Imperial College London
Department of Computing

Graphlet Correlations for Network Comparison and Modelling: World Trade Network Example

Ömer Nebil Yaveroğlu

December 2013

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College London
and the Diploma of Imperial College London

Declaration

I herewith certify that all material in this dissertation which is not my own work has been properly acknowledged.

Ömer Nebil Yaveroğlu

Copyright

Abstract

We propose methods on two fundamental graph theoretic problems: (1) *network comparison*, and (2) *network modelling*. Our methods are applied to five real-world network types, with an emphasis on world trade networks (WTNs), which we choose due to the world's current economic crisis.

Finding topological similarities of complex networks is computationally intractable due to NP-Completeness of the subgraph isomorphism problem. Hence, simple heuristics have been used for this purpose. The most sophisticated heuristics are based on graph spectra and small subnetworks including graphlets. Among these, graphlets are preferred since spectra do not provide a direct real-world interpretation of network structure. However, current graphlet-based techniques can be improved. We improve graphlet-based heuristics by defining a new network topology descriptor, *Graphlet Correlation Matrix (GCM)*, which eliminates all redundancies and quantifies the dependencies in graphlet properties. Then, we introduce a new network distance measure, *Graphlet Correlation Distance (GCD)*, that compares GCMs of two networks. We show that GCD has the best network classification performance, is highly noise-tolerant, and is computationally efficient. Using this methodology, we highlight a three-layer organization in the WTNs: core, broker, and periphery. Furthermore, we uncover the link between the dynamic changes in oil price and trade network topology.

Network models should shed light on the rules governing the formation of real networks. Using GCD, we identify models that fit five real-world network types. However, none of these standard network models fit WTNs. Hence, we introduce two new network models: one that mimics the Gravity Model of Trade, and the other that mimics brokerage / peripheral positioning of a country in WTN. Also, we show that economic wealth indicators of a country are predictive of its future brokerage position. Finally, we use exponential-family random graph modelling approach to build a generic framework that enables modelling based on any graphlet property.

To my parents, family and friends...
None of this would have been possible without your precious support.

Acknowledgements

I am grateful to my supervisor, Professor Nataša Pržulj, for her guidance and support on my research. Prof. Pržulj's enthusiasm on always targeting higher motivated me to be more productive in this study, and gain a lot more than I expected from my studies. Prof. Pržulj's experience in research also helped me to develop skills required for being an independent researcher.

I would like to thank my examination committee members, Prof. Duncan F. Gillies and Prof. Desmond Higham, for their time and insightful feedback on this dissertation. I also want to thank my second supervisor, Prof. Marek Sergot, for his helpful mentoring throughout my studies.

I am grateful to all the past and present members of Prof. Pržulj's research group: Dr. Noël Malod-Dognin, Dr. Joana Gonçalves, Kai Sun, Anida Sarajlic, Vuk Janjic, Miles Mulholland, and Yulian Ng. Apart from your scientific contribution to my work with our discussions, your friendship and support has been very valuable to me over the years.

I want to thank Dr. Aleksandar Stojmirovic for initiating great ideas for my studies, and also for guiding the implementation of these ideas with his insightful discussions. Our collaborators did not only help us develop excellent ideas, but also made our research more exciting and dynamic by supplying us with new ideas from different disciplines. I sincerely thank Prof. Athina Markopoulou, Prof. Carter Butts, Dr. Maciej Kurant, Sean M. Fitzhugh, Darren Davis, Prof. Rasa Karapandza, and Prof. Zoran Levnajic for their collaboration.

I would like to acknowledge Prof. Sinan Kalkan and Hande Çelikkanat for their patience and effort in proof-reading this dissertation.

I am extremely grateful to my family, whom not only supported me by all means during my studies but also encouraged me to chase my dreams at the expense of being far away from me.

Finally, I would like to thank to Imperial College London – Department of Computing, European Research Council (ERC) Starting Independent Re-

searcher Grant 278212 (2012-2017), and USA National Science Foundation (NSF) Cyber-Enabled Discovery and Innovation (CDI) grant OIA-1028394 for funding my studies.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Network Representations

- $G(V, E)$ = A network G with the node set V and edge set E.
- $|V|$ = The number of nodes in a network.
- $|E|$ = The number of edges in a network.
- A = Adjacency matrix of a network.
- D = Diagonal Degree Matrix of a network.
- L = Laplacian Matrix of a network.

Network Properties

- $D(v)$ = Degree of node v.
- $CC(v)$ = Clustering Coefficient of node v.
- $C_c(v)$ = Closeness Centrality of node v.
- $C_b(v)$ = Betweenness Centrality of node v.
- $E(v)$ = Eccentricity of node v.
- $C_e(v)$ = Eccentricity Centrality of node v.
- $C_i(v)$ = Graphlet degree of node v for orbit i. Note that, i is an integer and $i \in [0, 72]$.
- CE_i = The number of graphlets counted by placing edge orbit i on the flipped edge during the MCMC process of ERGM parameter estimation.

- GDV = Graphlet Degree Vector.
- GCM = Graphlet Correlation Matrix.

Random Network Model Abbreviations

- ER = Erdös Rènyi Random Network Model.
- $ER - DD$ = Generalized Random Network Model.
- $SF - BA$ = Scale-free Barabasi-Albert (Preferential Attachment) Model.
- $SF - GD$ = Scale-free Model with Gene Duplication and Divergence Events.
- GEO = Geometric Random Network Model.
- $GEO - GD$ = Geometric Random Network Model with Gene Duplication and Divergence Events.
- $STICKY$ = Stickiness-index based Random Network Model.
- GR = Gravitational Random Network Model.
- $ERGM$ = Exponential-family Random Graph Model.

Network Distance Measures

- GCD = Graphlet Correlation Distance.
- $GCD - 11$ = Graphlet Correlation Distance that is computed from the 11 non-redundant orbits of 2- to 4-node graphlets.
- $GCD - 73$ = Graphlet Correlation Distance that is computed from the 73 orbits of all 2- to 5-node graphlets.
- $RGFD = RGF$ Distance= Relative Graphlet Frequency Distance.
- $GDDA = GDD$ Agreement= Graphlet Degree Distribution Agreement (computed from the arithmetic mean of the distributions).

Economic Wealth Indicators

- $RGDPL$ = Gross Domestic Product - derivation method 1
- $RGDPL2$ = Gross Domestic Product - derivation method 2
- $RGDPCH$ = Gross Domestic Product - derivation method 3
- KC = Consumption Share
- KG = Government Consumption Share
- KI = Investment Share
- $OPENK$ = Openness
- POP = Population
- LE = Level of Employment
- BCA = Current Account Balance

Other Abbreviations

- $MCMC$ = Markov-Chain Monte-Carlo.
- $MPLE$ = Maximum Pseudo-Likelihood Estimation.
- MLE = Maximum Likelihood Estimation.
- PPI = Protein–protein interaction

1 Introduction

1.1 Motivation

A *network* (*graph*) is a mathematical representation of relational data in which *nodes* (vertices) correspond to the entities in a system, and *edges* (arcs) correspond to relations among those entities [145]. Networks are widely used for representing complex systems from many different domains, such as economics [10, 17, 26, 30, 52, 93, 170, 186], biology [46, 53, 65, 96, 167, 171, 197], sociology [28, 119, 123, 124, 193] and technology [116, 206]. Network based analysis of these systems sheds light on their organization, the mechanisms that govern their formation and evolution, and the relations among their elements. However, exact solutions of the network analysis problems that produce these insights are intractable as the number of possible network configurations increases exponentially with the size of the networks. In this dissertation, we propose solutions for two of these network analysis problems: (1) *Network Comparison*, and (2) *Network Modelling*.

Many real-world complex systems are dynamic which means that these systems have different configurations at different time points; e.g., world trade networks [34], gene expression networks [61, 103], autonomous networks [206]. The time points at which these systems change and possible causes of these changes can be identified by systematically comparing the topologies of networks that correspond to different snapshots of the system. It is also possible to transfer knowledge among different real-world domains by identifying the topological similarities among corresponding networks.

Identifying the complete list of topological differences between two networks requires solving the subgraph isomorphism problem [35]. Given two graphs G and H as input, the subgraph isomorphism problem asks whether G has a subgraph that is isomorphic to (i.e., has exactly the same topology as) H. This problem is shown to be NP-Complete [63], meaning that there are no polynomial-time exact solutions, but only approximate so-

lutions (i.e., heuristics) for this problem. The network comparison problem is NP-Complete due to the underlying subgraph isomorphism problem. The most sophisticated methods for the network comparison problem are based on graph spectra [190, 201] and small subnetworks including network motifs [140] and graphlets [156, 157]. Among these network properties, graphlets are defined as small, connected, non-isomorphic, and induced subgraphs of a network. We investigate the redundancy and dependency relations among different graphlet properties, and improve the available techniques further by proposing a new network topology descriptor based on this investigation. We use this new topology descriptor to quantify the topological similarities between two networks.

We apply our new methodology to the world trade network. The recent global recession and the unstable nature of the world economy is encouraging researchers to gain a deeper understanding of the functional mechanisms of the world economy. World trade is one of the factors that shape the world economy. Understanding the organizational principles of the world trade network sheds light on the dynamics of the world economy, and guides the economists to minimize the systematic breakdown risks of the world economy. With this aim, we investigate the topological organization of the world trade networks, the link between the changes in world trade network topology and the global recessions, and the effects of a country's position on its wealth.

Given a network $G(V, E)$ that contains $|V|$ nodes, there are $2^{\binom{|V|}{2}}$ possible network configurations that G can be in when the network is undirected; i.e., when each pair of nodes in the network may or may not be connected by an edge without any specific edge orientation. A network model is a set of rules that describes the formation and evolution of networks by picking a subset of the possible configurations [145]. The model-fitting problem asks whether an input network is in the subset of configurations that is picked by the evaluated network model or not. Network comparison methods can easily answer this question by quantifying the topological similarities among the input network and the network configurations defined by the model [76, 156, 163].

We use our new network distance measure to identify the models of five different network types: (1) Autonomous Networks [206], (2) Facebook Networks [193], (3) Metabolic Networks [98], (4) Protein Structure Networks

[143], and (5) World Trade Networks [34]. Furthermore, we propose two new random network models that describe the topology of the world trade networks better than existing models. One of these models is defined solely based on the graphlet properties of the network, and fits world trade networks better than the other models. The superior performance of this model encourages us to extend the applicability of graphlet properties from network comparison problems to network modelling, and to implement a new framework that enables network modelling based on any combination of graphlet properties. Moreover, this new network modelling framework enables defining models that uncover the links among the node attributes and their position in the network.

In the rest of this section, we first explain the different types of real-world networks that are analysed in the scope of this dissertation. Then, we introduce the relevant graph-theoretic concepts on network comparison and modelling. We provide a brief literature survey on the network comparison problem and the state-of-the-art heuristics on it. Following this, we describe well-known random network models, and the methodologies for evaluating their fit on an input network. As our main focus in this dissertation is the analysis of world trade networks, we provide a brief literature survey on the main properties and well-known models of world trade networks. We conclude this section with the dissertation outline.

1.2 Real-World Networks

Relational data from many different real-world domains are modelled and analysed as networks; e.g., financial and world trade networks from the economics domain, protein-protein interaction, genetic interaction, metabolic interaction, protein structure, and signalling networks from biological domain, friendship and collaboration networks from social domain, and autonomous networks from technological domain. These networks appear in many different forms, and represent different types of information about these systems. Mining the networks from these domains uncovers valuable insights into understanding the functional mechanisms embedded in them.

Networks can appear as directed or undirected; based on the existence of an ordering among the node pairs that form the edges. Similarly, the edges can be unweighted or weighted for representing their relative impor-

tance in the network. In this dissertation, we mainly focus on undirected and unweighted networks, since the networks of this form still carry valuable amount of information, while the methods for analysing the structural properties of these networks are much more advanced and scalable to large networks. For this reason, we process the datasets that appear as directed or weighted to obtain unweighted and undirected network representations.

In this section, we introduce different forms of networks from the four above listed real-world domains and explain how we collect and process the network datasets that are analysed in this study.

