Networks 1-4 in Table A.1 in Appendix A). Yeast networks are the preferred benchmark for the reasons already exposed in
Section 3.4, Chapter 3. The PPIs in these datasets are mainly physical interactions, but also include literature-curated and functional links.

4.2 Manifold Learning and Network Embedding

One means to better visualise and interpret high-dimensional data is to assume that it lies on a manifold embedded in a space of high dimensions. In particular, given \( N \) high-dimensional data points \( \{x_1, x_2, \cdots, x_N\} \) with \( x_i \in \mathbb{R}^D \) lying on a manifold embedded in a \( D \)-dimensional space, the objective of Manifold Learning is to find a mapping \( \mathcal{M} : \mathbb{R}^D \rightarrow \mathbb{R}^d \) such that the mapped points \( \{y_1, y_2, \cdots, y_N\} \) with \( y_i \in \mathbb{R}^d \) preserve some of the topological properties of the original manifold.

If it is assumed that PPINs lie on a high, unknown metric space shaped by the biological properties of the proteins that form them, then it is possible to map a network to a reduced space, in which proteins that are close to each other are more likely to interact (Boguñá et al., 2009). This problem is called Network Embedding and is very similar to Manifold Learning, however in this case we lack the coordinates of the points that we want to take to a low dimensional space (further details below).

Let us represent a PPIN by an undirected, unweighted graph \( G = (V, E) \) with a set of \( N = |V| \) nodes and a set of \( L = |E| \) edges, which is a set of 2-element subsets of \( V \). Network Embedding consists of finding a mapping \( \mathcal{M} : V \rightarrow X \), where \( X \) is a set of points \( \{x_1, x_2, \cdots, x_N\} \) with \( x_i \in \mathbb{R}^d \), i.e. each node of \( G \) is assigned a coordinate in a space of \( d \) dimensions such that the original topological properties of the network are preserved in this low dimensional space. Manifold Learning algorithms can be easily modified for Network Embedding, however not all the algorithms that learn manifolds are applicable for this task, only those able to embed a topology starting from a distance or adjacency matrix can be used. The reason is that in this particular work, our goal is to perform the above described Network Embedding with no information other than the network topology itself. In this case, nodes have no properties other than their connections to other nodes and because of this, the embedding process necessarily has to start with the only information available: node distances (i.e. graph SPs) or degrees based on network connectivity.

Embedding techniques that start from proximity matrices \( P \) defined in the high-dimensional space (with entries \( P_{i,j} \) storing the distance between the pair of nodes \( i, j \)), seek to find low-dimensional points such that their pairwise distances in the reduced space are equal or very close to those stored in \( P \). If \( p \) is the proximity matrix in the reduced space (with entries \( p_{i,j} \) storing the distance between the pair of points \( i, j \)), embedding can be seen as the following minimisation problem:
\[
\min \sum_{i<j} (p_{i,j} - p_{i,j})^2
\] (4.2.1)

There are different strategies to solve problem 4.2.1, one of the most common ones being Classical MDS (Abdi, 2007), and it is important to note that link prediction is now possible because, having coordinates for all network nodes in a space of low dimensionality, allows for the assignment of scores to pairs of nodes that are not connected in the original network topology. These scores are associated with distances in the reduced space and the list of non-adjacent pairs of nodes sorted by this measure of link likelihood is the output of the link prediction process by Network Embedding (further details in the following section).

We chose to compare MCE, ncMCE and Isomap against well-established unsupervised and supervisedly tuned manifold embedding algorithms that accept a distance or adjacency matrix as input. The unsupervised embedding techniques considered are Sammon mapping (a type of non-linear MDS) (Sammon, 1969) and two force-based embedding techniques: Stochastic Neighbourhood Embedding (SNE) (Hinton and Roweis, 2002) and tSNE (a variant of SNE) (van der Maaten and Hinton, 2008). The supervisedly tuned techniques are local MDS (Venna and Kaski, 2006) and Neighbour Retrieval Visualiser (NeRV) (Venna et al., 2010). These methods are also force based, but instead of using forces based on kernels (like SNE or tSNE), they use forces based on neighbourhood graphs (Shieh et al., 2011). Both local MDS and NeRV require a parameter \( \lambda \) to be tuned between 0 and 1. In this work, we assessed the performance of these two techniques using values of \( \lambda \) from 0 to 1 in steps of 0.1, and took the low-dimensional coordinates that yielded the best prediction result.

### 4.3 Minimum Curvilinear Embedding

MCE is a parameter-free and time efficient unsupervised algorithm for non-linear dimensionality reduction. It was originally introduced as a new form of non-linear MDS and has proven to be a powerful and robust tool in different applications (Cannistraci et al., 2010; Zagar et al., 2011; Ryu et al., 2012; Moitinho-Silva et al., 2013).

MCE consists of a first step in which a non-linear distance matrix (the minimum curvilinear distance matrix or MC-matrix) is calculated as pair-wise sample distances over the Minimum Spanning Tree (MST) in the feature space. The MST is computed according to the ED or the Correlation distance (or any other preferred distance) in this space. The embedding transformation is then
performed by classical MDS of the MC-matrix.

The fact that the MST is a neighbourhood graph that can well approximate the main network information - offering a general summary of the network topology - has been extensively shown in different applications (Cannistraci et al., 2010; Shaw and Jebara, 2009; Shieh et al., 2011) and in our case, this can be particularly useful for denoising the information present in PPINs. In fact, the False Positive (FP) rate of currently widely used experimental technologies to detect PPIs is significantly high, sometimes exceeding 60% (Kuchaiev et al., 2009). Minimum Curvilinearity (MC) tends to stress local topological distances and dilate large connectivity distances (Cannistraci et al., 2010). A consequence is that the use of MCE for embedding causes a sort of network deformation when the network structure is compressed in a reduced space of just a few dimensions. The deformation augments the separation between nodes far apart in the network topology and maintains or reduces the distances between nearby nodes (Cannistraci et al., 2010). This might be a point of weakness for network visualisation because it stretches the network shape in the reduced space. However, it is a point of strength for link prediction because it generates a non-linear soft-threshold effect - a type of gradual denoising (Cannistraci et al., 2010) based on a non-linear transformation - on the network connectivity distances measured in the reduced space. The soft-threshold discriminates between protein pairs far apart in the original network topology (which earn large score values because they are now connected by enlarged path distances in the embedded space) and candidate links connecting nearby proteins in the network topology (which earn small score values because they maintain or reinforce their topological proximity in the embedded space).

It is important to stress that in the case of Network Embedding, we lack coordinates or properties of the samples (nodes) in the feature space (high-dimensional space where the network lies). Given a network (PPIN in the case of this work) lying in a space of high dimensions (see Figure 4.1A), MCE extracts the MST directly from this available network (see Figure 4.1B) and generates the MC-kernel by computing pairwise distances over the MST alone (see Figure 4.1C). The embedding process by MDS would be the next step but in this work we propose a more versatile and efficient mapping procedure: embedding by Singular Value Decomposition (SVD) of the centred or non-centred MC-kernel (see Figure 4.1D). We refer to the former possibility simply as MCE (because in fact it is theoretically equivalent to embedding by MDS) and to the latter as ncMCE (see Algorithm 1).

As stated above, the hypothesis behind link prediction by Network Embedding is that once the network is mapped to a low dimensional space, nodes that are close to each other are more likely to interact (Boguña et al., 2009). Kuchaiev, You and their colleagues (Kuchaiev et al., 2009; You
et al., 2010) exploited this idea and assigned likelihood scores to candidate PPIs by means of EDs between nodes in the reduced space (see Figure 4.2A).

In this chapter, we propose a different scoring scheme. To maximally exploit the effect of such soft-thresholding, we propose the use of the SP distance in the low-dimensional space. As the network topology - mapped to a reduced space - should now be sufficiently denoised by means of the MCE device, the use of the SP (instead of the ED) appears to be a more appropriate way to assign distances between nodes because it obeys the denoised network topology. This is even more reasonable considering that each interaction is remapped to the reduced space with a positive and definite numerical weight (see Figure 4.2B). In conclusion, we expect the SP to be a congruous measure for converting the topological discrimination obtained by the MCE soft-threshold effect into a value. This computational engagement between MCE as a technique for embedding (useful for denoising networks affected by FP interactions) and SP for determining network-connectivity distances (effective when the networks are clean or denoised and present few FP interactions), gives rise to a synergy that can boost the separation between good and bad candidate links.
**Algorithm 1: Minimum Curvilinear Embedding (MCE)**

**input:** \( A, N \times N \) adjacency matrix representation of a PPIN (\( N \) is number of nodes in the network);
\( d \), the embedding dimension;
\( c \), a Boolean specifying whether the MC-kernel will be centred or not;

**output:** \( X, N \times d \) matrix whose rows are points with coordinates in a \( d \)-dimensional reduced space;

1. Extract the minimum spanning tree \( T \) out of \( A \);
2. Compute the distances between all node pairs over \( T \) to obtain the MC-kernel \( D \);
3. If \( c == \) TRUE then
   - Centre kernel \( D \), i.e., \( D = -\frac{1}{2}JD^{2}J \) with \( J = I - \frac{1}{N}11^{T} \);
4. Perform economy size singular value decomposition of \( D \), i.e., \( D = U_{d}\Sigma_{d}V_{d}^{T} \);
5. Return \( X = (\sqrt{\Sigma_{d}}V_{d})^{T} \);

*\( M^{T} \) indicates matrix transpose, \( I \) is the \( N \times N \) identity matrix and \( 1 \) is a column vector of ones.

---

Figure 4.2: **A** Scores for candidate links by computing Euclidean distances in the reduced space. **B** Our proposed scoring scheme requires the reconstruction of the original network in the low dimensional space. This network is now weighted (see edge thickness in the figure) and **cleaner** shortest-paths are the scores for the candidate interactions. Note how the score for candidate interaction A-O is different in the two schemes (red lines in the figure).

A few extra words about the advantages of the non-centred version of MCE. The expression *crowding problem* means that after dimension reduction, data clusters collapse on top of each other in the reduced embedding space (van der Maaten and Hinton, 2008). This problem has particular relevance in network embedding because we want to avoid diverse network components collapsing in the same region of the reduced space, as this can cause incorrect link predictions. For this reason, we decided to introduce the ncMCE and test its performance in solving the crowding problem.

There is no universal rule for when centring transformation should be used in the analysis. Nevertheless, non-centring has been shown to offer several advantages (Basnet, 1993; Jolliffe, 2002). This is particularly evident in visualisation tasks, when the set of points that forms each cluster is distributed around the centre of mass in the high-dimensional space. If we perform embedding in
two dimensions after the centring transformation, the points tend to overlap around the origin of the first two dimensions, which is a typical example of the crowding problem. However, in most cases, executing the embedding without centring can significantly reduce this issue. In addition, omission of the MC-kernel centring means that MCE has a time complexity of $O(N^2)$, and thus is more efficient than the other considered embedding techniques, such as MCE and ISO, that have a time complexity of $O(N^3)$. For this reason, ncMCE also offers a significant computational advantage for handling very large networks (more about computational time complexity of link predictors in Appendix B).

### 4.4 Determining the Dimension of Network Embedding

To determine the best dimension to embed the network into, we investigate the results generated by two approaches: dimension determination by AUC (see Figure 4.3A) and by Resolution (Res) (see Figure 4.3B).