1.2.1 Economic Networks

Networks are widely used for representing and analysing different types of complex micro-scale and macro-scale economic information; e.g., interbank relation networks where banks are the nodes and the edges represent the credit-debt relations among them [17, 93], investment (inter-company) networks where nodes represent companies and edges link the companies that co-invest on the same portfolio [10, 26], supply-chain networks where nodes correspond to organisations (e.g., companies) and edges represent the flow and movement of materials and information [30, 186], and world trade networks where nodes correspond to countries and edges correspond to the trade links among them [52, 170]. Among this variety of economic network types, we focus on the world trade networks because of their importance in representing the global money flow, and the macro scale information that can be mined through the topology of these networks. World trade networks naturally appear as directed and weighted networks, where the edge directions represent the import/export relations, and the edge weights represent the volume of trade. However, depending on the applied network analysis techniques, unweighted and undirected versions of these networks have also been used.

The *United Nations Commodity Trade Statistics* (UN Comtrade) database is the most reliable and complete source for the world trade data. UN Comtrade contains the world trade relations data since 1962 [34]. The records of the database are formed by the individual declarations of the countries. This method of dataset construction sometimes cause inconsistencies in the database; e.g., country A declares that it imported products of $X\$$ worth

from country B, while B declares that it exported products of $Y\$$ worth to country A. These inconsistencies need to be resolved while constructing the world trade networks from UN Comtrade [170].

The world trade data in UN Comtrade is grouped into categories with respect to their commodities, which enables constructing commodity-specific networks; e.g., trade network of Food and Live Animals, Mineral Fuels, Chemicals, Machinery and Transport Equipment. The trade data is organized with respect to 10 different commodity categorization standards; i.e., STIC (4 different versions), HS (5 different versions), and BEC standards.

Construction of Analysed Economic Networks. From the economic domain, we analyse only the world trade networks in the scope of this study. We obtain the world trade data from the UN Comtrade database [34], and construct commodity specific networks from this dataset using the Standard International Trade Classification (SITC) Revision 1 standard. The products that are traded in 1960s can be very different from what is being traded now; e.g., with the recent developments in the technological era, new products such as laptops, tablet computers, mobile phones appeared after 1990s. SITC Rev. 1 is preferred over the other commodity classification standards since it best covers the range trade products from 1960s to now. SITC Rev. 1 groups the trade products into 10 commodities. For each of these commodities and also for the total trade, we generate 49 trade networks, one for each year between 1962 and 2010 (producing a total of $11 \times 49 = 539$ networks; one network per each commodity - year combination).

In order to resolve the issues caused by the inconsistent import/export declarations to UN Comtrade, we assign confidence scores to each country's import/export declaration. The declaration confidence score of a country, X, is defined as the absolute difference between the sum of all imports/exports that are declared by X and the sum of all imports/exports that are declared by the trade partners of X. The countries with smaller declaration confidence scores are accepted to be more reliable. We determine the weight of a directed edge from country A to country B by taking the trade amount declared by the more reliable country.

The fact that most countries have both import and export trade makes the trade network inherently directional. However, since we are only interested in the presence or absence of an interaction between countries, we generated undirected networks and weighted the edges by summing import and export

trade volumes; e.g., given that country A exports X\$ worth of products to country B and country B exports Y\$ worth of products to A, the trade volume (i.e., the weight of the undirected edge) between A and B is equal to $(X + Y)$. For making the networks unweighted, we filter the lowest weighted edges until 90% of the total trade in the network remains. This filtering produces undirected and unweighted networks that represent the most important trade relations in the network, while covering at least 90% of the money flow in the world. This filtering is necessary for observing the graphlet properties of the world trade network better, since currently graphlets do not support analyses of weighted and directed networks.

1.2.2 Biological Networks

Different types of relational data in biology are analysed using networks. The main types of biological networks are protein - protein interaction networks, metabolic networks, protein structure networks, disease networks, genetic interaction networks, transcriptional regulatory networks, and signal transduction networks. These networks are described as follows:

Protein-Protein Interaction Networks: Proteins are the main building blocks of almost all processes in an organism. They almost never function alone but bind to each other. Protein-Protein Interaction (PPI) networks represent the binding information among all proteins of an organism; nodes representing the proteins and edges representing physical interactions (bindings) between two proteins. Protein interaction networks appear as undirected graphs. Although the edges of these networks are normally unweighted, some studies assign weights representing the confidence on the existence of the interaction [188].

The two main experimental techniques that most protein interaction information is obtained from are Yeast-2-Hybrid (Y2H) screening [94, 95, 165, 172, 183, 195] and Protein Complex Purification methods using Mass-Spectrometry (MS) experiments [33, 64, 108, 162]. Y2H screening experiments identify pairwise protein interactions. However, the interactions identified by this technique contain many false positives since the experiments are performed in yeast nucleus regardless of the organism the genes are taken from. The genes from different organisms may not behave as in their native

environment when they are in yeast nucleus. It is estimated that 50% of the interactions identified by a Y2H experiment are noisy [197], although the Y2H experiment systems have recently improved to produce more accurate results. MS Experiments do not identify binary protein interactions as Y2H experiments do, but they identify protein complexes. In this technique, bait proteins are tagged and used as hooks. The proteins that interact with the bait (i.e., the preys) are separated from the culture together with the bait protein, indicating the existence of a protein complex. The main problem with MS experiments is the extraction of binary interactions from identified complexes. The two models that are commonly used for this purpose are the spoke model and the matrix model. The spoke model assumes that the bait protein interacts with all prey proteins, and none of the prey proteins interact with each other. The matrix model assumes that all protein pairs in the identified complex interact with each other. It is obvious that these two models are abstractions over the underlying structure of the protein complex. The matrix model introduces many false positives while the spoke model introduces many false negatives together with some false positives.

Another problem with protein interaction networks is their incompleteness. For a network with n nodes, there exists $n(n-1)/2$ possible interactions. There are approximately 6,000 proteins in yeast, raising the need for testing $\sim$18 million interactions for their existence. In addition to this huge number of possibilities, most of the protein interaction identification studies are focussed on a certain process or disease, leaving the other parts of the protein interaction network uncovered. *Saccharomyces Cerevisiae* is the most well-studied organism for protein interaction networks. The total number of protein interactions in *Saccharomyces Cerevisiae* is estimated to be between 150,000 - 370,000 [75]. However, the number of protein interactions identified for *Saccharomyces Cerevisiae* as of August 2013 is 81,839 [181] (statistics of BioGRID database - version 3.2.103), showing that even the interactome for this well-studied organism is only $\sim$ 50% complete.

The main public databases that contain protein interaction networks are Saccharomyces Genome Database (SGD) [29], Munich Information Center for Protein Sequences (MIPS) [134], the Database of Interacting Proteins (DIP) [203], the Online Predicted Human Interaction Database (OPHID) [19], Human Protein Reference Database (HPRD) [152, 155], the General Repository for Interaction Datasets (BioGRID) [181, 182], and the Search

Tool for the Retrieval of Interacting Genes / Proteins Database (STRING) [188]. Some of these databases contain interactions that are predicted with computational techniques but not validated experimentally; e.g., OPHID, STRING. These predicted interactions should be used with caution or excluded in most analyses, since protein interaction networks already contain high levels of experimental noise which will exponentially increase with the inclusion of predicted interactions.

Although we also applied our methodology on the PPI networks that are collected from BioGRID, we keep the results of these experiments out of the scope of this dissertation, since our results were similar to the results of previous studies on these networks.

Metabolic networks: Biochemical reactions are crucial for keeping a cell in homoeostasis state (the stable state that a normal cell should be in). Metabolic networks explain the collection of all biochemical reactions that occur in a cell [96, 189]. A metabolic network is a bipartite network of metabolites and reactions, where each metabolite is connected with the reactions that it is involved in. Metabolites can be small molecules such as glucose, amino acids or larger molecules such as polysaccharides, glycan. The biochemical reactions are represented by directional edges since they represent chemical conversion of the metabolites from one form to another. However, most biochemical reactions are bidirectional; i.e., the effects of most reactions can be reversed. For this reason, it is also safe to represent metabolic networks as undirected networks.

The main data source for the metabolic networks is the KEGG database [98]. GeneDB [79], BioCyc [99], EcoCyc [103], MetaCyc [107], and ERGO [149] databases also contain biochemical reaction information for different species.

The metabolic network of all species can be viewed as a very large single network that contains all possible reactions in all species. Enzymes, which are proteins that catalyse the biochemical reactions and synthesized from the genome, cause the difference among the metabolic networks of different species. If the gene that produces the enzyme does not exist in a species, the corresponding biochemical reaction does not occur within the cell of the species. For this reason, it is common practice that reactions are replaced by the enzymes that catalyse them, or even by the genes and

proteins that produce that enzyme in metabolic networks. This replacement generates different metabolic network representations; e.g., networks in the form of metabolite – enzyme, metabolite – protein, and metabolite – gene interactions. The bipartite metabolic networks can be represented as simple graphs, by removing the reaction nodes or metabolite nodes, and connecting the remaining nodes if they are at distance 2 to each other in the bipartite network. This simplification produces metabolic networks in the form of metabolite – metabolite, reaction – reaction, enzyme – enzyme, protein – protein, and gene – gene networks. The particular choice on the network representation to be used depends on the focus of the study and the capabilities of the network analysis tools.

Construction of Analysed Metabolic Networks. We analyse the metabolic networks in the form of enzyme – enzyme interactions. We obtain the metabolic network information of $2,301$ species from KEGG database [98] (downloaded in February 2013), and construct a metabolic network for each species by linking a pair of enzymes if they catalyse reactions that share a common metabolite. We excluded networks containing less than 100 nodes from our analysis.

Protein Structure Networks: The tertiary (3D) structure of a protein provides insights into both characterization of the protein [53, 86, 125, 185, 202] and also identification of its binding domains [68, 135]. The information provided by the tertiary structure of the protein complements the information provided by its sequence. Protein structure networks represent the tertiary structures of proteins. The nodes in these networks correspond to the amino acids in a protein. Two amino acids are connected if they are in contact; i.e., the distance between their alpha-carbons is less than a distance threshold; a common threshold being 7.5 Å (Angstrom, that is 10^{-10} meters). The Structural Classification Of Proteins (SCOP) database is the main information source for the tertiary structure information of proteins [143]. This database contains coordinates that represent the relative positions of the alpha-carbons of each amino acid in a protein. Furthermore, it provides information about the classification of the protein in terms of class, fold, family and superfamily. RCSB Protein Data Bank (PDB) provides an interface for searching the structural information about specific proteins in SCOP [14].

Construction of Analysed Protein Structure Networks. We use the standard distance threshold of 7.5 Å while constructing these networks, and construct the networks of all protein structures in the Astral_40 compendium v1.75B [143] (Downloaded in January 2011). When we filter out the protein structures with more than 40% of sequence identity or less than 100 amino-acids, we obtain the protein structure networks of 8,226 proteins.

Disease Networks: So far, diseases have been grouped and studied in terms of the similarities of their symptoms and the organs they affect. This trend is shifting towards relating diseases based on their genetic origins, rather than their phenotypical similarities. In this respect, Goh et al. [65] defined the first disease – disease association network. In this network, nodes correspond to diseases and two nodes are connected when the two corresponding diseases are linked with at least one gene in common. They further extend the disease – disease network into a bipartite disease – gene network, where the genes and diseases are connected if there is a causal relationship between them. Hidalgo et al. [81] define disease – disease networks in a different manner, by evaluating the common occurence of the diseases in the same person at the same time, which is called commorbidity of diseases. Furthermore, Hu et al. [89] produced a disease – drug network by analysing the genomic expression profiles of human diseases and drugs.

Most disease networks are based on the known disease – gene associations. There are many databases that contain disease – gene associations; e.g., Online Mendelian Inheritance in Man (OMIM) [70], Functional Disease Ontology Annotations (FunDO) [148], Comparative Toxicogenomics Database (CTD) [42], Genetic Association Database (GAD) [12]. The DisGeNet database [11] integrates the disease – gene associations from many of these individual databases, and provides a single dataset containing all experimentally validated and predicted disease – gene associations.

Analyses of disease networks is out of the scope of this dissertation, but we have some ideas on constructing disease – disease association networks as a future direction (explained in Section 6.2.2).

Genetic Interaction Networks: Genetic interactions are defined based on the effect of combined gene deletions on a given phenotype. The multiplicative phenotype fitness model assumes that the combined deletion

of two independent genes is expected to show a phenotype which is the multiplication of the phenotype effects observed after single gene deletions [36, 46, 102, 147, 191, 192]. The most commonly used phenotype for measuring the effects of gene deletions is the colony size; i.e., the number of cells in the culture. If the deletion of two genes results with a phenotype worse than the expected phenotype, then these two genes are accepted to have *negative* genetic interactions. Synthetic lethality, synthetic sickness, synthetic growth defect interactions are examples of negative genetic interactions. Deletion of two genes may also result with a better phenotype than expected, showing a *positive* genetic interaction. Genetic interactions are identified by the synthetic genetic array (SGA) [191] or synthetic lethal analysis by microarray (SLAM) [147] experiments. Dixon *et al.* [46] provides a detailed survey of different genetic interaction types, experimental systems for extracting genetic interaction information, and possible scenarios for the occurrence of the genetic interactions. In genetic interaction networks, nodes represent the genes and edges connect two genes if the observed phenotype after the deletion of genes is unexpected. These network are undirected. Edges can be weighted based on the Z-scores of the observed phenotypes.

The public databases for obtaining genetic interaction data are BioGRID [181] and Flybase [194]. *Saccharomyces Cerevisiae*, *Schizosaccharomyces Pombe*, *Drosophila Melanogaster* and *Caenorhabditis Elegans* are the only well-studied organisms for genetic interactions in the last 10 years. However, a recent study by Lin *et al.* [121] revealed a genetic interaction network for *Homo Sapiens* indicating the forthcoming genetic interaction data from other species.

Genetic interaction networks are not analysed in the scope of this dissertation due to their limited availability.

Transcriptional Regulatory Networks: Transcription regulatory networks describe the relations between genes in terms of their effects on each other's transcription [171]. The nodes of these networks are genes. A directed edge is drawn from node A to node B if the product of gene A (protein A) regulates the transcription of gene B. Protein A binds to the regulatory DNA regions of gene B which may result with over-expression or under-expression of gene B. These interactions are identified by measuring

and comparing the relative mRNA levels of the genes. The well-studied organisms for their transcription regulation mechanisms are *Saccharomyces Cerevisiae* and *Caenorhabditis Elegans*. The databases that contain transcription regulation information are EcoCyc [103], KEGG [98], RegulonDB [61], Reactome [38], TransPath [167] and TransFac [131].

Analysis of transcription regulatory networks is out of the scope of this dissertation.

Signal Transduction Networks: These networks explain the complex signalling mechanisms inside a cell [167]. The nodes of these networks are proteins and the directed edges connecting these proteins represent the signals propagated from one protein to another. These networks are used for modelling the cellular responses to different internal and external stimuli by means of pathways. These networks are especially important for the analysis of diseases, since most diseases are caused by errors occurring in the transduction of the signals in these networks. However, the availability of these relations is limited. Therefore, analysis of signal transduction networks is out of the scope of this dissertation.

1.2.3 Social Networks

Networks have been used for representing a wide-range of complex social systems; e.g., friendship networks [132, 193], collaboration networks [47, 119], citation networks [28, 118], e-mail networks [119], co-authorship networks [119, 123], co-purchasing networks [117, 124]. Among these network types, friendship networks, in which nodes represent people and edges connect people with friendship relations, are of particular interest due to the recent boom in online social networking applications; e.g., Facebook, Twitter, Instagram, Google+. The recent developments in online social networking raised a new set of interesting network analysis questions; e.g., What are the main characteristics of social networks?, How do friendship networks form?, What are the principles governing the evolution of these networks?, How can the social media be used most effectively for viral advertising purposes? Though online social networking applications are important data sources for obtaining friendship networks, collecting these networks is an extremely challenging task. These networks contain millions of nodes and edges, and

the topology of these networks change dynamically by added and deleted users/connections at every second. It is very hard to take a snapshot of these networks at a particular time point. For this reason, these networks are mostly obtained by network crawlers [112], which are small software programs that sample different chunks of the network data in parallel in order to capture the network structure in a fast and accurate way. Network crawling based construction of friendship networks comes with the cost of high levels of noise and incompleteness in the obtained networks. Because of the difficulty of obtaining social networks, there are not many publicly available large-scale datasets.

Construction of Analysed Social Networks: We analyse the friendship networks that are collected Traud et al. [193]. These friendship networks are obtained from the Facebook friendship links of the members of $\sim$ 100 American universities. The nodes of these networks correspond to Facebook user accounts that are linked to an American University as a student or staff, and the links correspond to the Facebook friendship relations among the users. These networks are complete subnetworks of the whole Facebook network in September 2005.

Stanford Large Network Dataset Collection [116] contains some additional social networks of Facebook, Google+, and Twitter. However, these datasets are collected on a voluntary basis by some smartphone applications that the users need to install. For this reason, they are highly incomplete.

1.2.4 Technological Networks

The World Wide Web was developed in 1990 and has been one of the most significant inventions of all times since then. It is indeed one of the best examples of networks; nodes corresponding to electronics such as computers, laptops, mobile phones, satellites with different IP addresses, and edges corresponding to direct physical communication channels among them. Consisting of millions of dynamically changing nodes and edges, it is challenging to obtain a snapshot of this huge system. Autonomous systems provide an abstract representation of the World Wide Web; an autonomous system being a subset of routers in the World Wide Web. In autonomous system networks, nodes correspond to autonomous systems. The autonomous systems that exchange information are connected by edges, forming a "who-

talks-to-whom" network.

Construction of Analysed Technological Networks. The University of Oregon Route Views Project [206] produced one of the best datasets of autonomous networks. Analysing the Border Gate Protocol logs of autonomous systems in Oregon University on a daily basis, 733 networks representing the traffic flow on a single day are constructed [116]. We downloaded these 733 autonomous networks from SNAP database on 09/08/2012. Each of these networks represents daily communication data between autonomous systems of Oregon University for the time period between 8^{th} November 1997 and 26^{th} May 2001.

1.3 Concepts on Networks

A *graph* (also called *network*) is a mathematical representation of a set of objects and the relations among them. A graph is denoted by $G = (V, E)$ where V is the set of *nodes* that represent the objects, and E is the set of *edges* that define the relations among the elements of V. A graph is *undirected* if the edges of the graph have no orientation; i.e., $\forall(u, v) \in E : (u, v) = (v, u)$. A graph is *directed* when the edges of the graph are defined as a set of ordered tuples; i.e., $\forall(u, v) \in E : (u, v) \neq (v, u)$. A graph is *weighted* if a real-valued property is assigned to the edges of the graph. A *simple graph* is an undirected and unweighted graph which contains no self-loops ($\forall v \in V : (v, v) \notin E$) or multiple edges. The *neighbourhood* of node v, $N(v)$, is the set of nodes that are adjacent to v. A *path* between nodes u and v is an ordered set of edges that need to be traced for reaching from node u to node v without visiting any node more than once. A *cycle* is a path that starts and ends at the same node. A graph is *connected* if there exists a path from every node to every other node, otherwise it is disconnected. A graph $H(V', E')$ is a *subgraph* of $G(V, E)$ if $V' \subseteq V$ and $E' \subseteq E$. A subgraph $H(V', E')$ of G is *induced* if it contains all the edges in G between the nodes in V'; otherwise it is a *partial* subgraph. An undirected, connected graph is a *tree* when any two vertices are connected by exactly one simple path; i.e., there are no loops in the graph. Eliminating any edge from a tree makes the graph disconnected, and connecting any two disconnected nodes of a tree forms a cycle. A singly connected network (also known as a *polytree*) is a directed acyclic graph with the property that ignoring the directions on the

edges yields a tree.

Networks can be represented in computer memory using different data structures. The particular choice of a data structure depends on the computational requirements of the software in which the networks are used. The complex information encoded in these data structures do not provide a direct understanding of the network structure. Network properties summarize the main topological characteristics of the network and provide an easy-to-understand description of the network structure. Identifying the exact topological correspondence between two networks is computationally intractable, due to the underlying subgraph isomorphism problem that is NP-Complete [35]. For this reason, there are only approximate solutions (i.e. heuristics) to the network comparison problem. These heuristics use the network properties (statistics) that summarize the network topology. In the rest of this section, we first introduce different data structures for representing networks in computer memory, and discuss their advantages and disadvantages. Then, we describe the topological network properties that summarize the information encoded in these representations. We conclude this section by describing the network comparison heuristics that are based on the network properties.

1.3.1 Network Representations

There are two fundamental data structures for representing a graph $G(V, E)$ with $|V|$ nodes and $|E|$ edges [115]: (1) adjacency list, and (2) adjacency matrix. The *adjacency list* of $G(V, E)$ is a $|V|$ dimensional array AL, where each element of the array $AL[n]$ corresponds to a node n in the network and is linked to the list of nodes that are adjacent to n. For representing weighted networks, an extra list of edge weights should be kept for each node. The *adjacency matrix* of $G(V, E)$ is a $|V| \times |V|$ matrix A, where $A[u, v]$ is a non-zero value when nodes u and v are connected, and equal to 0 otherwise. A is a symmetric matrix when G is undirected. For representing weighted networks, the edge weights can be encoded in the value of $A[u, v]$.

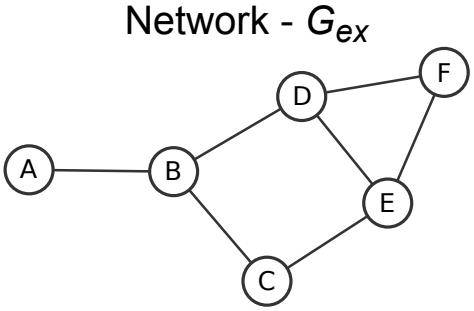
Both representations have their own advantages. Table 1.1 summarizes the worst-case space complexities of representing a network in computer memory with these data structures, together with the worst-case time complexities of common network operations when performed on these repre-

sentations; i.e., adding a node into the network, adding an edge into the network, deleting a node from the network, deleting an edge from the network, and searching for the existence of an edge in a network. Note that, in practice, these complexities are lower, especially when working with sparse graphs. For sparse graphs, the adjacency list representation is more memory efficient than adjacency matrices. Moreover, the computational cost of adding or deleting a node from the network is high for the adjacency matrices, since the size of the matrix changes and the whole matrix needs to be allocated again. On the contrary, edge operations are faster on adjacency matrices, as the existence and the weight of an edge can be directly changed from the relevant matrix element.

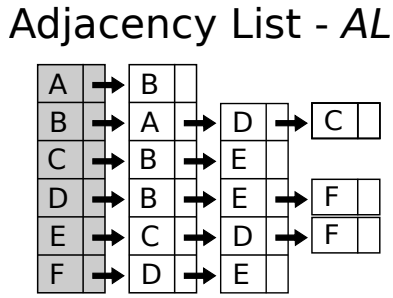
	Adjacency List	**Adjacency Matrix**
Storage	$O(\|V\| + \|E\|)$	$O(\|V\|^2)$
Add Node	$O(1)$	$O(\|V\|^2)$
Add Edge	$O(1)$	$O(1)$
Delete Node	$O(\|E\|)$	$O(\|V\|^2)$
Delete Edge	$O(\|E\|)$	$O(1)$
Search Edge	$O(\|V\|)$	$O(1)$

Table 1.1: Comparison of adjacency list and adjacency matrix representations with respect to the space complexities and time complexities of performing simple graph operations. These complexities are based on the assumption that node indexes are known.

The space allocated for the adjacency matrix can be used more effectively by combining different types of information about the network in this representation. For example, the diagonal elements of the adjacency matrix of a simple graph are all equal to 0 since the graph does not contain self-loops. Therefore, the space allocated for the diagonal elements can be efficiently used for representing other node-specific information. The *Laplacian matrix* of a graph L does this by encoding the degrees (i.e., the number of links that the nodes have) into the diagonal elements of the adjacency matrix. Let D be the diagonal degree matrix of a network; which is a $|V| \times |V|$ matrix with diagonal elements, $D[u, u]$, being equal to the node degrees and all other elements being equal to 0. The standard combinatorial Laplacian matrix,



(A)



(B)

Adjacency Matrix - *A*

	A	B	C	D	E	F
A	0	1	0	0	0	0
B	1	0	1	1	0	0
C	0	1	0	0	1	0
D	0	1	0	0	1	1
E	0	0	1	1	0	1
F	0	0	0	1	1	0

(C)

Laplacian Matrix - *L*

	A	B	C	D	E	F
A	1	-1	0	0	0	0
B	-1	3	-1	-1	0	0
C	0	-1	2	0	-1	0
D	0	-1	0	3	-1	-1
E	0	0	-1	-1	3	-1
F	0	0	0	-1	-1	2

(D)

Figure 1.1: The adjacency list AL, adjacency matrix A, and Laplacian matrix L representations of a small network, G_{ex}. Panel A illustrates the small network, G_{ex}. Panel B, C, and D respectively correspond to the adjacency list, adjacency matrix, and Laplacian matrix representations of the network G_{ex}.