#### 4.4.1 The AUC criterion

In the AUC approach (based on the work of You et al. in 2010), the network is embedded into dimension 1 and the distances between all nodes are computed in the reduced space. These distances are sorted by increasing length and a threshold $\varepsilon$ is varied from 0 to the longest distance to quantify the number of True Positives (TPs) (links from the original network that pass the $\varepsilon$ cut), False Negatives (FNs) (links from the original network that do not pass the $\varepsilon$ cut), True Negatives (TNs) (non-directly connected nodes that do not pass the $\varepsilon$ cut) and FPs (non-directly connected nodes that pass the $\varepsilon$ cut). With these numbers at each $\varepsilon$, we can compute the TP rates and FP rates and generate a Receiver Operating Characteristic (ROC) curve. The process is repeated for higher dimensions, until the difference between the AUCs for dimension $\text{dim}$ and dimension $\text{dim}-1$ is less than 0.001 (resulting in $\text{dim}$ as the selected dimension, see Figure 4.3A). In several tests we found that 0.001 is sufficiently small to suggest that no better performance would be achieved if nodes were mapped to higher dimensions, for this reason we chose it as a stopping flag in the AUC criterion. The problem with this approach is that it takes the original network as ground truth, which may not be accurate given the amount of non-real interactions included in the topology due to experimental bias or defects (You et al., 2010).
4.4.2 The Resolution criterion

The Res approach addresses the above-mentioned issue. The best dimension according to this criterion is the one that provides good discrimination between good and bad candidates for interaction, i.e., the more different the distances in the reduced space, the higher the resolution and the better the dimension. Application of Res requires the use of Equation 4.4.1. If \( \text{score}_{\text{dim}} \) is the set of scores assigned to the candidate interactions in dimension \( \text{dim} \), \( \text{unique}(\text{score}_{\text{dim}}) \) is a function that discards the duplicates in the set and returns only its distinct elements. Later, the standard deviation \( \sigma \) of the unique scores is computed to assess the quality of the resolution provided, to finally divide the result by \( \text{dim} \) to penalise high dimensions, which have been shown not to provide any relevant increase in performance (You et al., 2010). The dimension that generates greater resolution is then selected (see Figure 4.3B and Equation 4.4.1):

\[
\text{Dimension} = \arg\max_{\text{dim}=1,\ldots,d} \frac{\sigma(\text{unique}(\text{score}_{\text{dim}}))}{\text{dim}} \tag{4.4.1}
\]

One can think that for Res to work, the network has to be embedded up to a sufficiently large dimension \( \text{dim} \) to determine the one with the best resolution. After several tests (data not shown) we found that this resolution is attained for the lowest dimensions. Embedding the network up to dimension 15 would therefore be enough.

We specifically designed this criterion to fit with the power of MCE, which provides more of a soft-threshold effect in low dimensions. This is why we only tested the resolution criterion in combination with it. We also applied a variation of Equation 4.4.1 to check whether dimension determination using only the top-100 candidate interactions would generate a better AUP. The difference between Equations 4.4.1 and 4.4.2 is that in Equation 4.4.2 we compute \( \sigma \) on the unique scores from the top-100 candidate protein pairs:

\[
\text{Dimension} = \arg\max_{\text{dim}=1,\ldots,d} \frac{\sigma(\text{unique}(\text{scores}_{\text{top-100}}_{\text{dim}}))}{\text{dim}} \tag{4.4.2}
\]

4.5 Testing the Proposed Innovations

4.5.1 Fraction of FPs visited by MC and SP

Random Geometric Graphs (RGGs) are important because there is indication that they can be good models for networks such as PPINs (Przulj et al., 2004). We generated RGGs by accommodating
1000 points uniformly at random in the 100-dimensional unitary cube and then connected them if and only if the dot product (similarity) between the vectors with tails in the origin and heads over these points was above a connectivity threshold \( r \). We set the threshold by ensuring that properties common to real biological networks (small-world and scale-free topologies) and connectivity were present. The advantage of using RGGs to test our innovations is that the sets of true and spurious interactions are clearly defined: true interactions are those that comply with the threshold and spurious links are those that do not. Based on this, we generated noise in the structure of 1000 different RGGs (noisy networks) in amounts typical of PPINs (40% FNs and 60% FPs) and counted the fraction of unique FPs visited when computing SPs between all node pairs in the networks (first step of the ISO algorithm) and when computing MC distances (first step of the MCE algorithm).

The results presented in Figure 4.4 for the RGGs suggests that the estimate of non-linear connectivity measure using the MST (i.e. MC) takes into account only a small proportion of FPs, offering a denoised estimate of the network connectivity. In contrast, using the SP involves visiting all FP edges at least once, which introduces a lot of noise to the link prediction process. The same investigation was conducted on the four yeast networks mentioned before, in which the FP links were identified using the same GO-based strategy of the ERPE (see Section 2.2.2 in Chapter 2). The outcome of this second analysis (see Figure 4.4 ) converged to the same result obtained for the artificial networks. These findings support the hypothesis that MCE should be a powerful tool for link prediction in noisy networks. In fact, as noisier networks present more FP interactions, the use
of ncMCE/MCE in such cases should produce an even greater increase in performance over the use of ISO.

Figure 4.4: Fraction of FPs used by shortest-path (SP) and minimum curvilinearity (MC). For the artificial networks (RGGs), mean and standard error values on 1000 network realizations are provided.

4.5.2 Solving the Crowding Problem

Although it is not a biological dataset, the radar signal dataset is a point of reference in machine learning and is an important benchmark for testing the ability of embedding techniques to solve the crowding problem (Shieh et al., 2011). Instead of creating an artificial dataset, we decided to use a real one to test whether ncMCE is able to solve the crowding problem and, as a result, better embed networks into low dimensions. The radar signal dataset is highly non-linear and has 351 samples characterised by two classes: good radar signals that are highly similar and bad radar signals that are highly dissimilar (Shieh et al., 2011).

ncMCE (Figure 4.5A) offered the best embedding of this dataset and attained high linearisation (Figure 4.5E) in both the first (AUC = 0.95) and the second dimensions (AUC = 0.96). The ROC curve is used to evaluate the discrimination power along a dimension of projection: if the dimension offers a linear discrimination between the good and bad signals, the respective AUC will be 1. We also tested the performance of ISO, which is a reference algorithm for non-linear dimension reduction, but its embedding was highly crowded (Figure 4.5C). In contrast, Tree Preserving Embedding (TPE), a recent parameter-free algorithm for non-linear dimensionality reduction (Shieh et al., 2011), produced non-linear discrimination (good signals in the centre and bad signals on the periphery) of the clusters around the origin of the axis (Figure 4.5D). This demonstrates that TPE can address the crowding problem but cannot solve the non-linearity of the dataset. MCE solved the non-linearity in the second dimension (Figure 4.5E), but only partially addressed the crowding prob-
lem (Figure 4.5B). The only algorithm that was able to simultaneously solve both the non-linearity and the crowding problem in this dataset was ncMCE (Figure 4.5A, E). Interestingly, on the basis of the embedding offered by ncMCE and MCE, one might speculate that the high dissimilarity between the bad radar signals pointed out in previous studies (Shieh et al., 2011) could be interpreted as the presence of at least two different kinds of bad radar signal clusters that are difficult to embed due to their high non-linearity (elongated and/or irregular high-dimensional structure). The possible different bad-signal clusters are indicated in grey and black in Figure 4.5. Finally, whereas only a few seconds were needed to run ncMCE, MCE and ISO, TPE took several hours to embed this small dataset, and its current implementation can be prohibitively slow for large datasets. As TPE is inefficient for embedding networks composed of thousands of nodes, we could not evaluate its performance in the present study.

4.5.3 Discrimination Between Good and Bad Candidate PPIs

Given the embedding of any PPIN, if the hypothesis that nodes closer to each other in the reduced space are more likely to interact is true, the distribution of low-dimensional distances between connected network nodes $p(\text{distance}|\text{original})$ should have higher peakedness (kurtosis) than the distribution of distances between candidate-links $p(\text{distance}|\text{candidate})$; in addition, $p(\text{distance}|\text{original})$ should be shifted towards zero. Following Kuchaiev et al. and You et al. (Kuchaiev et al., 2009; You et al., 2010), for each of the four considered yeast networks, we fitted a non-parametric estimate to the above two distributions respectively and we used the Mann-Whitney non-parametric test to determine whether there was a statistically significant difference between them over different dimensions of embedding.

The results in Figure 4.6A-D show that in all networks, ncMCE-SP had the highest kurtosis and shift towards zero. Moreover, links from the original network topology that are distant from the origin are likely to represent FPs, while non-adjacent nodes whose distance is close to zero are good candidates for interaction. In addition, the rightmost panel of Figure 4.6 shows that in the four considered networks, there was a statistically significant difference between $p(\text{distance}|\text{original})$ and $p(\text{distance}|\text{candidate})$. This significant difference was conserved across the different dimensions analysed and was much larger when the SP scoring scheme was used in the reduced space. This indicates that SPs should work better than EDs for scoring proximity distances between network nodes (proteins) in the reduced space.
Figure 4.5: Embedding of the radar signal dataset. The red spots indicate good radar signals. The grey and black spots indicate bad radar signals, which might be interpreted as two diverse sub-categories of bad signals. 

A ncMCE.  
B MCE.  
C ISO.  
D TPE.  
E ROC and respective AUC computed for evaluating the linear discrimination performance of the first (Dim1) and second (Dim2) dimensions. The evaluation was repeated for each of the four techniques on the two dimensions of embedding. To facilitate the visualisation, we do not report the ROC for TPE due to its poor performance.

4.5.4 Link Prediction Evaluation on Random Geometric Graphs

Using the same 1000 noisy RGGs generated for the experiment of Section 4.5.1, we applied the PRPES (see Section 2.2.2 in Chapter 2) to assess the prediction power of ncMCE/MCE and ISO for link prediction over these artificial networks. During this test, the networks were embedded into dimensions 1 to 10, which is the recommended range for testing the performance of ISO (You et al., 2010). Next, we repeated the experiment to assess the performance of the embedding predictors on a sparsification experiment over 1000 different RGGs without noise (clean networks). Figure
Figure 4.6: Distribution of shortest-path scores in the reduced space (dimension 3 displayed) for the four datasets analysed in this work (A Network 1, B Network 2, C Network 3 and D Network 4). Network links $p$ (distance | original) (solid line) and candidate links $p$ (distance | candidate) (dashed line) after ncMCE (left) and ISO (middle) network embedding. The insets show the distribution of ED scores. The rightmost panel shows the resulting p-values from the statistical test for difference in the distributions over different dimensions.
4.7A-C shows that the two variations of MCE (especially ncMCE-SP using dimension one) were the strongest approaches for rediscovering true interactions in the noisy RGGs. Figure 4.7D-F shows that in the same experiment but in clean RGGs, ncMCE-SP still had the best performance (using dimension 1), but ISO-SP came significantly closer.

It is important to mention that the low precision values obtained in these last two experiments are very common in link prediction due to the large amount of candidate interactions compared to the small amount of pruned interactions a technique is trying to rediscover (see for example (Liben-Nowell and Kleinberg, 2007), in which precisions in the range of 0.0015-0.0048 are reported when predicting links in coauthorship networks). We generate RGGs with 1000 nodes and around 12600 edges (we say around because RGGs are random and the number of edges is not fixed), which means that the number of candidate interactions is $1000(1000 - 1)/2 - 12600 = 486900$. At the last sparsification level (right before the network loses connectivity and when the largest amount of pruned links should be rediscovered), the number of pruned links is $\sim 11000$, this means that the probability that a random prediction is correct is $11000/(486900 + 11000) = 0.02$. The precision of, for example, ncMCE is 0.06 at this level for noisy RGGs (see Figure 4.7A), which is a three-fold improvement over the analytical precision of a random predictor.