L, of a network is computed from the adjacency matrix A and diagonal degree matrix D as:

$$L = D - A. \tag{1.1}$$

The adjacency list (AL), adjacency matrix (M), and Laplacian matrix (L) representations of a small example network, G_{ex}, is illustrated in Figure 1.1.

1.3.2 Network Properties

The properties that summarize the topological characteristics of a network fall into two categories: (1) *Global Network Properties*, that give an overall view of the network with respect to all nodes and edges (i.e., degree distribution, clustering coefficient, shortest path lengths, centrality measures, and graph spectrum), and (2) *Local Network Properties*, that evaluate the topol-

ogy of a network in terms of its subgraphs (i.e., network motifs, graphlets). Global network properties are useful statistics that provide a simplified description of the network topology. However, these properties sometimes fail to differentiate between networks with completely different topologies. Independent of the amount of information that is embedded in the global network properties, these properties are very sensitive to noise in the network data, as they evaluate the topology of a network as a whole. The local changes in the network (e.g., deletion of a node, removal of an edge) might cause these properties change tremendously, although the structure of the network is still preserved for the rest of the network. Local network properties, which describe the network in terms of its subgraphs, would not suffer from these problems as most of the subgraphs in the network would not be affected from these local changes.

In the rest of this section, we describe the global and local network properties in detail, and illustrate them on the example network, G_{ex}, that is shown in Figure 1.1–A.

Global Network Properties

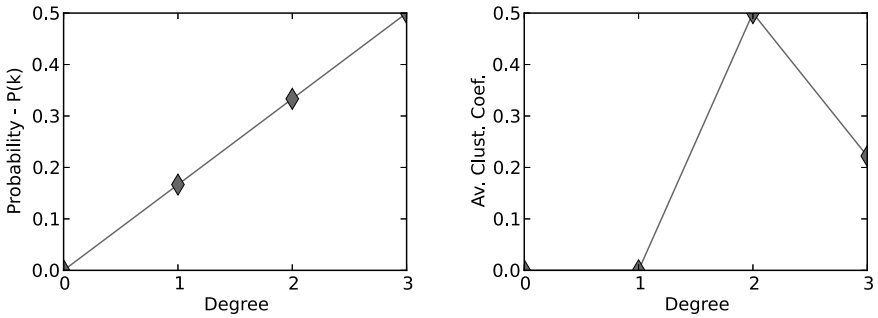
The simplest global network property is the node degree. The *degree* of a node is the number of links that the node has to other nodes in the network. For example, in G_{ex}, the degree of node A is 1 and the degree of node B is 3. *Average degree* of a network is the arithmetic average of the degrees of all nodes in the network. The average degree of G_{ex} is equal to $(1+3+2+3+3+2)/6 = 2.333$. If the network is directed, then two different degree definitions apply: (1) *In-degree* of a node is the number of links which point to the node, and (2) *Out-degree* of a node is the number of links which originate from the node. The *degree distribution* of a node is the distribution of $P(k)$, where $P(k)$ is the probability that a randomly selected node has degree k. Figure 1.2–A illustrates the degree distribution of G_{ex}. The highest degree nodes of a network are called *hubs*.

The *clustering coefficient* of a node v, C_v, is the probability that two neighbours of a node are linked by an edge. When defined, it is computed as:

$$C_v = \frac{T(v)}{\binom{deg(v)}{2}} = \frac{2 \times T(v)}{deg(v) \times (deg(v)-1)}, \tag{1.2}$$

where $deg(v)$ is the degree of node v and $T(v)$ is the number of triangles

through node v. Clustering coefficient is a measure of the degree to which nodes in a graph form transitive relations. For example, in G_{ex}, the clustering coefficient of node B is 0 since its neighbours are not connected, while the clustering coefficient of node D is equal to 0.333 as there is one link between the three neighbours of D. *Average clustering coefficient* is the arithmetic average of the clustering coefficients of all nodes in the network. It represents how densely connected the network is. The average clustering coefficient of G_{ex} is equal to 0.278. The *clustering spectrum*, $C(k)$, is the distribution of average clustering coefficients of all degree k nodes, over all k. Figure 1.2–B illustrates the clustering spectrum of G_{ex}.



(A) **(B)**

Figure 1.2: Global network properties of the network, G_{ex} (Figure 1.1–A). The illustrated network properties are: Panel A – degree distribution, Panel B – clustering spectrum.

A *shortest path* between two nodes is a path that contains the minimum number of edges. The *distance* between two nodes is the length of a shortest path between two nodes; i.e., the number of edges in a shortest path. For example, in G_{ex}, there are two shortest paths between nodes A and E: (1) the path A-B-D-E, and (2) the path A-B-C-E. The lengths of these shortest paths are 3 since these paths contain 3 edges. The distances between the nodes are used for describing how spread the network is. The *diameter* of a network has two definitions: (1) the maximum shortest path distance among all pairs of nodes (e.g., the diameter is 3 for G_{ex}), and (2) the average of shortest path distances of all node pairs (e.g., the diameter is 1.667 for G_{ex}). In this dissertation, we use the first definition of diameter unless otherwise is explicitly stated. The *spectrum of shortest path lengths*

is the distribution of probabilities $P(d)$, where $P(d)$ is the probability that the distance between two randomly selected nodes are separated from each other with distance d, over all d. Figure 1.3–A illustrate the spectrum of shortest path lengths for G_{ex}.

Centrality of a node measures the relative topological importance of a node within a graph. There are five well-known centrality measures: (1) degree centrality, (2) closeness centrality, (3) betweenness centrality, (4) eccentricity centrality, and (5) K-shell decomposition. The simplest centrality definition is the *degree centrality* that is defined as the number of links incident upon a node. The degree centrality assumes that the importance of a node increases together with the number of its neighbours. *Closeness centrality*, $C_c(v)$, is another centrality measure that evaluates the distances from a node to all other nodes. It is computed as:

$$C_c(v) = \frac{1}{\sum_{u \in V} dist(u, v)}, \tag{1.3}$$

where $dist(u, v)$ is the distance between nodes u and v. For example, in G_{ex}, the closeness centralities of nodes A and D are respectively 0.091 and 0.143; higher values representing more central nodes. *Betweenness centrality*, $C_b(v)$, is a more detailed centrality measure that evaluates the number of shortest paths in the network that pass through the node. The betweenness centrality is computed as:

$$C_b(v) = \sum_{s \neq t, s \neq v, v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{1.4}$$

where σ_{st} is the total number of shortest paths between nodes s and t and $\sigma_{st}(v)$ is the number of shortest paths between nodes s and t that pass through v. In G_{ex}, the betweenness centrality of node A is 0 since none of the shortest paths in the network pass through A. On the contrary, the betweenness of node D is 0.3 highlighting its central role in connecting nodes. The *eccentricity* of a node is the maximum of the shortest path distances between the node and all other nodes in the network. *Eccentricity centrality*, $C_e(v)$, is computed as:

$$C_e(v) = \frac{1}{E(v)}, \tag{1.5}$$

where $E(v)$ represents the eccentricity of node v. In G_{ex}, the eccentricities of nodes A and D are respectively 3 and 2, and the corresponding eccentricity centralities are 0.333 and 0.5. *K-Shell decomposition* is another centrality measure which divides the nodes of a network into groups based on their degrees [27]. The K-shell decomposition of a network is computed iteratively, by first removing all nodes with 1 connection (i.e., degree 1) until no such nodes are left. All the removed nodes form the 1-shell of the network. Then, the same process is repeated for nodes with two or less connections forming the 2-shell. The decomposition process is iterated until all nodes are assigned to one of the k-shells. Nodes which are assigned to higher degree shells are more central in the network. In G_{ex}, the 1-shell is $\{A\}$. Once node A is removed from the network, the 2-shell of the network is defined by iteratively removing degree 2 or less nodes. In this respect, first, nodes B, C, and F are removed from the network. As a result, nodes D and E are both degree 1 in the remaining network. For this reason, nodes D and E are also included into the 2-shell of the network. Therefore, the 2-shell of the network contains nodes $\{B, C, D, E, F\}$, and there are no higher degree shells of this network.

Spectral network theory encodes the complexity of a network's topology using the eigenvalues and eigenvectors of matrices associated to the network; e.g., adjacency matrix, Laplacian matrix, heat kernel, path length distribution [201]. Let X be the matrix associated with the graph. The eigendecomposition of X is:

$$X = \phi\lambda\phi^T, \tag{1.6}$$

where $\lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ is the diagonal matrix with the sorted eigenvalues as elements and $\phi = (\phi_1|\phi_2|...|\phi_n)$ is the matrix with the sorted eigenvectors as columns. The *graph spectrum* is defined as the set of eigenvalues $s = \{\lambda_1, \lambda_2, ..., \lambda_n\}$, where $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$. The eigenvalues of a matrix are real numbers when the matrix is symmetric; i.e., $A = A^T$. This property indicates that the spectra of undirected networks are real numbers. Two networks are *cospectral* if they have the same eigenvalues with respect to the used matrix representation. Note that, more than one graph may share the same spectrum, especially when the graph is in a tree form. Figure 1.3–B illustrates the adjacency matrix and Laplacian matrix

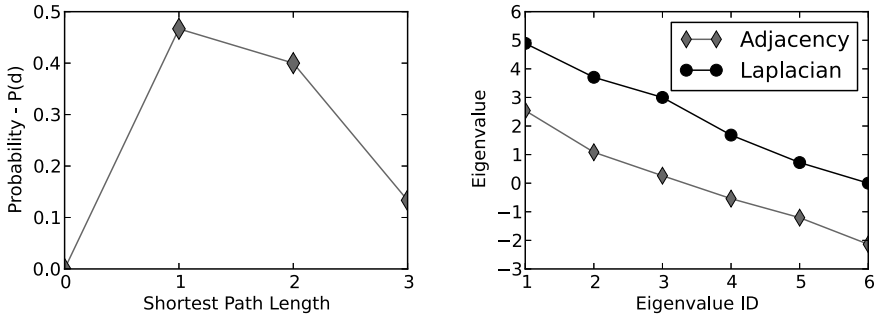


Figure 1.3: Global network properties of the network, G_{ex} (Figure 1.1–A). The illustrated network properties are: Panel A – spectrum of shortest path lengths, Panel B – adjacency and Laplacian spectra.

spectrums of G_{ex}.

Local Network Properties

Network Motifs (or simply called *motifs*) are small partial subgraphs of a network that occur more frequently than expected in random [139, 140]. The null model for network motif identification is the Erdös - Renyi (ER) random network model, in which every pair of nodes are randomly connected with probability p (detailed description is provided in Section 1.5). The significance of the over-representation or under-representation of a network motif is evaluated by its Z-score, Z_i:

$$Z_i = \frac{N_{real_i} - < N_{rand_i} >}{std(N_{rand_i})}, \tag{1.7}$$

where N_{real_i} is the number of appearances of subgraph i in the real network, and $< N_{rand_i} >$, $std(N_{rand_i})$ are the mean and standard deviation of the number of appearances of subgraph i in same size and density ER networks. Z-scores of subgraph patterns in larger networks tend to be higher; therefore, they need to be normalized depending on the network size. The normalized

Z-score of a subgraph i is called the significance profile of i, SP_i:

$$SP_i = \frac{Z_i}{\sqrt{\sum_j Z_j^2}}. \tag{1.8}$$

Network motifs uncover the main organizational principles of networks. For example, the feed-forward loops are found to be overrepresented in signalling networks [3] explaining the way signals are propagated in such a network.

Artzy-Randrup *et al.* [7] criticize the dependence of network motifs on ER network models. They claim that most real-world networks do not have random topology, and comparing the frequency of the subgraphs of input network with the frequencies in the ER networks contains some bias as the random network model is not a good model for the real network. On the other hand, network motifs are partial subgraphs. For this reason, their ability to capture the structural similarities is not as strong as that captured by the induced subgraphs.

Przulj *et al.* [157] introduce *graphlets*; that are small, induced, connected, and non-isomorphic subgraphs of a large network. They also annotate the nodes of all 2- to 5-node graphlets with automorphism *orbits* (simply called orbits), where each automorphism orbit defines a group of nodes that are topologically symmetrical in a graphlet [156]. Thirty 2- to 5-node graphlets and their 73 automorphism orbits are illustrated in Figure 1.4. Using the automorphism orbits of graphlets, Przulj et al. [156] generalize the notion of node degree to graphlet degree: the i^{th} graphlet degree of a node N is the number of graphlets that N touches at orbit i. With this definition, the 0^{th} graphlet degree corresponds to the standard definition of node degree. The *Graphlet Degree Vector (GDV)* (also known as graphlet signature) of a node is a 73-dimensional vector where each value represents the graphlet degree of the node for a particular orbit. The GDV computation for node A in G_{ex} is illustrated in Figure 1.5. The GDV of a node represents the topological structure around a node [138]. Graphlet statistics can be used in two different ways for describing the topology of a network: (1) the number of appearances of the thirty graphlets in the network, and (2) the distributions of the graphlet degrees for each of the 73 orbits. For example, G_{ex} contains 1 of each graphlet in $\{G_2, G_4, G_5, G_9, G_{10}, G_{13}, G_{16}, G_{21}\}$, 2 of graphlet G_6, 5 of graphlet G_5, and 8 of graphlet G_1. The graphlet degree

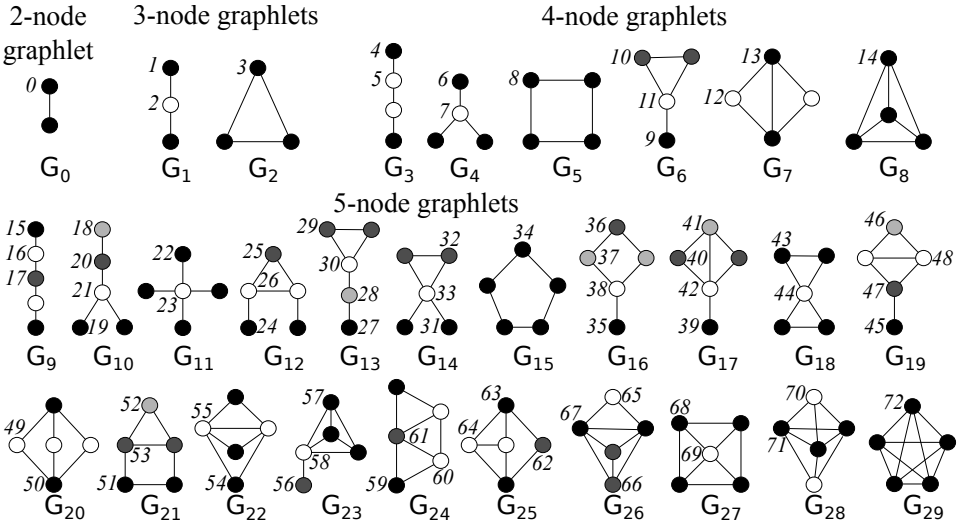
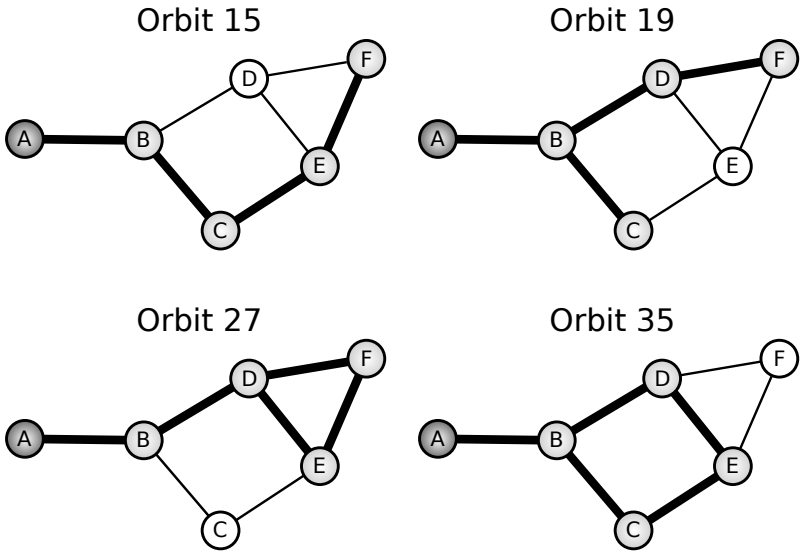


Figure 1.4: All 2- to 5-node graphlets $G_0, G_1, G_2, \ldots, G_{29}$, and their automorphism orbits $0, 1, \ldots, 72$. Nodes belonging to the same automorphism orbit are of the same shade in each graphlet.

distributions of 2- and 3-node graphlet orbits (i.e., orbits 0, 1, 2, and 3) are illustrated in Figure 1.6.

In comparison to network motifs, graphlets are more powerful in capturing the underlying topology because they are defined as induced subgraphs of a network. Furthermore, they are not defined in comparison to a random network model but only on the observed counts of subgraphs, without any assumptions on a null network model. The statistics of 2- to 5-node graphlets are detailed enough to capture the topological similarities between networks, as most real-world networks are small-world networks, and 5-node graphlets capture most of their topological properties. However, the necessity of using 5-node graphlets is an open question that needs to be investigated. Furthermore, the statistics obtained from different graphlets are not completely independent of each other. There are redundancies and dependencies among graphlet statistics, i.e., the statistics of some graphlets can be inferred from a different set of graphlet statistics. Current graphlet-based methods [156, 157] do not handle these issues accurately, and they need to be improved further. Finally, graphlets are defined only for undirected networks, while network motifs also include directed subgraph statistics. Development of directed graphlet statistics is still an open research topic



0	1	2-3	4	5	6	7-14	15	16-18	19	20-26	27	28-34	35	36-72
1	2	0	3	0	1	0	1	0	1	0	1	0	1	0

Figure 1.5: Graphlet degree vector of node A in G_{ex} (Figure 1.1–A) and its computation for 5-node graphlets. The number of 5-node graphlets associated with node A is 4. Notice that, the path *A-B-D-F-E* does not increase the graphlet degree of orbit 27, since graphlets are induced subgraphs and the induced subgraph on these nodes also contains the edge (D, E).

that is not in the scope of this dissertation.

1.4 Introduction to Network Comparison

The network comparison problem consists of three sub-problems: (1) network topology comparison, (2) network alignment, and (3) network querying. The network topology comparison problem focus on defining distance measures that evaluate the overall topological correspondence between two networks. The network alignment problem requires a more detailed comparison that would produce a mapping between the nodes of two networks such that the correspondence between the edges of the two networks is maximized. Finally, the network querying problem searches for a small topolog-

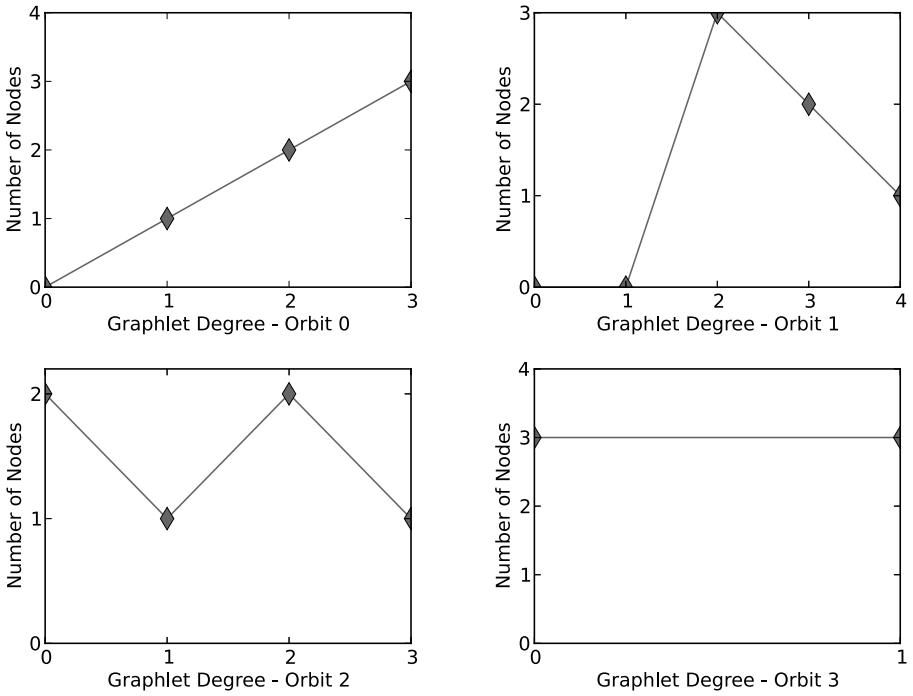


Figure 1.6: The graphlet degree distributions for orbits 0, 1, 2, and 3 for G_{ex} (Figure 1.1–A). Note that, the graphlet degree distribution of orbit 0 corresponds to the degree distribution of the network (Figure 1.2–A). The topology of a network is described by the 73 graphlet degree distributions, one for each orbit in Figure 1.4.

ical pattern in a large graph. There are no polynomial-time exact solutions for any of these problems because of the underlying subgraph isomorphism problem that is NP-Complete [35]. For this reason, heuristic approaches that produce approximate solutions in polynomial-time are proposed for these problems. In this dissertation, we focus on the network topology comparison problem because of its applicability on network modelling.

The simplest heuristics for the topological network comparison problem compare the global network properties of the two networks that are described in Section 1.3.2. The single-valued global network properties (i.e., average degree, average clustering coefficient, diameter) can be directly compared by taking their absolute difference; i.e., given two global network properties p_1 and p_2, their absolute difference is $|p_1 - p_2|$. When the global network properties are in the form of distributions (e.g., degree distribu-

tion), the most direct approach is to compute the Euclidean distance of the two distributions; e.g., given the two degree distributions d_i and d_j, the Euclidean distance, $Dist(d_i, d_j)$, is computed as:

$$Dist(d_i, d_j) = \sqrt{\sum_{k=0}^{max(d_i,d_j)} (d_i(k) - d_j(k))^2}. \quad (1.9)$$

The distributions may be re-weighted or normalized before the computation of the Euclidean distance, in order to highlight a specific part of the distribution. As an alternative, standard statistical tests that compares two distributions such as Kolmogorov-Smirnov [175] or Mann-Whitney-U [126] test can be used for evaluating the similarities between the two distributions, with the cost of increased computational time.

Given the spectrums of two graphs s^1 and s^2 (see Section 1.3.2 for the definition of graph spectrum), the *spectral distance* between the two graphs, $d_s(G, H)$, is defined as the Euclidean distance between their spectrums [201]:

$$d_s(G, H) = \sqrt{\sum_i (s_i^1 - s_i^2)}. \quad (1.10)$$

When the lengths of the spectrums for two graphs are different, 0 valued eigenvalues are added into the smaller spectrum while preserving the correct magnitude ordering. Note that, the graph spectrum can be computed using adjacency matrix, Laplacian matrix, normalized Laplacian matrix, heat kernel, or the shortest path length matrix. Wilson et al. [201] provide a detailed evaluation of these alternative graph spectrum definitions, and show that the spectral distance computed from the Laplacian matrices of two networks is the best measure for classification and clustering purposes. Later on, Thorne et al. [190] used the spectral distance of Laplacian matrices for analysing the evolution of protein interaction networks. In parallel to these studies, we chose spectral distance from Laplacian matrices as the benchmark representing the performance of spectral distance measures in this dissertation.

As explained in Section 1.3.2, there are two different graphlet statistics that describe the topology of a network: (1) the number appearances of 30 graphlets in the network, and (2) the 73 graphlet degree distributions, each

corresponding to a graphlet orbit. *RGF Distance* between two networks, $RGF(G, H)$, uses the first of these properties for comparing two networks [157]:

$$T(G) = \sum_{i=1}^{29} N_i(G), \tag{1.11}$$

$$F_i(G) = -log\left(\frac{N_i(G)}{T(G)}\right), \tag{1.12}$$

$$RGF(G, H) = \sum_{i=0}^{29} |F_i(G) - F_i(H)|, \tag{1.13}$$

where $N_i(G)$ is the number of times that the graphlet G_i appears in graph G, $T(G)$ is the total number of 3- to 5-node graphlets that appear in the network (edge count is excluded in the computation), and $F_i(G)$ is the relative graphlet frequency for graphlet i.