Altogether, these results indicate that (i) the use of ncMCE presents a clear advantage over MCE; (ii) the lower dimensions (especially dimension 1 for ncMCE) are very effective when using ncMCE/MCE-based algorithms; (iii) the gap between ncMCE-SP and ISO-SP increases in the presence of noise, which especially encourages the use of ncMCE-based algorithms in noisy networks such as PPINs; and (iv) the use of SP (to score in the reduced space) generally offers a clear advantage over the use of ED. RGGs are crucial for designing a ground-truth evaluation that allows us to directly observe the effect of introducing noise (false interactions) in the rediscovery of the real/original network topology. Because GO-free evaluations are essential for demonstrating the performance of link predictors in the absence and presence of network noise, the findings up to this section represent very important results.

### 4.5.5 Link Prediction Evaluation using the Gene Ontology

The results obtained so far support our intuition that network embedding by ncMCE/MCE combined with the SP connectivity distance in the reduced space can boost the performance in topological prediction of candidate PPIs. We used the ERPE using the GO as external referee (see Section 2.2.2 in Chapter 2) to assess the performance of MCE in its two flavours against that of ISO.
Figure 4.7: Mean rediscovery precision of TP interactions for different sparsification levels of noisy networks with 60% FP interactions in their topology: embedding dimensions 1 (A) and 4 (B) are displayed. The standard error bar is reported for each point. Analogous plots for clean networks (which do not present FP interactions in their topology) are reported again for dimensions 1 (D) and 4 (E). The Area under the Sparsification Curve is reported for each dimension of embedding, considering the rediscovery of TP interactions in noisy (C) and clean (F) networks. The arrow indicates the overall best performance (given by ncMCE-SP in dimension one). The percentage improvement in respect to the best ISO performance (ISO-SP) is reported.

Figure 4.8 provides experimental confirmation of our intuition. MCE and ncMCE combined with SP outperformed both ISO and pure SP (computed on the original network without embedding) in all networks. ISO performed even worse than pure SP in the first network. Besides, the simulation in Figure 4.8 suggests that in general ncMCE-SP slightly outperforms MCE-SP. These results hold even if the MF ontology is not considered in the GO relevance test (for cases in which, for example, enzymes interact but do not perform similar functions), if more than 10% of the links in the networks are taken from the top of the list (see Chapter 2) or if proteins that are part of large complexes (such as the ribosome, the proteasome and the exosome) are removed from the network (given the fact that large protein complexes tend to be the most numerous in available PPINs, the performance of prediction techniques can be biased if the predictor mostly recognises interactions between proteins taking part in these complexes (Reyes, 2009)).

In addition to the above results, there is evidence (Figure 4.9) that although we used several advanced techniques for dimensionality reduction (both unsupervised and supervisedly tuned), ncMCE-SP remained the overall leader, which represents our second important finding. Surprisingly, we discovered that even though local MDS and NeRV were supervisedly tuned to achieve their best
Figure 4.8: Performance comparison between ncMCE, MCE, ISO and pure SP computed in the high-dimensional space. The x-axis indicates how many interactions are taken from the top of the candidate interaction list (sorted decreasingly by score) and the y-axis indicates the precision of the technique for that portion of protein pairs. Solid lines indicate the use of SPs in the reduced space to assign scores and dashed lines the use of EDs.
performance, they could not equal ncMCE-SP. This result suggests that force-based methods for embedding are not appropriate in this context, at least when combined with ED or SP in the reduced space. The reason for their poor performance is that these algorithms perform an embedding that finely preserves the network topology, thus also preserving the noise. In contrast, ncMCE provides a soft-threshold effect (discussed in Section 4.3), which boosts the separation between good and bad candidate links.

For completeness, we compared ncMCE-SP with the bio-inspired link predictors FSW and CDD (see Table 3.1 in Chapter 3), two of the most powerful node neighbourhood techniques (You et al., 2010). ncMCE-SP ranked first, with a notable improvement, in the first two networks (Network 1 and Network 2, Figure 4.10) and second in the third network (Network 3, Figure 4.10). In the fourth network, all of the techniques produced similar performances (Network 4, Figure 4.10). According to the minimum AUP attained in the four different networks, ncMCE-SP was also the most robust technique. FSW ranked first in the third network, while in the first two networks its performance was similar to that of CDD. Given these results, we can conclude that ncMCE-SP offers a general improvement, particularly in robustness, compared with the other techniques.

4.5.6 Testing the Criteria for Dimension Determination

So far, in the simulations showed in Figures 4.8-4.10, we have used the AUC criterion, which was designed to work with any algorithm for embedding. Unlike the AUC criterion, the Res criterion was designed to fit better with MCE, which provides more of a soft-threshold effect (thus stronger denoising) in the lowest dimensions.

This is experimentally confirmed in Figure 4.11A, where the peaks of the resolution criteria (both ResAll and Res100) are always in one of the first two reduced dimensions. From Figure 4.11B, we gather that the AUC criterion and the ResAll criterion selected the same dimensions, and thus show equal precisions. However, in terms of robustness (Figure 4.11C), the Res100 criterion slightly outperformed the others. These results corroborate our intuition to invent a new and radically different criterion based on the resolution of the unique score values, which is an easy and time-efficient strategy.

4.5.7 Network Sparsification Evaluation

To present a more refined vision of the potential offered by topological link-prediction techniques, we used the ERPES using the GO as external referee (see Section 2.2.2 in Chapter 2). This approach was
Figure 4.9: ncMCE-SP against advanced unsupervised and supervisedly tuned embedding techniques. The x-axis indicates how many interactions are taken from the top of the candidate interaction list (sorted decreasingly by score) and the y-axis indicates the precision of the technique for that portion of protein pairs. Solid lines represent the performance of techniques that use SPs in the reduced space to assign scores and dashed lines represent techniques that use EDs. Although ncMCE-SP (red solid line) is an unsupervised approach, it appears on both sides for reference.
used to generate the results shown in Figure 4.12, which compares the main embedding techniques (ncMCE, MCE and ISO) and the reference node neighbourhood techniques (FSW, CDD, IG1 and SP). All of the embedding methods were tested in combination with the same distance (SP) to measure link likelihood in the reduced space. Figures 4.12A and B display the sparsification curves of the first two networks for ncMCE-R (R indicates the use of the resolution criterion) and FSW, which were the highest ranked methods overall in their respective categories. Although ncMCE-A (A indicates the use of the AUC criterion) attained the same result as ncMCE-R in each network, for the sake of clarity, we display only the curve of the latter. The methods were ranked considering the area under the sparsification curve (AUS). To evaluate the general performance of the methods, we considered the minimum AUS performance of each method for all networks (Figure 4.12C).

A special variation of this experiment was performed on each network to investigate whether the extraction of different MSTs from the networks resulted in important changes in ncMCE performance (data not shown). This is a possibility because all the network links have a weight of 1 in unweighted networks like the PPINs under study. For this test, only ncMCE was used because it generally outperformed MCE, as shown in Figures 4.8 and 4.9. Here, for each percentage of link deletions, 100 different MSTs were extracted (by random initialisation). The AUPs attained by the different ncMCEs (each of which uses a different MST) were averaged and their standard error bars included in the sparsification curve. The standard error bars for the ncMCE’s sparsification curves showed that the difference in the performance of ncMCE when using different MSTs was negligible.

In general, MCE-based embedding techniques (red bins in the histogram of Figure 4.12C) outperformed the node neighbourhood techniques (green bins in the histogram of Figure 4.12C) and ncMCE was again the best method. Taken together, our experiments suggest that ncMCE-SP rep-
Figure 4.11: A Dimension determination curves for ncMCE-SP over all the candidate interaction scores (Res_{All}) and the top-100 ones (Res_{100}). The x-axis indicates the different dimensions tested and the y-axis the measure of resolution for a specific dimension. The upper panel shows the curves for up to 100 dimensions and the lower panel is the zoomed-in portion of the plot for dimensions 1 through 10. B Precision curves that show the performance achieved by the dimensions determined using the AUC criterion, Res_{All} and Res_{100}. C Performance Robustness (minimum AUP amongst all networks) for the different criteria combined with ncMCE-SP.

represents a new benchmark for robustness in the topological prediction of PPIs, and this is the third main result of our study.

As a further investigation, starting with the final set of sparsified networks generated in the
previous experiment, we re-densified their topologies by random addition of links and applied two approaches (ncMCE-R and FSW) at each percentage of densification. As we can see in Figures 4.12A and B, this process was unable to re-create a meaningful topology that might have been shaped by evolutionary features in the history of the protein interactome. If a topology analogous to the original had been recovered, the prediction techniques would have been able to achieve a performance comparable with that reached before network sparsification.

This finding emphasises the presence of preferential bio-information stored in the PPIN topology that cannot be modelled by uniform random sampling of new interactions. Therefore, the simple unweighted topology can be highly informative for different purposes, one of them being the prediction of new interactions or alternatively, as recently shown, the structural controllability of any complex network (Liu et al., 2011).

4.5.8 In silico Validation

As mentioned in the Introduction, the experimental detection of PPIs can be very expensive in terms of both time and money. The computational approaches we investigated to predict novel interactions are meant to guide wet-lab experiments rather than to complete the interactome of the organism under study. Currently, the Y2H validation of 100 protein pairs can represent a challenging upper limit to simulate a real scenario for the budget of many labs. We decided to suggest different sets of candidate interactions to test in wet-lab experiments, and we report the evaluations for different thresholds: 20, 40, 60, 80 and 100. We executed an in silico validation to verify the quality of the candidate interactions proposed by the best techniques. The top-100 candidate interactions for ncMCE-SP-Res100 and FSW were tested on the STRING database, which is the most complete PPI database (Szklarczyk et al., 2011). The results for the different thresholds are reported in Figure 4.13E. ncMCE-SP-Res100 attained promising results in this last test, surpassing FSW for GO precision (Figures 4.13A and D), GO robustness (Figure 4.13B) and STRING confidence robustness (Figures 4.13C and D).

To search for the biological information related to the interactions predicted by ncMCE-SP-Res100 and validated in STRING, for each network we performed a pathway enrichment analysis using DAVID Bioinformatics Resources 6.7 (Huang et al., 2009a,b). For each network, the list of proteins involved in the predicted and validated interactions was tested against all network proteins as background. This kind of background choice was motivated by the fact that it tends to produce more conservative p-values and, in fact, a general guideline for the enrichment analysis is to use a
narrowed-down list of genes instead of all genes in the genome (Huang et al., 2009a,b). In addition, the Benjamini correction for multiple hypotheses test was applied. The results of the analysis emphasise that the lists of predicted and STRING-validated protein interactions have significant biological meaning in at least one pathway for each of the investigated networks. Interestingly, the predicted proteins were involved in cellular processes (e.g. cell cycle), nucleotide metabolism (e.g. pyrimidine and purine metabolism) and genetic information processes (e.g. RNA polymerase and RNA degradation). This evidence suggests that the proposed method predicted interactions in different network modules that are related to significant and heterogeneous pathways in yeast.
**Figure 4.13:**

**A** GO precision curves for the top-100 candidate PPIs proposed by ncMCE-SP combined with Res100 (referred to as ncMCE-Res100 in the figure) and FSW. The x-axis indicates how many interactions are taken from the top of the candidate interaction list (sorted decreasingly by score) and the y-axis indicates the precision of the technique for that portion of protein pairs.