The second graphlet based network statistic, the graphlet degree distribution, is used for defining a more detailed network distance measure, called *Graphlet Degree Distribution Agreement* (also known as GDD-Agreement or GDDA) [156]. Unlike RGF distance, GDD-Agreement is a similarity measure, quantifying how topologically similar two networks are. GDD-Agreement between two networks, $A^j(G, H)$, is computed for an orbit j as:

$$S_G^j(k) = \frac{d_G^j(k)}{k}, \tag{1.14}$$

$$T_G^j = \sum_{k=1}^{\infty} S_G^j(k), \tag{1.15}$$

$$N_G^j(k) = \frac{S_G^j(k)}{T_G^i}, \tag{1.16}$$

$$D^j(G, H) = \frac{1}{\sqrt{2}}\left(\sum_{k=1}^{\infty}[N_G^j(k) - N_H^j(k)]^2\right)^{\frac{1}{2}}, \tag{1.17}$$

$$A^j(G, H) = 1 - D^j(G, H), \tag{1.18}$$

where the number of orbits touching the j^{th} orbit k times, $d_G^j(k)$, is first scaled, $S_G^j(k)$, and then normalized, $N_G^j(k)$, in order to decrease the effect of larger degrees in GDD-Agreement. Euclidean distance between the scaled and normalized distributions, $D^j(G, H)$, is used for identifying the distances

between the networks based on orbit j. The computed distance is divided to $\sqrt{2}$ in order to produce a distance value between 0 and 1. The distance value is converted to a similarity (agreement) score by subtracting it from 1. The overall similarity between the two networks, G and H, are computed from the 73 different GDD-Agreement scores by either taking the arithmetic mean:

$$GDDA_{arith}(G,H) = \frac{1}{73}\sum_{j=0}^{72} A^j(G,H), \tag{1.19}$$

or the geometric mean (Equation 1.20):

$$GDDA_{geo}(G,H) = \left(\prod_{j=0}^{72} A^j(G,H)\right)^{\frac{1}{73}}. \tag{1.20}$$

The performance evaluation of a network distance measure depends on the application that they are intended to be used for. For model identification and clustering purposes, these network distance measures have not been systematically compared against each other. In this dissertation, we perform this systematic evaluation using networks generated from different network models, and test which of the distance measures best group networks from the same model.

1.5 Introduction to Network Modelling

A network model is a collection of rules for generating random networks with specific topological properties. A well-fitting network model gives insights into understanding the functional mechanisms in the real-world system and enables performing more effective data mining in the network. Network models have been used with different purposes; e.g., for identifying the over-represented subgraphs (network motifs) in the network [139, 140, 171], denoising biological networks by predicting the confidence levels of interactions in the network [111], guiding interactome detection experiments [113].

In this section, we first describe the standard random network models that are widely studied for the modelling of biological networks; namely Erdös - Rényi Model, Generalized Random Model, Scale-free Barabàsi-Albert (Preferential Attachment) Model, Scale-free Gene Duplication and Divergence Model, Geometric Model, Geometric Model with Gene Duplication and Di-

vergence, Stickiness Index-Based Model. Then, we describe a more flexible set of network models called Exponential-family Random Graph Models (ERGMs). Finally, we provide details on methods for evaluating how well a network model fits a network.

1.5.1 Random Network Models

The first and simplest network model is the **Erdös - Rényi Model (ER)** [51]. In ER model, an edge between any pair of nodes is drawn uniformly at random with probability p. For generating an ER network with n nodes and probability p, each pair of nodes are connected randomly with probability p, resulting with $p \times (n(n-1)/2)$ edges in the network. Many topological properties of ER networks can be theoretically computed [16]. The degree distribution of an ER network follows a Poisson distribution. The average degree of an ER network is $(n-1) \times p$. The average clustering coefficient of ER networks are small since the edges in the network are distributed uniformly at random. The average diameter of these networks are also small which is an order of $\log(n)$.

A variation of ER model, called **Generalized Random Model (ER-DD)**, matches the degree distribution of the generated network to a given distribution using the "stubs" method [146]. The number of "stubs" to be filled by edges are assigned to each node randomly based on the given degree distribution. Edges are added by randomly picking node pairs that have available "stubs" and connecting them. After each edge addition, the number of available stubs of the connected nodes are decreased by one. Therefore, the degree distribution of these models match with the given distribution when all "stubs" are filled. Similar to ER models, the clustering coefficient of ER-DD models are low because of the random distribution of the edges in the network.

Scale-free networks are characterized by their power-law degree distributions, meaning that a small number of nodes have high degrees while most of the nodes have low degrees. **Barabàsi-Albert Preferential Attachment Model (SF-BA)** is the most well-known among scale-free network models [9]. This model uses the rich-gets-richer principle for generating scale-free networks: Starting with a small seed network (e.g., a network containing a single node), new nodes are added into the network by connecting them

with existing nodes with probabilities proportional to their degrees:

$$p(v_i) = \frac{d_i}{\sum_{j=0}^{n} d_j}, \tag{1.21}$$

where d_i is the degree of node i. Hormozdiari et al. [88] shows that the seed network configuration strongly influences the resulting network. The clustering coefficient and average diameter of SF-BA networks are low. SF-BA networks are very robust to noise, as deletion and addition of most nodes do not affect the connectivity of the network. However, high degree nodes (hubs) are open for targeted attacks which results with the overall failure of the network.

Another scale-free model is the **Scale-free Gene Duplication and Divergence Model (SF-GD)** [196]. SF-GD is a biologically motivated model that imitates the gene duplication and mutation events for the scale-free network generation. SF-GD model generation consists of two main steps: In the duplication step, a node in the network is selected uniformly at random, and a new node that has the same set of connections with the selected node is added into network. The selected node and the new node are also connected with probability p. In the divergence step (also known as mutation step), each edge of the new node is deleted with probability q. This procedure is repeated until the generated network contains the same number of nodes with the input network.

In a **Geometric Model (GEO)**, the nodes are independently and uniformly distributed in a unit space [151]. Two nodes are connected if the distance between them is smaller than or equal to a distance threshold, r. The distance threshold is chosen to adjust the number of edges in the model networks. GEO model can be altered based on the dimensionality of the metric space and the distance measure among the nodes. The degree distribution of GEO networks follows a Poisson distribution, unlike scale-free networks. Their clustering coefficients are high and their diameters are small.

Pržulj et al. [159] adapt geometric models to imitate the gene duplication and mutation events that occur during the evolution of a biological network, defining **Geometric Model with Gene Duplication and Divergence (GEO-GD)**. The GEO-GD model is based on the fact that the nodes of a biological network (i.e., proteins) share the same bio-chemical space. When

a gene is duplicated, it is in the same location with its ancestor. As time progresses, the node diverges from the ancestor node by moving in the biochemical space and forming new connections with the other nodes in the network. Inspired by this principle, GEO-GD model generation is initiated with a small number of nodes that are distributed randomly in a metric space. New nodes are added into the metric space by duplicating existing nodes and moving them randomly in the metric space. As a duplicated node moves further away from its ancestor, it differs from the originating node by forming more diverse connections. GEO-GD models are characterized by power-law degree distributions, high clustering coefficients, and low average diameters. Two alternative methods are suggested for generating GEO-GD models: (1) GEO-GD Expansion (GDE) model, and (2) GEO-GD with probability cut-off (GDP) model. In the GDE model, when a node is duplicated, the new node moves in a random direction for a random distance; the maximum distance being $2r$ where r is the distance threshold that the two nodes are connected in the geometric model. If the node moves less than r, then it shares most of its ancestor's functions and neighbours. In the GDP model, there are two possibilities that a duplicated node can move: (1) it can move in a random direction for a maximum distance of r with probability p, or (2) it can move in a random direction for a maximum distance of $10r$ with a probability of $1-p$. In this dissertation, we consider only the GDE model for generating GEO-GD networks.

Another biologically motivated network model is the **Stickiness Index-Based Network Model (STICKY)** [158]. The STICKY model is based on two main assumptions: (1) High degree proteins have many binding domains and these domains are highly involved in interactions, and (2) A pair of proteins are more likely to interact if they both have high degrees (many domains). The model uses stickiness indices of nodes for defining the probability of two nodes being connected. The stickiness index of node i is defined as:

$$\theta_i = \frac{deg(i)}{\sqrt{\sum_{j \in V(G)} deg(j)}}, \tag{1.22}$$

where $V(G)$ is the set of all nodes in network G and $deg(i)$ is the degree of node i. The edges of the model network are randomly chosen based on the probabilities defined by the multiplication of the stickiness indexes of the corresponding nodes. The STICKY model generates networks that have the

same degree distribution as the input network.

In Figure 1.7, we illustrate networks that are generated from the seven network models. We use a SF-BA network that has 500 nodes and 1% edge density as the seed network, and generate one network from each model that share the model-specific characteristics of the seed network. As illustrated in Figure 1.7–A, the ER network has a uniform distribution of edges among all node pairs. The ER-DD network follows the same trend, but this time, the network has more visible hubs since the ER-DD model imitates the degree distribution of the seed network (Figure 1.7–B). The SF-BA model has a topology similar to the ER-DD model, few nodes being connected to all other nodes, and the rest of the nodes distributed as peripheries around them (Figure 1.7–C). Due to the imitated duplication and mutation events, the SF-GD model produces networks that have a few strongly clustered components (Figure 1.7–D). The network from the GEO model highlights the position-specific clustering of the nodes in the unit space (Figure 1.7–E). The position-specific clustering pattern is also observable in the GEO-GD network (Figure 1.7–F), together with the highly clustered connected components pattern caused by the duplication and mutation events. Finally, the network from the STICKY model shows a strong core-periphery structure, with a tightly connected core and many peripheral nodes (Figure 1.7–G).

1.5.2 Exponential-family Random Graph Models

Exponential-family random graph models (ERGMs, also known as p* models) are probabilistic network models that are parametrized in terms of sufficient statistics based on graph-theoretic properties [85, 150, 164]. In ERGMs, the conditional probability of an edge's existence is determined by the effect of the edge on the values of one or more network properties (i.e., sufficient statistics or functions), given the rest of the graph in which it resides. The network properties that define a model are conventionally called model *term*s.

ERGMs are specified via three elements: (1) a vector of model terms (i.e., sufficient statistics), (2) a vector of real-valued model coefficients, and (3) a support [92, 105]. Let $\mathbf{Y}$ be a random variable that represents an n-by-n adjacency matrix of an unweighted, loopless (no self-edges) network with n nodes. $\mathbf{Y}$ can have 2^{n^2-n} different values (configurations), where