**B** GO performance robustness for the above methods.

**C** In silico validation of the top-100 candidate PPIs proposed by the above methods. Sub-networks formed by the top-100 candidate PPIs proposed for each network by ncMCE-Res100. The red edges indicate links validated in the STRING database. The number of validated PPIs, their average STRING confidence along with standard deviation and their average GO confidence along with standard deviation, appear on top of each network.

**D** Validation in STRING for different thresholds over the top-100 candidate links.
4.6 Theoretical Support for the MC Framework

All of the above simulations provide experimental evidence of the power of the MC concept to tackle link prediction, especially in noisy networks like PPINs. In addition to this, there is also theoretical work, based on complex network geometry, that help us better understand the reason why MC performs that well in the problem posed in this chapter.

As already stated in Section 2.1.1 in Chapter 2, Krioukov and colleagues proposed that a hyperbolic geometry underlies the structure of complex networks and explains the presence of topological features common to several of them, such as heterogeneous degree distributions and high clustering (Krioukov et al., 2010). It is in the light of this idea that we can understand where the power of MC stems from.

Let us define $\mathbb{H}^2_\zeta$ as the hyperbolic, two-dimensional space of constant curvature $K = -\zeta^2 < 0$ with $\zeta > 0$ equipped with a metric $d_{\mathbb{H}} : \mathbb{H}^2_\zeta \times \mathbb{H}^2_\zeta \to \mathbb{R}$, such that the hyperbolic distance between two points at polar coordinates $a_1(r_1, \theta_1)$ and $a_2(r_2, \theta_2)$ is defined as:

$$d_{\mathbb{H}}(a_1, a_2) = \frac{1}{\zeta} \cosh \left( \cosh(\zeta r_1) \cosh(\zeta r_2) - \sinh(\zeta r_1) \sinh(\zeta r_2) \cos \theta_{12} \right)$$ \hspace{1cm} (4.6.1)

where $\theta_{12} = \pi - |\pi - |\theta_1 - \theta_2||$.

Based on 4.6.1, it is not difficult to see that in $\mathbb{H}^2_\zeta$, the length of a hyperbolic circle and its area are given respectively by:

$$L_{\mathbb{H}}(R) = 2\pi \sinh(\zeta R)$$ \hspace{1cm} (4.6.2)

$$A_{\mathbb{H}}(R) = 2\pi (\cosh(\zeta R) - 1)$$ \hspace{1cm} (4.6.3)

where $R$ is the radius of the hyperbolic circle of interest.

Since $\sinh x = \frac{e^x - e^{-x}}{2}$ and $\cosh x = \frac{e^x + e^{-x}}{2}$, it is clear that both $L_{\mathbb{H}}(R)$ and $A_{\mathbb{H}}(R)$ grow exponentially (i.e. as $e^{\zeta R}$) with $R$ (see Equations 4.6.2 and 4.6.3).

Complex networks are usually so sparse (like the PPINs analysed in this chapter) that can be thought of as trees and trees can be thought of as discrete hyperbolic spaces. In a complete, balanced $b$-ary tree $T$ (a tree with branching factor $b$) with depth $R$, the analogies of the hyperbolic circle’s length and area are respectively:

$$L_T(R) = b^R$$ \hspace{1cm} (4.6.4)
The total number of nodes: \( A_T(R) = \frac{b^{R+1} - 1}{b - 1} \) \hspace{1cm} (4.6.5)

Notice how both \( L_T(R) \) and \( A_T(R) \) in Equations 4.6.4 and 4.6.5 grow as \( b^R \) with \( R \) but if we apply natural logarithms to both sides of the equations we end up with \( L_T(R) = e^{R \ln b} \) and \( A_T(R) = e^{R \ln b + \ln(\frac{\ln b}{b - 1})} \). As a result, if we set \( \zeta = \ln b \), from a purely metric perspective, \( H^2_{\ln b} \) and \( b \)-ary trees are equivalent.

The above means that when measuring distances between nodes over the MST as dictated by the MC-framework, one is in fact navigating one of the discrete representations of the hyperbolic geometry underlying the network under study. Further support for the importance of spanning trees in complex networks stems from the study of network skeletons, also known as communication kernels (spanning trees of maximum betweenness), which play important roles for traffic and communication between nodes and preserve the scale-free topology of the networks from which they are extracted (Kim et al., 2004, 2005; Goh et al., 2006).

### 4.7 Conclusions and Perspective

Considering the difficulty of dealing with sparse and noisy protein networks (You et al., 2010), our results represent a promising achievement and encouragement to deepen the investigation of network embedding techniques for topological prediction of protein interactions. In our tests, ncMCE showed enhanced performance in network embedding-based link prediction compared with other dimension-reduction algorithms. In addition, ncMCE has a time complexity of only \( O(N^2) \), which is lower than the complexity of the other considered machine learning techniques and is a valid candidate for handling very large networks. Finally, our experiments revealed that the shortest path works significantly better than the Euclidean distance for scoring network protein-pairs embedded in the reduced space. We envision that network-embedding techniques for predicting novel PPIs might play an important role in the development of systems biology tools, such as those used for network-based inference of disease-related functional modules and pathways (Cannistraci et al., 2013a). The real biological interactions could be complemented with the \textit{in silico} predicted ones to boost the inference of functional modules. In the near future, this last point will become increasingly important for patient classification, diagnosis of disease progression and planning of therapeutic approaches in personalised medicine (Ammirati et al., 2012).
Chapter 5

Exploitation of Genetic Interaction Network Topology for the Prediction of Epistatic Behaviour

Abstract

GI detection impacts the understanding of human disease and the ability to design personalised treatment. The mapping of every GI in most organisms is far from complete due to the combinatorial amount of gene deletions and knockdowns required. Computational techniques to predict new interactions based only on network topology have been developed in network science but never applied to GI networks. We show that topological prediction of GIs is possible with high precision and propose a graph dissimilarity index that is able to provide robust prediction in both dense and sparse networks. Computational prediction of GIs is a strong tool to aid high-throughput GI determination. The dissimilarity index we propose in this article is able to attain precise predictions that reduce the universe of candidate GIs to test in the lab.

GIs or *epistasis* (see Section 2.1.4 in Chapter 2 for definitions) play an important role in untangling the relationships between genotype and phenotype, advance our understanding of human disease (Wiltshire et al., 2006; Combarros et al., 2009; Evans et al., 2011) and improve our ability
to design personalised treatment plans. Nevertheless, mapping every GI is far from complete, even in model organisms. Here is where computational prediction of novel GIs comes into play.

Computational techniques to predict undetected GIs aim to reduce costs incurred by their experimental detection and target good candidates to test in the lab. Current attempts to computationally predict GIs depend on biological information that, for some organisms, might not be available. Examples of these efforts are the integration of data that characterise epistasis, such as gene expression, gene product physical interaction or functional annotations to train probabilistic decision trees (Wong et al., 2004) or to apply logistic regression (Zhong and Sternberg, 2006). Other endeavours involve the overlap of data coming from different networks (Protein Interaction Networks, Gene Ontology Networks, Co-expression Networks, etc.) and the application of random walks (Chipman and Singh, 2009) or an ensemble of classifiers (Pandey et al., 2010).

We propose the exploitation of the biological information stored exclusively in the network topology that should be shaped by the genomic properties characterising the organism under investigation. To this effect, parameter-less neighbourhood-based (both general-purpose and bio-inspired) and network-embedding techniques are applied to the GINs of two different organisms (worm and yeast) the first of which is sparser than the second one. The reliability of these techniques and the impact of sparse network architecture on their prediction performance are analysed and discussed. We also propose a graph dissimilarity index that proves to perform better in GI prediction, for the networks here considered.

5.1 Network Datasets and Link Predictors

This work focuses on negative interactions due to their known impact on essential biological functions (Costanzo et al., 2011). The datasets used correspond to GIs in *Saccharomyces cerevisiae* (budding yeast) and *Caenorhabditis elegans* (nematode worm), detected by Costanzo and colleagues and Byrne and colleagues (Costanzo et al., 2010; Byrne et al., 2007) respectively and downloaded from BioGRID 3.1.85 (Stark et al., 2006) (Networks 9 and 10 in Table A.1 in Appendix A). Self-interactions and redundant links were removed from these datasets to constitute a Worm GIN of 457 nodes and 1242 links (average node degree = 5.44) and a Yeast GIN of 3842 nodes and 52179 links (average node degree = 27.16). Notice the latter is \( \sim \)5 times denser than the former.

In this chapter we evaluate the performance in link prediction of all the neighbourhood-based and bio-inspired techniques listed in Table 3.1 in Chapter 3. The performance of the centred and non-centred versions of ISO and MCE were also investigated using dimension determination by AUC
and Res (see Chapter 4 and Figure 4.3 in the same chapter).

5.2 Network Sparsification Evaluation

We used the ERPES with the GO as external referee and found that precise topological prediction of GIs is possible: most AUP values before network sparsification are over 0.7 (see Figure 5.1) whereas the random predictor is only able to provide an AUP of ∼0.3 in the Worm GIN (Figure 5.1A) and of ∼0.57 in the Yeast GIN (Figure 5.1B). It is important to note that techniques such as JC and IG1 are unable to advocate good candidates for interaction (they perform worse than random) when the network is sparse (see Figure 5.1A) due to their dependence on information that is scarce in this setting (see their formulation in Table 3.1 in Chapter 3); however, when the network is dense they are able to predict new links (see Figure 5.1B). Another interesting result is that of PA, whose performance seems to improve as the networks get sparser. This behaviour goes in line with the situation in which link formation over a very sparse network, where nodes are close to being isolated entities, is accurately predicted by a rich-get-richer phenomenon (Newman, 2001a; Barabási and Albert, 1999).

In general, we can say that when the networks are sparsified it is evident that neighbourhood-based techniques lose precision because of their dependence on the number of node-neighbours, which is reduced as links are removed from the network topologies (this is especially clear in Figure 5.1B). Conversely, embedding techniques are more robust to link deletion regardless of the dimension determination approach used, particularly MCE (Cannistraci et al., 2010, 2013b), which works over a very sparse graph: the MST it extracts from the network. This smart sparsification mechanism allows MCE to be very robust to link deletion by mining information from the core of the network itself. In fact, the MST is constructed using a greedy process to navigate the network, thus being an efficient and reliable way to estimate proximity distances over the network topology (Boguñá et al., 2009). ISO’s performance turned out to be unstable given that computing shortest-paths over a network is not a greedy process, which can lead to prediction of irrelevant GIs. This is especially true when the network is randomly sparsified, which can eliminate important links and generate even noisier network topologies (see Figure 5.1A).

It is important to mention that, since the networks we are working with are unweighted (only 1s and 0s indicate link existence), more than one MST can be extracted from them. The plots under MCE’s sparsification curves in Figure 5.1 show that the difference in ncMCE’s performance (the best variation of the MCE approach) when using different MSTs is so small (see standard error bars
in the plots) that it can be neglected.

A final point about network embedding techniques arises from the fact that their non-centred versions perform slightly better than the centred ones. Both MCE and ISO rely on the effect that makes nodes with higher likelihood to interact closer to each other in the reduced space and the effect seems to be lessened if the shortest-path lengths over the network or the MST are corrected by centring (see Figure 5.1). In addition, the non-centred versions are much faster given their quadratic time complexity (see Appendix B) and not the cubic one of the centred versions.

In summary, our results on the Yeast (dense network) and Worm (sparse network) GINs suggest that previous findings that indicate that neighbourhood-based techniques perform better on dense protein interactomes while embedding-based approaches can deal specifically with very sparse and noisy networks (You et al., 2010; Cannistraci et al., 2013b) are also valid for GINs.

Of special interest is FSW, the best technique from the neighbourhood-based group according to our assessments. We found that it has a high dependency on its penalty terms $\lambda_{x,y}$ and $\lambda_{y,x}$ (which penalise nodes with very few direct neighbours) and if they are removed from its formulation, performance is significantly reduced (see ‘FSW no $\lambda$’ in Figure 5.1). This occurs because FSW’s score is very low when nodes $x$ and $y$ do not share many neighbours making all scores very similar in this situation (very close to zero, see Figure 5.1A). The mechanism used by FSW to provide a good discrimination between good and bad candidates for interaction is therefore the penalty term. This mechanism also plays a very important role when the network is very dense (see ‘FSW’ and ‘FSW no $\lambda$’ in Figure 5.1B): it helps FSW to give a low probability of interaction to unbalanced nodes (i.e. nodes with quite different degrees).

5.3 The Adjusted Czekanowski-Dice Dissimilarity

Driven by the benefits of having a penalty term to assign lower scores in case one node has very few neighbours, we propose a robust graph dissimilarity index for GI prediction based on CDD’s formulation: the Adjusted Czekanowski-Dice Dissimilarity or ACDD (see Equation 5.3.1 and Table 3.1 in Chapter 3). To our surprise, a formulation equivalent to ours was independently reached by Liu and colleagues in 2009 for the discovery of protein complexes using weighted protein interaction networks (Liu et al., 2009). Although the motivations that led each group to ACDD’s formulation were different (computational detection of epistasis in our case, detection of protein complexes in theirs), we think that this emblematic example of serendipity gives further support to the good results we obtained using this index.
Figure 5.1: Sparsification experiment in the Worm GIN (A) and the Yeast GIN (B). The ▼ symbol indicates where the simulation starts. The plots under the MCE results correspond to sparsification of one network using the best performing MCE but extracting 100 different MSTs at each percentage of link deletion. The error bars report standard error. The grey regions are an indicator of the difference in sparsity levels between the two networks (27.16 against 5.44).

\[
ACDD(x, y) = \frac{|\Gamma(x)\Delta\Gamma(y)|}{|\Gamma(x)\cup\Gamma(y)|} + \lambda_{x,y} + \lambda_{y,x}
\]

where \(\Gamma(x)\) is the set of neighbours of \(x\) including \(x\) and

\[\lambda_{x,y} = \max(0, n_{avg} - |\Gamma(x)|)\]

with \(n_{avg}\) = the average node degree in the network.

(5.3.1)

To understand why the penalty terms are better accepted by CDD than FSW most of the time, we have to analyse their extreme cases. CDD is a dissimilarity index that ranges between 0 and 1, with 0 representing its best possible score. Contrariwise, FSW is a probability, i.e. it also ranges between 0 and 1, but 1 represents its best score. Let’s consider the case where \(\Gamma(x) \cap \Gamma(y) = \emptyset\); in
this situation FSW would automatically give 0 probability of occurrence to edge \( \{x, y\} \) and, since it will be positioned at the bottom of the candidate list, it might not even be considered in the precision curve analysis. Conversely, CDD assigns a score of 1 (the worst possible) but the penalty terms added to its formula allow for a better discrimination between bad candidates because they automatically add extra value to the score, which now depends solely on the degree of the seed nodes. Note that the way in which the penalty terms are used in ACDD, resemble the way in which the PA index works but in a less stringent manner (having a sum and not a product) and this is important to have a more stable performance that does not drop when the network is close to its sparsification limit.

Indeed, ACDD impressively boosts the performance of the neighborhood-based approaches. Figure 5.2A shows how it outperforms both FSW and CDD, improving the latter by +10% if the areas under the sparsification curves are compared. This might indicate a possible direction of research to develop a valid solution to the problems that neighborhood-based predictors face when the topology is sparse (You et al., 2010). In Figure 5.2B, the improvement is +16%, which is considerable if we take into account that the Yeast GIN is much denser and has a more stable topology for which it is more difficult to make predictions (the saturation of the topology leaves little chance for prediction of putative additional interactions). Additionally, ACDD was recently compared against FSW and CDD in a study conducted by our group on sparse and noisy protein networks, and it showed very close performance to that of FSW in protein interaction prediction (Cannistraci et al., 2013b).

Figure 5.2: Sparsification experiments showing only FSW, CDD and including ACDD. The ▼ symbol indicates where the simulation starts. The grey areas indicate the improvement of ACDD over CDD. A Results for the Worm GIN. B Results for the Yeast GIN.
We took the best performing prediction techniques from each category (i.e. neighbourhood-based, bio-inspired and network embedding) and, starting from the sparsified networks generated in the previous experiment, we re-densified the topologies by random addition of links and applied the different approaches at each percentage of densification (see Section 4.5.7 in Chapter 4). As we can see in Figures 5.3A and B, this process is unable to re-create the topology shaped by the biological properties of the genes and links that form both networks. If the original topology were recovered, the different prediction techniques would have been able to achieve the same performance they reached before network sparsification. This is an important finding because it is confirmation that there is important information stored in the network that can be exploited for different purposes, one of them being prediction of new interactions.

5.4 *In silico* Validation and Intersection with DRYGIN

Performance evaluation by GO is the preferred tool to measure precision in interaction reliability assessment and prediction (Saito et al., 2002, 2003; Chen et al., 2005, 2006a,b; You et al., 2010). Its use has been motivated by the guilt-by-association principle (Oliver, 2000) and, in fact, the GO
is used after high-throughput detection of epistasis to discriminate between true and false positives (Baryshnikova et al., 2010). Nevertheless, GO annotations may present experimental bias or inherent errors (Rhee et al., 2008). To address this issue and as an additional performance evaluation method, we took the top-100 candidate links from the most robust approach (namely ACDD) and verified their overlap with STRING 9.0 (Szklarczyk et al., 2011), a database of gene and protein interactions derived from different sources.

Out of the 100 links taken from the Yeast GIN, 88 were validated with an average STRING confidence of $0.95 \pm 0.13$ (average ± standard deviation). Out of the 100 GIs taken from the Worm GIN, 7 were validated with an average STRING confidence of $0.74 \pm 0.16$. 7/100 may look like a low ratio but if we compare it with the ratio of the random predictor (see below), we take into account that their STRING confidence is high and that there is not much information on epistasis for worm, these 7 GIs represent a valuable short-listed set of candidates for test in the lab. Moreover, CN’s top candidates also overlap with 7 STRING interactions and 6 of them coincide with the ones advocated by ACDD. This is an important cross-validation, because ACDD and CN follow very different prediction approaches and yet, their results for this experiment converge.

If we compare these results with those coming from a random predictor, whose top 100 candidate links matched only 3 STRING interactions with average confidence of $0.51 \pm 0.37$ in the Yeast GIN and 0 STRING interactions in the Worm GIN (even after several trials), we can say that the epistatic behaviour between genes detected by ACDD can be considered reliable in both the Yeast and the Worm GINs.

Along with one of the most important papers on GI detection in yeast (Costanzo et al., 2010), Costanzo and colleagues published a huge dataset with all the yeast double mutants that they tested for GI (~6 million) and a p-value for each one, indicating their likelihood of existence (in fact the Yeast GIN used in this chapter is a subset of that list with a very stringent p-value cutoff). This list of GIs, tested experimentally by means of the Synthetic Genetic Array (Tong et al., 2001), is available on the Data Repository of Yeast Genetic INteractions (DRYGIN) (Koh et al., 2010) and we took advantage of this information to provide a more wet-lab oriented validation of the top-100 candidate interactions advocated by the best prediction techniques in the Yeast GIN (namely ACDD, CN, ncMCE AUC and ncISO AUC).

In Figure 5.4, we report the number of interactions (amongst the top-100 for each technique) that overlapped with DRYGIN as a deviation from random prediction (we intersected the top 100 interactions advocated by 100 different random predictions and averaged for comparison with the best link predictors). This deviation is provided in dB, which is commonly used to specify the ratio
of a quantity (in this case the precision, computed as the number of top-100 candidates that overlap with DRYGIN) relative to a reference level (in this case the average precision of a random predictor, computed as the average number of randomly proposed candidates that overlap with DRYGIN: $10 \cdot \log_{10} \frac{\text{overlap}_{\text{link predictor}}}{\text{overlap}_{\text{random predictor}}}$). We can see that the technique whose overlap with DRYGIN deviates the most from a random prediction baseline is ACDD.

All double mutants from DRYGIN are accompanied by a p-value that indicates how likely they are to be real (interactions with p-values that are at most 0.05 are more likely to occur). Figure 5.4 also shows that the GIIs advocated by ACDD and CN have a low average p-value ($\langle p \rangle$), which indicates that they are good candidates for genetic interaction. This last result provides a more wet-lab oriented validation of the top-100 candidate interactions advocated by the best prediction techniques presented in this chapter.

![Figure 5.4: Overlap of the top-100 yeast GIIs advocated by the best prediction techniques with DRYGIN. The overlap is reported in dB, which indicates the level of deviation from a random prediction baseline (0 dB). The average p-value ($\langle p \rangle$) reported in DRYGIN for the overlapped interactions is also provided.](image)

### 5.5 Conclusion

Computational prediction of GIIs is a powerful tool that can aid high-throughput GI determination by reducing the universe of possibilities to test in the lab. In this work, we propose a graph dissimilarity index that achieves high precision on a dense Yeast GIN and a very sparse Worm GIN (according to GO, STRING and DRYGIN), thus being able to point out good epistatic candidate interactions. Further studies in this regard would involve validation using gene expression across
different conditions and, of course, experimental proof. Moreover, prediction of GIs using ACDD in higher organisms, such as human, could advance our understanding of human disease and our ability to design personalised treatment plans.

Finally, we remarkably proved that the network topology is shaped according to the information given by the biological properties of its components (genes and negative genetic interactions in this case). This is an important finding that confirms that computational approaches can be applied to mine such biological information from the mere structure of the network and utilise it for different purposes, prediction being one of them.
Chapter 6

Exploring the Genetics Underlying Autoimmune Diseases with Network Analysis and Link Prediction

Abstract

Ever since the first Genome Wide Association Study (GWAS) was carried out we have seen an important number of discoveries of biological and clinical relevance. However, there are some scientists that consider that these research outcomes and their utility are far from what was expected from this experimental design. We instead believe that the thousands of genetic variants associated with complex disorders by means of GWASs are an extremely valuable source of information that needs to be mined in a different way. Based on this philosophy, we followed a holistic perspective to analyse GWAS data and explored the structural properties of the network representation of one of these datasets with the aim to advance our understanding of the genetic intricacies underlying autoimmune human diseases. The simplicity, computational efficiency and precision of the tools proposed in this paper represent a new means to address GWAS data and contribute to the better exploitation of these rich sources of information.
Despite the thousands of genetic variants associated with human complex traits since the first GWAS (The Wellcome Trust Case Control Consortium, 2007), the independent contribution of variants to genetic risk has failed to fully explain the heritability of human diseases (Bloom et al., 2013; McKinney and Pajewski, 2011). It has been long recognised that bio-process regulation is possible only thanks to the relationships between genes and the proteins they produce (McKinney and Pajewski, 2011), which advocates the analysis of GWAS data from a network perspective.