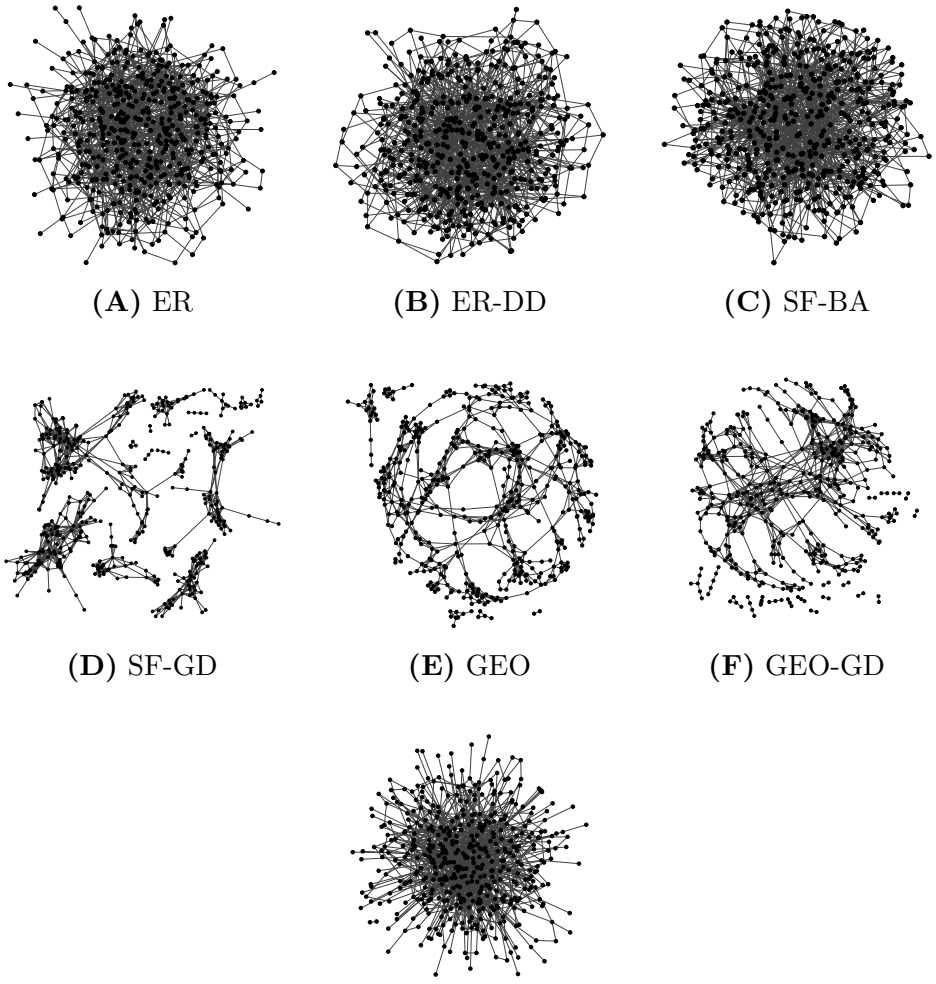


Figure 1.7: Illustration of networks that have 500 nodes and 1% edge density and generated from the seven network models. The corresponding models are: Panel A – ER model, Panel B – ER-DD model, Panel C – SF-BA model, Panel D – SF-GD model, Panel E – GEO model, Panel F – GEO-GD model, Panel G – STICKY model.

each value represents a different network having n nodes. The number of configurations is 2^{n^2-n} because the adjacency matrix of the graph contains binary values (unweighted graph) and the diagonal values of the matrix are all equal to 0 (no self-edges). The set of all possible configurations is called the *support* for $\mathbf{Y}$ and represented by $\mathcal{Y}$. Any element of $\mathcal{Y}$ is a *realization* of $\mathbf{Y}$ and is represented by $\mathbf{y}$. An ERGM describes the probability of observing a realization, $\mathbf{y}$, conditional on several network properties (sufficient statistics). The probability of observing a realization is computed as:

$$P_{\theta,\mathcal{Y}}(\mathbf{Y} = \mathbf{y}|\theta, t) = \frac{\exp\{\theta^\top t(\mathbf{y})\}}{\sum_{z\in\mathcal{Y}} \exp\{\theta^\top t(z)\}}, \mathbf{y} \in \mathcal{Y}, \tag{1.23}$$

where θ is the vector of model coefficients (i.e., the weights for the model terms) and t is the vector of sufficient statistics for the model terms (i.e., the values of the considered network properties for all possible realizations) [55, 199]. Generalization of the above to more general cases (e.g., graphs with loops, digraphs, etc.) is immediate given alternative choice of $\mathcal{Y}$. Since any probability mass function for $\mathbf{Y}$ on finite $\mathcal{Y}$ can be written in this form, ERGMs are fully general representations for random graphs of finite order.

The denominator of Equation 1.23 is a normalizing factor. The computation of the normalizing factor in the general case requires computation of the exponent term for all possible realizations of $\mathbf{Y}$, which typically has computational complexity of order 2^{n^2}. For this reason, computation of the normalizing factor is intractable. However, for many purposes one can work with ratios of graph probabilities, i.e.,

$$\frac{P_{\theta,\mathcal{Y}}(\mathbf{Y} = \mathbf{y}'|\theta, t)}{P_{\theta,\mathcal{Y}}(\mathbf{Y} = \mathbf{y}|\theta, t)}, \tag{1.24}$$

rather than with the probabilities themselves. In this case, the normalizing factor cancels and we are left with an expression in terms of the differences in model statistics under the respective graphs. The vector $t(\mathbf{y}') - t(\mathbf{y})$ is known as the vector of change statistics for $\mathbf{y}'$ versus $\mathbf{y}$ under t, and plays a critical role in ERGM computation. Of particular importance are the change statistics resulting from the perturbation of $\mathcal{Y}$ by a single edge state (i.e., adding or removing a specific edge). The change statistics under such a perturbation may be derived as follows. Let $\mathbf{y}$ be a realization of $\mathbf{Y}$. $\mathbf{y}^+_{\mathbf{ij}}$ represents the configuration that contains all the edges of $\mathbf{y}$ and the

edge between nodes i and j. Similarly, $\mathbf{y}_{\mathbf{ij}}^{-}$ represents the configuration that contains all the edges of $\mathbf{y}$ excluding the edge between nodes i and j. Then, the change statistics of $\mathbf{y}$ for nodes i and j under perturbation of the edge (i, j), $\delta_t(\mathbf{y})_{i,j}$, is defined as:

$$\delta_t(\mathbf{y})_{i,j} = t(\mathbf{y}_{ij}^{+}) - t(\mathbf{y}_{ij}^{-}). \tag{1.25}$$

The normalizing factor in Equation 1.23 can be eliminated by dividing $P_{\theta,\mathcal{Y}}(\mathbf{Y} = \mathbf{y}_{\mathbf{ij}}^{+}|\theta, t)$ by $P_{\theta,\mathcal{Y}}(\mathbf{Y} = \mathbf{y}_{\mathbf{ij}}^{-}|\theta, t)$. The derivation from this division produces the conditional odds for the existence of edge (i, j):

$$\frac{P_{\theta,\mathcal{Y}}(\mathbf{Y} = \mathbf{y}_{\mathbf{ij}}^{+}|\theta, t)}{P_{\theta,\mathcal{Y}}(\mathbf{Y} = \mathbf{y}_{\mathbf{ij}}^{-}|\theta, t)} = \exp\{\theta^T \delta_t(\mathbf{y})_{i,j})\}. \tag{1.26}$$

The conditional odds given in Equation 1.26 can be used for deriving the probability of the existence of an edge given the remainder of the graph. The conditional probability for the existence of an edge (i, j) is then computed as:

$$P_{\theta,\mathcal{Y}}(\mathbf{Y}_{ij} = 1|\mathbf{Y}_{ij}^{c} = \mathbf{y}_{ij}^{c}, \theta, \delta_t) = \text{logit}^{-1}(\theta^T \delta_t(\mathbf{y})_{i,j}), \tag{1.27}$$

where $\text{logit}(p) = \log(p/(1-p))$, and $\mathbf{y}_{ij}^{c}$ is the realization that contains all the edges of $\mathbf{y}$ except the edge $(i,\ j)$ (See [85] for details).

In an inferential context, ERGM models of a network are typically fit by estimating the model coefficients, θ, that maximize the conditional probability, $P_{\theta,\mathcal{Y}}(G|\theta, t)$. The most common methods for the estimation of model coefficients are Maximum Pseudo-Likelihood Estimation (MPLE) or Maximum Likelihood Estimation (MLE). Current MLE methods typically rely on Markov-Chain Monte-Carlo (MCMC) algorithms that simulate ERGM draws without computing normalizing factors. Although implementations differ, a typical MCMC algorithm for ERGM simulation randomly perturbs the edge states in the simulated network one-by-one and uses the change statistics of these edge flips to compute the change in acceptance probabilities of the realizations using Equations 1.26 or 1.27. In both estimation strategies, change statistics are employed for avoiding explicit normalization factor computation. Indeed, the model statistics themselves never need to be directly computed; only the change scores are necessary for most purposes. Computing the change score, δ, rather than the actual property

value, t, yields substantial savings for commonly used model terms (e.g., degree statistics, k-stars, triad counts).

The *ergm* package [92] for R statistical computing system [161] provides a set of tools for analysing networks within an ERGM framework. The *ergm* package allows the users to define ERGMs based on a wide range of network properties, estimate model coefficients of the ERGMs with respect to input networks using the likelihood-based methods, simulate (generate) random networks from a given ERGM, and perform graphical goodness-of-fit tests of the type described by [72, 90]. The *ergm* package provides a large but limited number of model terms. The complete list of these natively supported model terms are listed and explained in [66].

New user-defined model terms can be included into the *ergm* package using the *ergm.userterms* package [73]. A new modelling term is defined by implementing an R function and a corresponding C function. The R function acts as an interface for the model term and pre-processes the term parameters before the computation of the change statistics. The C function performs the computation of the change statistics for the model term when an edge is flipped in the network; e.g., for defining "the number of edges" term, the C function should return +1 when a new edge is added into the network and -1 when an edge is removed. The code for calculating the change statistics should be time-optimized, as it is likely that this computation will be performed millions of times during a typical MCMC run. Due to *ergm*'s modular design, model terms that are employed to *ergm* package this way work in precisely same manner as natively supported terms, and are transparent from an end-user perspective.

1.5.3 Evaluating Model Fit on Real-World Networks

Well-fitting network models give insights into understanding the rules governing the emergence and evolution of real-world networks. In order to assess the fit of a network model to a given network, the network should be compared with the networks that can be generated from the model. In particular, given an input network G, the first step of assessing the model fit is generating several networks from the evaluated network model. Each network model is capable of producing a different range of networks; e.g., less parametric models such as ER are theoretically capable of producing any

observable network over n nodes, while more stringent models that require more parameters can only generate a small range of networks. The range of observable networks also changes depending on the size and density of the generated networks. For this reason, the number of model networks that need to be generated from a model for model-fitting experiments should be chosen to allow observation of a significant range of different configurations. On the other hand, generating more model networks increases the required computational time for the model-fitting tests. Generating a minimum of 30 networks per model was previously accepted to be sufficient for observing a significant range of networks that can be generated from a model [82, 135, 159]. After generating a sufficient number of model networks, the topologies of the generated networks are compared with the input network, G [157]. As explained in Section 1.4, topological network comparison is a NP-Complete problem, for which there are only approximate polynomial-time solutions [35]. Therefore, the comparison between the topologies of the input network and the model generated networks are performed using the heuristic approaches. Any of the global or local network properties that are explained in Section 1.4 can be used to perform these comparisons; e.g., degree distribution, clustering coefficient, shortest path length distribution, graph spectra, network motifs, and graphlets.

The most intuitive method for comparing the topologies of the input network and model networks is contrasting their global network properties (e.g., degree distribution, spectrum of shortest path lengths). A visual model-fitting assessment can be obtained by computing the averages and standard deviations of the global network properties for all generated model networks, and plotting them together with the properties of the input network. This method was previously applied for evaluating the fit of ERGM models in Statnet package [66]. However, global network properties are not detailed enough to capture the exact topologies of networks. For example, a graph that is composed of 3 disconnected triangles and a 9-node cycle have the same degree distributions while their topologies are completely different. For this reason, testing the model based on global network properties is not a strong model-fit assessment method. Furthermore, the results of these tests do not quantify the level of topological correspondence between two networks.

The graph spectra, network motifs, and graphlets capture the local sub-

graph patterns better than the global network properties. Therefore, the comparison of these network properties produce more accurate model identification results. It is hard to interpret the spectral statistics of a network, since the spectrum of a graph cannot be translated into everyday language directly. Furthermore, more than one graph may have the same spectral profile, resulting with the failure of spectral methods in network comparison [201]. The information encoded in network motif and graphlet statistics can be translated into everyday language easily, as they represent which subgraph patterns appear in the network and which patterns do not. Among these two network properties, we focus on the graphlet statistics, since the interpretation of the motif-based methods is highly dependent on the chosen random network model to identify the over-represented and under-represented patterns [7].

Przulj et al. use graphlet-based network distance measures (i.e., RGF distance [157] and GDD-Agreement [156]) for identifying the best fitting network model among a number of alternatives. They compute the RGF distances and GDD-Agreements between the input network and the generated model networks, and accept the the model with the minimum average distance to the input network as the best-fitting model. Note that, although this method is suggested and widely-applied using the graphlet-based network distances measures, any other network distance heuristics can be applied in a similar way.