Day by day, more comprehensive and complex biological datasets become available and prompt us to make use of smart strategies to quickly and accurately analyse these valuable sources of information. One such a strategy, the emerging field of Network Medicine, not only allows for the systematic study of the molecular intricacies underlying complex traits but it also serves as a proxy to identify the relationships between apparently different phenotypes (Barabási et al., 2011).

While metabolic, protein interaction, virus-host and regulatory networks are important to understand the interdependencies between molecular components in the cell (Vidal et al., 2011; Barabási et al., 2011), genetic interaction networks (or epistatic networks) represent abstract relationships between genes associated with genetic buffering. It is thought that these networks can be the key to uncover the missing heritability that single gene effects cannot explain (McKinney and Pajewski, 2011). Unfortunately, although well-established techniques exist to detect gene interactions in model organisms (Lehner, 2013), their determination in human is still work-in-progress.

In 2007, Goh and colleagues (Goh et al., 2007) proposed a work around this shortcoming and built a human disease gene network by means of a method that we here exploit to construct an Autoimmune Disease Network (ADN). We decided to focus on autoimmune disorders due to their relevance in terms of people affected, direct health care costs and research funding when compared to, for example, cancer (23.5 million vs 9 million Americans, 100 billion vs 57 billion USD and 591 million vs 6.1 million USD, respectively) (American Autoimmune Related Diseases Association, Inc., 2012). We study the basic topological properties of the ADN, the GO and expression similarities of its constituting genes and the modules they form. We introduce the weighted version of link predictors that we proposed in recent publications and benchmark them against classical prediction techniques to later choose the best approach and analyse the biological implications of its top-scored candidate interactions.
6.1 Network Representation of GWAS Data

6.1.1 Autoimmune Disease Network

Based on the list of autoimmune diseases curated by the American Autoimmune Related Disease Association (American Autoimmune Related Diseases Association, Inc., 2012), we extracted genes associated with autoimmune disorders from the GWAS Catalog (Hindorff et al., 2012) and constructed a bipartite network of diseases and their associated genes (Figure 6.1, left panel). We then projected this network to its one-mode form, where a pair of genes is linked with a weighted edge indicating the number of diseases they are both associated with (Figure 6.1, right panel). We refer to this undirected, weighted network as the ADN.

![Bipartite network and weighted one-mode projection](image)

**Figure 6.1:** One-mode projection of the bipartite network of genes and diseases. In the highlighted example, gene c is associated with diseases B and D whereas gene e is associated with diseases A, B, C and D. Since they have two diseases in common (B and D), they are linked with a weight of 2 in the projection of the bipartite network to the gene space.

6.1.2 Structural Properties of the ADN

Most real networks, representing dynamic and heterogeneous systems, present the small-world property (high clustering coefficient and low average shortest path length when compared to ER random graphs) (Watts and Strogatz, 1998), have heavy-tailed node degree distributions (which indicate a high degree of self-organisation and presence of hubs, i.e. highly connected entities) (Barabási and Albert, 1999) and their structure follows the LCP (common neighbours of connected node pairs are also connected) (Cannistraci et al., 2013c). We compared the small-world features of 1000 ER graphs with the same number of nodes and edges as the ADN (840 and 49,485 respectively) and found that the latter is far from being random: it has a clustering coefficient of 0.902 (ER graph
= 0.14 ± 0.007) and an average shortest-path of 2.009 (ER graph = 1.86 ± 0.006), i.e. it does present the small world property.

Although the node degree distribution of the ADN does not decay as a power law, it presents a right heavy tail that indicates the presence of network hubs (Figure 6.2, left panel). Furthermore, the correlation between the number of common neighbours of connected genes in the network and the square root of the number of connections between them (LCP-corr) is very high (0.97) as depicted by the LCP-DP of Figure 6.2, right panel. The very high values of clustering (quite common in one-mode projections of bipartite networks (Opsahl, 2013)) and LCP-corr indicate that the network is organised in tightly connected communities of genes and we can thus take advantage of this to discover disease modules and perform neighbourhood-based link prediction.

Figure 6.2: The node degree distribution and LCP-DP of the ADN.

### 6.2 Community Detection

After the application of a community detection algorithm based on Modularity Optimisation (maximisation of community quality by finding groups of nodes that are densely connected internally by heavy edges and that, at the same time, are connected between them by only a few light edges) (Blondel et al., 2008), we found that genes associated with the same or with related diseases clustered together (Figure 6.3).
6.2.1 Biological Properties of the Detected Communities

The majority of the detected communities are part of the main component of the ADN and contain genes associated with related diseases. Nevertheless, a couple of modules formed by genes associated with a particular disease are also present (Figure 6.3) and disconnected from the main network component, as it is the case for the Restless Legs Syndrome or Dilated Cardiomyopathy communities. This is indicative that such genes are very specific of those diseases.

GO Semantic Similarity (Wang et al., 2007) and Gene CoExpression (Obayashi et al., 2008) computation for each community and the complete network, shows large values of the former and low of the latter (see bars next to each community in Figure 6.3), which indicates that the analysed pairs of genetic variants can be expressed in, for example, different tissues and yet have important phenotypic impact.

As mentioned above, community detection grouped genes associated with related diseases (e.g. the green community of genes associated with inflammatory disorders) but this was not always the
case, like in the pink and grass coloured communities (Figure 6.3). The first one clusters genes associated with Rheumatoid Arthritis (RhA) and Celiac Disease (CD), two disorders with two very different phenotypes: chronic joint inflammation and intestine inflammation respectively but that can nevertheless show some common features: HLA molecule association, T-cell infiltration in target organs and high degree of co-morbidity (Coenen et al., 2009). The diseases represented in the second one, Type I Diabetes (TID) (autoimmune destruction of insulin-producing β-cells) and Vitiligo (skin depigmentation due to melanocyte malfunction), are also phenotypically different but it is known that patients with TID tend to present skin pigmentation problems, especially in areas of insulin injection (Burge and Carey, 2004).

6.3 Link Prediction

We assessed the performance of multiple topological link predictors (indices that assign likelihood scores to non-adjacent pairs of nodes based on the network topology) when applied to the communities that are not completely connected (namely the blue, pink, green, grass and magenta modules). This evaluation was done using the PRPES procedure. ERPES using the GO was not used because the genetic relationships in the ADN indicate co-association with autoimmune disorders, which is not necessarily correlated with shared genetic function or participation in similar biological processes.

6.3.1 Weighted Link Predictors

As mentioned in Section 6.1.1, the ADN is a weighted network and because of this, the weighted version of the classical prediction techniques (Murata and Moriyasu, 2007) - namely Weighted Common Neighbours (wCN), Weighted Preferential Attachment (wPA) and Weighted Adamic & Adar (wAA) - was considered in this chapter (see Table 6.1, Equations 6.3.1-6.3.3). Besides, we introduce the weighted version of two techniques that we proposed in recent publications for the prediction of epistatic interactions (Alanis-Lobato et al., 2013) (Weighted Adjusted Czekanowski-Dice Dissimilarity (wACDD), Table 6.1, Equation 6.3.4) and links in general (Weighted Cannistraci-Alanis-Ravasi (wCAR), Table 6.1, Equation 6.3.5) (Cannistraci et al., 2013c).

<table>
<thead>
<tr>
<th>Weighted Predictor</th>
<th>Formula</th>
<th>Eq. Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>wCN(x, y)</td>
<td>$wCN(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{w(x, z)}{2} \frac{w(z, y)}{2}$</td>
<td>(6.3.1)</td>
</tr>
<tr>
<td>wPA(x, y)</td>
<td>$wPA(x, y) = k_x \times k_y$</td>
<td>(6.3.2)</td>
</tr>
<tr>
<td>wAA(x, y)</td>
<td>$wAA(x, y) = wCN(x, y) \times \frac{\sum_{z \in \Gamma(x) \cap \Gamma(y)} w(x, z) w(z, y)}{\sum_{z \in \Gamma(x) \cap \Gamma(y)} w(z, y)}$</td>
<td>(6.3.3)</td>
</tr>
<tr>
<td>wACDD(x, y)</td>
<td>$wACDD(x, y) = \frac{2wCN(x, y)}{k_x + k_y}$</td>
<td>(6.3.4)</td>
</tr>
</tbody>
</table>

Table 6.1: Weighted link predictors

Continued on next page
Table 6.1 – continued from previous page

<table>
<thead>
<tr>
<th>Weighted Predictor</th>
<th>Formula</th>
<th>Eq. Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>wCAR</td>
<td>$w_{CAR}(x, y) = w_{CN}(x, y) \times w_{LCL}$</td>
<td>(6.3.5)</td>
</tr>
</tbody>
</table>

The penalisation parameters in wACDD correspond to $\lambda_x = \max(0, n_{avg} - k_x')$, $n_{avg}$ is the weighted average node degree of the network, $k_x'$ is the weighted degree of node $x$ and wLCL is the sum of the weights of the links connecting the common neighbours of $x$ and $y$ (local community links).

### 6.3.2 Prune-rediscover Performance Evaluation

Figure 6.4 depicts the very good performance of all indices and highlights the sensitivity of wCAR to the loss of the local community structure of the network: as links are removed from the topology, local communities disappear and wCAR loses precision. The power of neighbourhood-based approaches is expected, given the high values of clustering and LCP-corr of the network. These structural characteristics explain why wPA is unable to mimic the performance of the rest: simple weighted degree products are not enough for prediction in a network dominated by gene pairs that share lots of neighbours.

Although wCN, wAA and wACDD present very similar performance patterns, the latter outperforms the other techniques in most of the disease modules. We analysed the biological relevance of the top-3 candidate interactions advocated by this approach for each of the gene communities in which prediction was possible (see next Section).

### 6.3.3 Biological Interpretation of Best Candidate Interactions

The biological analysis that led to the conclusions made in this section, was performed with DAVID Bioinformatics Resources 6.7 (Huang et al., 2009a,b).

The top candidate links in the blue community correspond to interactions of genes AFF1, AGXT2L1 and AG1P7 with INSR. The products of these genes are involved in proper blood filtering and toxin secretion. Thus, their simultaneous mutations may explain the blood filtering problems in IgA Nephropathy or the affection of blood vessels in Kawasaki Disease or Lupus.

The top candidates in the pink community are links between HLA-DQA2, HLA-DQB1 (T-cell multiplication) and LPP (cell adhesion and motility), which could explain chronic inflammation, and the link between C1QTNF6 (regulation of immune cells) and IL2 (regulation of Th17 cells), which could explain the high levels of Th17 cells in the affected tissues of RhA and CD patients.

The top candidates edges in the green community are TMEM17-COG6, COG6-B3GNT2 and B3GNT2-PARK7. Genetic variants in TMEM17, COG6 and B3GNT2 would directly impact protein transport and their interaction with PARK7 could produce aberrations in the proper protection against cell stress.
Figure 6.4: Performance evaluation of classical and proposed weighted link predictors.

For the grass community, the top candidate links are the interactions of ADPRH, ATP8B1 and CASP7 with C14orf181. The latter is an uncharacterised gene that has been recently removed from several databases due to the presence of a more interesting gene in the same chromosome. However, not only was it part of our top-3 candidates but also of the top 10. Given the association of ATP8B1 with bile disorders and the high expression of CASP7 in liver, it would be interesting to put more attention to C14orf181.

Finally, the set of top candidate interactions for the magenta community are CAST-CCR5, CCR5-ANXA6 and ANXA6-CCRL2, all involved in Calcium (Ca) regulation and intracellular mobilisation. This means that mutations in this set of genes may impact signalling by Ca ions and explain Ca-associated problems in AIDS or Psoriasis. Ca-regulation by the thyroid gland could also be affected.
6.4 Conclusions

We analysed GWAS data from a systems perspective, which deems autoimmune disease genes not as isolated entities but as important parts of a whole. The value of the disease modules detected by Modularity Optimisation and the precision of the link predictors introduced, indicate that they are powerful and trustworthy tools. We encourage researchers to integrate Systems Biology and Network Science into their projects because the knowledge generated by these scientific devices could help us untangle the intricate relationships between genotype and phenotype.
Chapter 7

Concluding Remarks

7.1 Summary and Significance of this Work

This document presents different means by which we can take advantage of the structural organisation of complex networks (especially PPINs and GINs) to tackle the link prediction problem. Link prediction is important not only to determine which acquaintances are more likely to emerge in social networks but also to reduce the universe of protein or genetic pairs to test for interaction in the lab and, consequently, produce more complete and reliable biological interactomes, a crucial task in systems biology.

The LCP, along with CAR, the LCP-DP and the LCP-corr, represent a first attempt to advance a link/community-based interpretation of the epitopological learning component that appears in many cognitive, social and evolutionary processes. Our results with ncMCE represent a promising achievement and encouragement to deepen the investigation of network embedding techniques for topological prediction of protein and genetic interactions and highlight the importance of spanning trees as complex network skeletons and discrete representations of the geometry underlying them. ACDD represents an example of how simple neighbourhood-based approaches can be improved by combining the ideas from different network paradigms and encourages the investigation of how the power of some link predictors can be boosted when in combination with others. Finally, the network representation of autoimmune disease data shows how valuable and informative network-based techniques are and how a holistic perspective to GWAS data can advance our knowledge of complex diseases. Proof of the impact of the tools developed in this thesis, is the attention that they are gaining from other research groups, who are already comparing their approaches with ours and are building upon them to improve their predictive power (Liu et al., 2013; Wang et al., 2013).

The link predictors, network-based pipeline and link formation paradigm here introduced can be
applied to networks from different domains but they acquire more significance when one considers that their application to PPINs and GINs could lead to a better understanding of the complex wiring diagram underlying the living cell. Only when we have complete protein and genetic interaction maps will we be able to perturb certain system components without affecting others and only then will we be able to design more effective treatments against rare human disorders.

7.2 Future Research Work

As with any project, the work here presented could be extended in different directions. From a general perspective, the link prediction problem could be more formally addressed as a positive-unlabelled data problem (Elkan and Noto, 2008; Zhang and Lee, 2008). The amount of positive (the links in the observed network $A^0$) and unlabelled (the disconnected node-pairs in $A^0$) examples in network datasets, make of this problem the ideal candidate for training of semi-supervised classifiers to determine the set of links from the unlabelled examples that are more likely to be real interactions. However, in the case of biological interactomes, one has to account for the possible FPs inside the set of positive examples. The performance evaluation process also needs to be carefully revisited because uniform random removal of links could lead to an ill-posed evaluation if one considers that real networks do not grow links uniformly at random (see Sections 4.5.7 and 5.3).

As Liben-Nowell and Kleinberg already stated it (Liben-Nowell and Kleinberg, 2007), the best way to evaluate link predictors is to assess their ability to predict links that appear in a newer temporal snapshot of a network when applied over an older topology. The issue here is that access to the temporal evolution of a network is not easy, especially in the case of biological interactomes. In addition, once the performance evaluation process is refined, the field of topological link prediction would benefit from an extensive comparison between the predictors proposed in this thesis and the majority of the extant ones. This could result in the generation of a more adaptive link prediction technique able to advocate for good candidates for interaction in different topologies, regardless of their sparsity or level of noise.

From the neighbourhood-based predictors’ perspective, a proper analysis of how they may be combined to achieve higher precisions should be carried out. This is not an easy task due to the great diversity of ways in which a predictor can be normalised (take CN and its two different normalisations, JC and CDD, as an example).

From the embedding-based predictors’ perspective, provided the success of $H^2$ and Popularity-Similarity models (see Sections 2.1.1 and 4.6) it is worth trying to develop embedding techniques
able to map networks directly to hyperbolic spaces in an efficient manner.

From the weighted link predictors’ perspective, the advantages of predicting using and dismissing the weights of a network’s links should be scrutinised. Especially taking into account that more reliable biological network datasets with edge information are being made available (Koh et al., 2010; Szklarczyk et al., 2011).

Finally, the protein and genetic interactions advocated by the predictors introduced here (like CAR, ncMCE and ACDD) must be validated in the wet-lab to give further support to our findings.
REFERENCES


A Network Datasets

This study used six types of real interaction networks, for a total of 48 networks, which represent differing systems and have their own meaning and characteristics: biological networks, social networks, food webs, atomic-level networks, power grids and road networks. Detailed information about each network is listed in Table A.1 and the LCP-DPs for most of them appear in Figures A.1, A.2 and A.3. The rest are part of the main text of this thesis.

Almost all the categories mentioned above presented a high LCP-corr, ranging from 0.84 to 0.99. The Power Grid and the Karate Club network are borderlines with LCP-corr = 0.78 and LCP-corr = 0.75 respectively, as well as the Grassland Species Food Web with LCP-corr = 0.42. The road network values range from 0 to 0.16. Other networks analysed (most of them representing atomic interactions), present a clear LCP-corr of 0 (there are either no common neighbours between their interactors or there are no links between common neighbours if they exist).

<table>
<thead>
<tr>
<th>ID</th>
<th>Dataset</th>
<th>N</th>
<th>L</th>
<th>LCP-corr</th>
<th>Directionality</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yeast PPIN (Ben-Hur and Noble, 2005)</td>
<td>4036</td>
<td>10411</td>
<td>0.92</td>
<td>Undirected</td>
<td>Interactions between proteins in Saccharomyces cerevisiae (yeast). Redundant and self-interactions were removed from this network and only the largest connected component was considered.</td>
</tr>
<tr>
<td>2</td>
<td>Yeast PPIN (Chen et al., 2006a)</td>
<td>4385</td>
<td>12234</td>
<td>0.95</td>
<td>Undirected</td>
<td>Interactions between proteins in yeast. Redundant and self-interactions were removed from this network and only the largest connected component was considered.</td>
</tr>
<tr>
<td>3</td>
<td>Yeast PPIN sparse (You et al., 2010)</td>
<td>3645</td>
<td>12934</td>
<td>0.90</td>
<td>Undirected</td>
<td>Interactions between proteins in yeast. Redundant and self-interactions were removed from this network and only the largest connected component was considered.</td>
</tr>
<tr>
<td>4</td>
<td>Yeast PPIN dense (You et al., 2010)</td>
<td>3883</td>
<td>29922</td>
<td>0.97</td>
<td>Undirected</td>
<td>Interactions between proteins in yeast. Redundant and self-interactions were removed from this network and only the largest connected component was considered.</td>
</tr>
<tr>
<td>5</td>
<td>Worm PPIN (Razick et al., 2008)</td>
<td>4743</td>
<td>18752</td>
<td>0.91</td>
<td>Undirected</td>
<td>Interactions between proteins in Caenorhabditis elegans (worm).</td>
</tr>
<tr>
<td>6</td>
<td>Fly PPIN (Razick et al., 2008)</td>
<td>7809</td>
<td>71211</td>
<td>0.86</td>
<td>Undirected</td>
<td>Interactions between proteins in Drosophila melanogaster (fly).</td>
</tr>
<tr>
<td>7</td>
<td>Mouse PPIN (Razick et al., 2008)</td>
<td>2969</td>
<td>4033</td>
<td>0.85</td>
<td>Undirected</td>
<td>Interactions between proteins in Mus musculus (mouse).</td>
</tr>
<tr>
<td>8</td>
<td>Human PPIN (Razick et al., 2008)</td>
<td>11816</td>
<td>83422</td>
<td>0.91</td>
<td>Undirected</td>
<td>Interactions between proteins in Homo sapiens (human).</td>
</tr>
<tr>
<td>9</td>
<td>Yeast GIN (Costanzo et al., 2010)</td>
<td>3842</td>
<td>52179</td>
<td>0.85</td>
<td>Undirected</td>
<td>Interactions between genes in yeast. Redundant and self-interactions were removed from this network and only the largest connected component was considered.</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>ID</th>
<th>Dataset</th>
<th>N</th>
<th>L</th>
<th>LCP-corr</th>
<th>Directionality</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Worm GIN (Byrne et al., 2007)</td>
<td>457</td>
<td>1242</td>
<td>0.90</td>
<td>Undirected</td>
<td>Interactions between genes in worm. Redundant and self-interactions were removed from this network and only the largest connected component was considered.</td>
</tr>
<tr>
<td>11</td>
<td>Worm nervous system connectome (Watts and Strogatz, 1998)</td>
<td>297</td>
<td>2345</td>
<td>0.91</td>
<td>Directed</td>
<td>Synaptic interactions between neurons in worm.</td>
</tr>
<tr>
<td>12</td>
<td>Worm global connectome (Kaiser and Hilgetag, 2006)</td>
<td>277</td>
<td>1918</td>
<td>0.93</td>
<td>Directed</td>
<td>Global connectome of worm.</td>
</tr>
<tr>
<td>13</td>
<td>Worm rostral ganglia neuro-synaptic connectome (Kaiser and Hilgetag, 2006)</td>
<td>131</td>
<td>687</td>
<td>0.84</td>
<td>Undirected</td>
<td>Rostral ganglia (anterior, dorsal, lateral, and ring) synaptic interactions in worm.</td>
</tr>
<tr>
<td>14</td>
<td>Worm chemical synaptic connectome (Varshney et al., 2011)</td>
<td>279</td>
<td>1961</td>
<td>0.94</td>
<td>Directed</td>
<td>Chemical synapse network of worm neurons.</td>
</tr>
<tr>
<td>15</td>
<td>Worm gap junction connectome (Varshney et al., 2011)</td>
<td>279</td>
<td>514</td>
<td>0.87</td>
<td>Directed</td>
<td>Gap junction network of worm neurons.</td>
</tr>
<tr>
<td>16</td>
<td>Macaque brain connectome (Kötter, 2004)</td>
<td>94</td>
<td>1515</td>
<td>0.97</td>
<td>Directed</td>
<td>Macaque cortical connectivity connectome within one hemisphere.</td>
</tr>
<tr>
<td>17</td>
<td>Mouse visual cortex neuro-synaptic connectome (Bock et al., 2011)</td>
<td>18</td>
<td>42</td>
<td>0.86</td>
<td>Directed</td>
<td>Synaptic interactions between neurons in the primary visual cortex (layers 1, 2/3 and upper 4) in mouse.</td>
</tr>
<tr>
<td>18</td>
<td>Autoimmune disease network (Alanis-Lobato et al., 2014)</td>
<td>840</td>
<td>49485</td>
<td>0.97</td>
<td>Undirected</td>
<td>One-mode projection to the gene space of the bipartite network of autoimmune diseases and their associated genes. This network is weighted.</td>
</tr>
<tr>
<td>19</td>
<td>Food web (Cohen et al., 2009)</td>
<td>51</td>
<td>241</td>
<td>0.90</td>
<td>Directed</td>
<td>Which taxon eats which in Tuesday Lake, Michigan.</td>
</tr>
<tr>
<td>20</td>
<td>Dolphin associations (Lusseau et al., 2003)</td>
<td>62</td>
<td>318</td>
<td>0.89</td>
<td>Undirected</td>
<td>Frequent associations between dolphins in New Zealand.</td>
</tr>
<tr>
<td>21</td>
<td>Political blogosphere of the US Elections (Adamic and Glance, 2005)</td>
<td>1490</td>
<td>19025</td>
<td>0.93</td>
<td>Directed</td>
<td>Incoming and outgoing posts on blogs at the time of the 2004 US presidential election.</td>
</tr>
<tr>
<td>22</td>
<td>WikiVote (Leskovec et al., 2010)</td>
<td>7115</td>
<td>103689</td>
<td>0.93</td>
<td>Directed</td>
<td>Who votes whom to be a Wikipedia administrator.</td>
</tr>
<tr>
<td>25</td>
<td>American/Canadian Flight Map (Frey and Dueck, 2007)</td>
<td>456</td>
<td>37947</td>
<td>0.99</td>
<td>Undirected</td>
<td>Flights between American and Canadian cities.</td>
</tr>
<tr>
<td>26</td>
<td>College Football (Girvan and Newman, 2002)</td>
<td>115</td>
<td>613</td>
<td>0.89</td>
<td>Undirected</td>
<td>Network representation of the schedule of American football games.</td>
</tr>
<tr>
<td>27</td>
<td>Hydrogen bonds between residues in a protein (Martin et al., 2011)</td>
<td>164</td>
<td>876</td>
<td>0.90</td>
<td>Undirected</td>
<td>Hydrogen bond network of human GPX4.</td>
</tr>
<tr>
<td>28</td>
<td>van der Waals contacts between residues in a protein (Martin et al., 2011)</td>
<td>248</td>
<td>1979</td>
<td>0.88</td>
<td>Undirected</td>
<td>van der Waals contact network of human TIM barrel.</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>ID</th>
<th>Dataset</th>
<th>N</th>
<th>L</th>
<th>LCP-corr</th>
<th>Directionality</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>Power grid (Watts and Strogatz, 1998)</td>
<td>4941</td>
<td>13188</td>
<td>0.78</td>
<td>Undirected</td>
<td>Power grid of the western states of the USA.</td>
</tr>
<tr>
<td>30</td>
<td>Zachary’s karate club (Zachary, 1977)</td>
<td>34</td>
<td>78</td>
<td>0.75</td>
<td>Undirected</td>
<td>Friendships between members of a karate club in the US.</td>
</tr>
<tr>
<td>31</td>
<td>Grassland species (Dawah et al., 1995)</td>
<td>75</td>
<td>113</td>
<td>0.42</td>
<td>Directed</td>
<td>Food web of grassland species.</td>
</tr>
<tr>
<td>32</td>
<td>San Joaquin road network (Brinkhoff, 2002)</td>
<td>18263</td>
<td>23797</td>
<td>0</td>
<td>Undirected</td>
<td>Road map of San Joaquin County, California.</td>
</tr>
<tr>
<td>33</td>
<td>San Francisco road network (Brinkhoff, 2002)</td>
<td>174956</td>
<td>221802</td>
<td>0.16</td>
<td>Undirected</td>
<td>Road map of San Francisco, California.</td>
</tr>
<tr>
<td>34</td>
<td>California road network (Li et al., 2005)</td>
<td>21048</td>
<td>21693</td>
<td>0</td>
<td>Undirected</td>
<td>Road map between cities in California.</td>
</tr>
<tr>
<td>35</td>
<td>Oldenburg road network (Brinkhoff, 2002)</td>
<td>6105</td>
<td>7035</td>
<td>0</td>
<td>Undirected</td>
<td>Road map of Oldenburg, Germany.</td>
</tr>
<tr>
<td>36</td>
<td>North America road network (<a href="http://www.cs.fsu.edu/SpatialDataset.htm">http://www.cs.fsu.edu/SpatialDataset.htm</a>)</td>
<td>175813</td>
<td>179102</td>
<td>0</td>
<td>Undirected</td>
<td>Road map of North America.</td>
</tr>
<tr>
<td>37</td>
<td>California road network (Leskovec et al., 2009)</td>
<td>1965206</td>
<td>2766607</td>
<td>0.13</td>
<td>Undirected</td>
<td>Road map of California, USA in which nodes are both cities and road intersections.</td>
</tr>
<tr>
<td>38</td>
<td>Texas road network (Leskovec et al., 2009)</td>
<td>1379917</td>
<td>1921660</td>
<td>0.16</td>
<td>Undirected</td>
<td>Road map of Texas, USA.</td>
</tr>
<tr>
<td>39</td>
<td>Pennsylvania road network (Leskovec et al., 2009)</td>
<td>1088092</td>
<td>1541898</td>
<td>0.15</td>
<td>Undirected</td>
<td>Road map of Pennsylvania, USA.</td>
</tr>
<tr>
<td>40</td>
<td>German highway (Kaiser and Hilgetag, 2004)</td>
<td>1168</td>
<td>1243</td>
<td>0</td>
<td>Undirected</td>
<td>German highway system.</td>
</tr>
<tr>
<td>41</td>
<td>Ice</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>Undirected</td>
<td>Bonds between atoms in ice.</td>
</tr>
<tr>
<td>42</td>
<td>Diamond</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>Undirected</td>
<td>Bonds between atoms in diamond.</td>
</tr>
<tr>
<td>43</td>
<td>Graphite</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>Undirected</td>
<td>Bonds between atoms in graphite.</td>
</tr>
<tr>
<td>44</td>
<td>Fullerene</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>Undirected</td>
<td>Bonds between atoms in fullerene.</td>
</tr>
<tr>
<td>45</td>
<td>DNA</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>Undirected</td>
<td>Bonds between atoms in DNA.</td>
</tr>
<tr>
<td>46</td>
<td>$\beta$-sheet</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>Undirected</td>
<td>Bonds between residues in a $\beta$-sheet-like protein structure.</td>
</tr>
<tr>
<td>47</td>
<td>$\alpha$-helix</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>Undirected</td>
<td>Bonds between residues in an $\alpha$-helix-like protein structure.</td>
</tr>
<tr>
<td>48</td>
<td>Abiraterone drug</td>
<td>–</td>
<td>–</td>
<td>0</td>
<td>Undirected</td>
<td>Bonds between atoms in the chemical structure of abiraterone drug.</td>
</tr>
</tbody>
</table>

N and L is the number of nodes and links in the network respectively. Cells with a – correspond to values that have not been computed because the dataset has not been assembled or that are not applicable to that network. Cells with green background correspond to biological networks, cells with red background correspond to social networks, cells with brown background correspond to food webs, cells with orange background correspond to atomic-level networks, cells with blue background correspond to power grids and cells with grey background correspond to road maps.
Figure A.1: **LCP-DP for networks of biological origin.** The x-axis indicates the number of common neighbours and the y-axis the square root of the number of links between them. Each point in the plot is an interaction from the network. The black, dashed line represents the LCL’s upper bound.
Figure A.2: **LCP-DP for social networks.** The x-axis indicates the number of common neighbours and the y-axis the square root of the number of links between them. Each point in the plot is an interaction from the network. The black, dashed line represents the LCL’s upper bound.
Figure A.3: **LCP-DP for networks of different origins.** LCP-DP for atomic-level networks (top-left and top-centre) with significant LCP-corr. The Power Grid network (top-right), the Karate Club network (bottom-left), and the Grassland Species Food Web (bottom-centre) represent the borderline cases that we identified. The x-axis indicates the number of common neighbours and the y-axis the square root of the number of links between them. Each point in the plot is an interaction from the network. The black, dashed line represents the LCL’s upper bound.
B Computational Time
Complexity of Prediction Techniques

Neighbourhood-based Approaches

Let an undirected interaction network be represented by an undirected graph $G = (V, E)$, where $V$ is the set of nodes and $E$ is the set of links or interactions between such nodes ($E$ is a set of 2-element, unordered subsets of $V$). Let $N = |V|$ be the total number of nodes in the network and $L = |E|$ be the total number of links. Then the total number of candidate links (i.e. the non-directly connected pairs of nodes in $G$) a prediction technique has to score is:

$$\frac{N(N - 1)}{2} - L$$  \hspace{1cm} (B.0.1)

B.0.1 can be a very large number if $G$ is very sparse, however since most neighbourhood-based prediction techniques depend on the number of common neighbours between the non-directly connected nodes in the network (see Table 3.1), not all candidate links resulting from B.0.1 above need to be scored: non-adjacent nodes separated by more than two edges have an empty common neighbourhood and they can account for up to 90% of B.0.1 (Lü et al., 2009). If the number of node-pairs separated by more than two edges is $k$, the number depicted in Equation B.0.1 is brought down to

$$\frac{N(N - 1)}{2} - L - k.$$

Set operations like the intersection and the union between two sets (say $A$ and $B$) can be implemented so that they run in $O(|A|)$ time (given $|A| \leq |B|$), using hash-tables with $O(1)$ look-up operations (see Algorithm 2). Thus, for each of the candidate links to be scored, $O(|A|)$ computational time is required for set operations (with $|A|$ being the typical size of the smallest neighbourhood of the seed nodes in the network), which being a constant, leaves the overall computational time
complexity for a neighbourhood-based predictor as $O(N^2 - L - k)$. Notice this bound does not apply to indices like IG1 or ACDD because their formulation does not depend only on the set intersection (see Table 3.1), which means that they have to score all non-adjacent nodes and their computational time complexity is $O(N^2 - L)$.

**Algorithm 2:** Sample pseudocode for set intersection

```plaintext
function setIntersection(A, B):
    foreach e in A do
        if e in B then
            C.add(e);
        end
    end
    return C;
```

**Network Embedding Approaches**

The first step of network embedding techniques is the computation of all-pair shortest paths over the graph representation of the network (or over the MST, which can be found in $O(L \log_2 N)$ time if Kruskal’s algorithm is used, like in MCE/ncMCE). Johnson’s algorithm for all-pair shortest-path computation runs in $O(N^2 \log_2 N + NL)$ (Cormen et al., 2001) which can outperform Floyd-Warshall algorithm when the network is sparse (if $L$ is very large, $L \approx N^2$ and $NL \approx N^3$ which makes the algorithm run in $O(N^3)$ time).

If network centring is performed, some matrix multiplications are needed, which, if implemented naïvely, run in $O(N^3)$ time.

SVD computation of an $N \times N$ matrix requires $O(N^3)$ computational time in the worst case (Navab, 2005). However, economy size SVD is able to embed the matrix to dimension $d$ in $O(d^2N)$ time.

Finally, all-pair euclidean distance calculation between points embedded in the low dimensional space requires $O(N^2)$ plus an additional $O(N^2 \log_2 N + NL)$ computation of all-pair shortest paths over the embedded network (when the SP approach for candidate scoring is used).

In conclusion, network embedding techniques run in $O(N^3)$ computational time when network centring is performed and in $O(N^2 \log_2 N + NL)$ when the network is not centred. This last bound can be considered to be in the order of the square of the number of nodes for the networks analysed in this document.
C Published and Submitted Papers