Rito et al. [163] criticizes the methodology of Przulj et al. [156], claiming that the method is good for comparing alternative models with each other but the network model that is at minimum distance to the input network does not necessarily fit the network. In other words, the obtained results are all relative to the compared models; even if none of the models actually fit the data, a well-fitting model is identified with this method. They suggest a non-parametric methodology for testing whether a model truly fits a network. This methodology is based on two distributions: (1) distribution of data-vs-model distances: represents the distances between the input network and the model networks, (2) distribution of model-vs-model distances: represents the distances between all model network pairs. If these two distributions intersect, this indicates that the model differs within itself as much as it differs from the input network. Therefore, the intersection between the two distributions is an indicator of model fit. Later on, Hayes

et al. [76] apply the non-parametric method to analyse the topologies of the seven network models that are listed in Section 1.5. They find out that the topology of the model networks are unstable below a certain sizes and edge densities.

The above discussed methods assess the network models for their ability to reproduce the observed structure of an input network. Another problem in network modelling is assessing the trade-off between the complexity of a model (i.e., the number of parameters that are necessary to define the model) and its goodness-of-fit. Network models that are able to reproduce the observed topology of an input network with less number of parameters are desired over more complex models. Given two network models M_1 and M_2, the trade-off between the goodness-of-fit and complexity of the models can be assessed by two statistical measures that are based on information theory: (1) Akaike Information Criterion (AIC) [1], and (2) Bayesian Information Criterion (BIC) [168]. Akaike information criterion is defined as:

$$AIC = 2k - 2\ln(L), \tag{1.28}$$

where k is the number of model parameters, and L is the maximized value of the likelihood function for the estimated model. AIC penalizes the high number of parameters while rewarding the goodness-of-fit determined by the maximum likelihood. Therefore, network models that have smaller AIC values are preferred. Bayesian Information Criterion (BIC) is another measure that evaluates the trade-off between the model complexity and its goodness-of-fit. BIC penalizes the number of model parameters more strongly than AIC, and it is defined as:

$$BIC = -2\ln(L) + k\ln(n), \tag{1.29}$$

where L is the maximized value of the likelihood function for the estimated model, k is the number of model parameters, and n is the number data points in the observed data. Unlike AIC, BIC depends on the number of data points in the observed data; e.g., number of nodes in the modelled networks. Similar to AIC, models with lower BIC scores are desired. For both models, the likelihood function of the estimated model is defined based on the goodness-of-fit statistics for the networks generated from the models. It should be noted that AIC and BIC scores only quantify the trade-off

between the goodness-of-fit and the model complexity; they do not evaluate the fit of a network model. For this reason, these scores should only be used when making a comparison between two well-fitting network models. We use AIC and BIC scores to compare the estimated exponential-family random graph models in Chapter 5.

1.6 Previous Studies on World Trade Networks

The world economy has never been a stable and easy-to-predict system as it is composed of many independent components that affect each other with their individual actions. The recent global recession has once again shown that a local malfunctioning in these economic components may have uncontrollable consequences on the world economy on a global scale. Insights into the functioning of the world economy can be mined from the flow of money between countries, which is woven into their trade relations. Network theory provides powerful methods for the analysis of world trade: countries are represented by nodes and trade relations between them are represented by edges (Section 1.2.1). These networks enable a global view of the complex system of world trade. Serrano et al. [170] show that in trade networks the majority of countries have a small number of trading partners while only a few countries have many trading partners (i.e., the networks have power-law degree distributions), the distances between countries are small (i.e., the networks have small-world property), the trade partners of a country also tend to trade among themselves (i.e., the networks have high clustering coefficient), and countries with many trade partners tend to connect to countries with a small number of trade partners (i.e., the networks are disassortative). Similarly, Kastelle et al. [101] evaluate the effects of globalisation on the world trade network topology by defining a measure of "globalisation". Their analysis show that some aspects of the world trade network have substantially changed over time, though the main network properties of the world trade network is stable over time, opposing to the idea of globalisation that assumes "everything is different now".

One of the main challenges in the world trade network analyses is defining network models that explain the observed topology of world trade networks. The Nobel Prize winning Gravity Model of Trade is the most well-known model for describing the rules of trade link formation [4]. This model

proposes that trade weight between two countries is proportional to their economic sizes, e.g., Gross Domestic Products (GDP), and inversely proportional to their geodesic distance. The success of this model in explaining the formation of world trade networks is evaluated by numerous studies [15, 43, 48, 62]. Garlaschellli et al. [62] evaluate the Gravity Model of Trade through standard network statistics (namely, degree distribution, clustering coefficient, and average nearest neighbourhood degree) without properly comparing them against the observed statistics of real-world networks. Biggiero et al. [15] test the correlation between the expected trade volumes produced by the Gravity Model of Trade and the observed trade volumes in real trade networks, concluding that due to the low correlation (~ 0.5), the model only roughly approximates but does not provide a complete explanation of the world trade network. Benedictis et al. [43] analyse the correspondence between the model network and the real trade network using density, degree distribution, closeness centrality and betweenness centrality, and conclude that the model networks and the real networks agree with respect to these properties. Finally, Dueñas et al. [48] show that the Gravity Model of Trade can partially replicate the topology of the weighted trade network, but only when the observed binary topology is kept fixed. However, they also show that the model is not able to explain the observed high clustering coefficient and cannot correctly predict the existence of a trade link. Overall, these studies suggest that the gravitational model can approximate some basic characteristics of world trade networks, but it is still an imperfect model that cannot fully explain all the topological properties of these networks.

Another well-accepted model of world trade networks is the Core-Periphery model [32, 44, 77, 84, 153, 176]. This model suggests a hierarchical organisation of countries, based on their trade relations: the richest countries form the core of the networks where all countries trade with each other, while the poor countries are located on the periphery of the network and trade only with core countries but not among themselves. There is an ongoing debate about the number of layers that this Core-Periphery model should contain; some studies recognise only two main layers — the core and the periphery [84] — while other studies argue for the need of an additional, semi-peripheral, layer [32, 153, 176]; yet others propose a hierarchical model without a definite number of layers [44, 77]. Even the definition of core and

periphery differs among studies. Piana et al. [153] define the core, semi-peripheral, and peripheral countries based on the domination power of a country over other countries in terms of trade, while Clark et al. [32] define the coreness of a country based on the local density around it. The study of He et al. [77] differs from the others in that it defines a measure of hierarchical organisation in the world trade network and uses this measure to evaluate the effect of globalisation and global recessions on the structure of the world trade network. They show that the hierarchical organisation of the world trade network is decreasing with the globalisation and that global recessions are followed by a recovery (increase) in the hierarchical organisation. A similar measure of core-periphery organisation in a network is proposed by Rossa et al. [44]. Their method uses a random walker on the network to rate the coreness of a country, and describes the core-periphery organisation in the network based on the distribution of these country ranks.

Network models are grouped into two: (1) descriptive models, which explain the structure of an input network, and (2) generative models, which are sets of rules for producing random networks with similar topological characteristics. Both the Gravity Model of Trade and the Core-Periphery models have been mostly used as descriptive models in the above listed studies. To the best of our knowledge, no generative random network models have been proposed so far that are based on the main principles of these two models.

So far, all models of world trade networks have been analysed independently, without a proper comparison among them that would evaluate which model best fits the world trade network. Performing a systematic comparison about the models of world trade networks (as explained in Section 1.5.3), and analysing which of these models best explain the topology of world trade networks is still an open research question, that may shed light on our understanding of the functional mechanisms in the world economy. In the light of these goodness-of-fit analyses, better network models can be proposed for explaining the topological structure of world trade networks.

1.7 Dissertation Outline

In this dissertation, we present solutions for comparing and modelling networks, and analyze five different types of real-world networks (i.e., networks

of autonomous systems, Facebook, metabolic, protein structure, and world trade) with a special emphasis on the world trade networks.

In Chapter 2, we introduce a new network topology statistic, Graphlet Correlation Matrix, and make use of this statistic to derive a network distance measure, the Graphlet Correlation Distance. The graphlet correlation matrix provides a description of a network's topology with respect to the dependencies among the graphlet degrees of non-redundant orbits. Comparing these topological descriptors for different networks, we obtain the best network comparison heuristic for model clustering. We show that graphlet correlation distance is noise-tolerant, performs surprisingly well even with partial node properties, and has lower computational complexity than any of the previous graphlet-based network distance measures.

In Chapter 3, we analyse the world trade networks in detail with our new methodology. We question the organizational principles of world trade networks using graphlet correlation matrices, and link the changes in world trade network topology with the changes in crude oil price. As the crude oil price is a direct indicator of global recessions, we analyse the causes of observed changes in world trade network topology during crisis years, based on the change in the number of graphlets on these networks. Then, we link the position of a country on the world trade network with its economic wealth in the light of the organisational principles obtained from the graphlet correlation matrix.

In Chapter 4, we test different network models for their fit on five different types of real-world networks; i.e., autonomous systems, Facebook, metabolic, protein structure, and world trade. None of the tested models fit to world trade networks, raising the need for defining new models of world trade networks. We propose two such models and show that these models fit world trade networks. The best of these two models is built based on our observations on the graphlet correlation matrices of world trade networks and forms a three-layer organization by maximizing the number of a broker-type graphlet, in particular G_{23} (Figure 1.4), in the network. We analyse the world trade networks further based on the properties of G_{23}, showing the predictive power of the wealth of a country on its future broker position.

In Chapter 5, being encouraged by the success of graphlet based modelling on world trade networks, we introduce a new generic framework for

network modelling based on a wide-range of graphlet based network properties. We exploit the exponential-family random graph models for generating this framework, and introduce four different graphlet-based change score functions for use with this network modelling method. These new ERGM terms not only test the significance of certain graphlet frequencies, but also relate node attributes with graphlet patterns in the context of an ERGM.

Finally, in Chapter 6, we conclude the dissertation by providing a brief summary of our contributions, and introduce our preliminary results on four different research problems as future work.

2 Network Analysis & Comparison: Graphlet Correlations Approach

In this chapter, we explain the redundancies and dependencies in the graphlet degree vectors of nodes (Section 2.2), and use these redundancies and dependencies to introduce a new network topology statistic (Section 2.3.1) and a new topological network distance measure (Section 2.3.2). This new distance measure outperforms all of the state-of-the-art network distance measures in model identification, and is computationally less expensive than the other graphlet based measures.

2.1 Motivation

The descriptive power of graphlets – small, connected, non-isomorphic, and induced subgraphs of a large network (Figure 1.4) – have been widely exploited for comparing network topologies and mining networks for local topological similarities [136, 156, 157]. Though current graphlet based methods are shown to be successful, there is still room for improving these techniques. First, since smaller graphlets appear in larger graphlets (e.g., graphlet G_1 appears in graphlet G_3 two times), graphlet statistics are not independent. The statistics of larger graphlets are bound by the statistics of smaller graphlets, creating redundancies and dependencies in the graphlet degrees of nodes. These redundancies and dependencies in graphlet statistics are not correctly tackled by current graphlet based network comparison methods (i.e., RGF distance, GDD Agreement – Section 1.4), causing uneven weighting of different graphlet statistics during the computation of network distances. Second, the computation of 5-node graphlet statistics increases the computational complexity, reducing the applicability

of graphlet based techniques on very large networks such as online social networks. Despite the high computational cost, the contribution and necessity of 5-node graphlet statistics for network comparison have not been systematically evaluated before.

In this chapter of the dissertation, we first identify all redundant statistics in the graphlet degree vectors of nodes. After eliminating the redundant statistics, there still remain dependencies in the graphlet degrees of different orbits, due to the existence of smaller graphlets in larger ones. We quantify the level of dependencies among the non-redundant orbits using Spearman's Correlation Coefficient. Interestingly, networks with different topologies show different levels of orbit dependencies. We exploit this observation for defining a new network topology statistic, called Graphlet Correlation Matrix, which explains the topology of a network in terms of relative graphlet appearances. Furthermore, we use this network statistic to contrast network topologies, defining a new network distance measure called Graphlet Correlation Distance (GCD). We test the model identification performance of GCD in detail, and systematically compare it with the state-of-the-art network distance measures. Moreover, we contrast the performance of these network distance measures in the existence of noise in the networks, and also based on subsets of network statistics. Finally, we analyse the computational complexities of these network distance measures, highlighting the obtained improvement on graphlet based network distance measures.

2.2 Redundancies and Dependencies in Graphlet Degree Vectors

Graphlets are small, connected, non-isomorphic and induced subgraphs of a large network (Figure 1.4). Graphlet based network statistics, such as the number of times they appear in a network or the number of times they touch a node at a specific orientation, provide a detailed description of the network topology. The statistics of different graphlets are not independent of each other. This is mainly due to fact that smaller graphlets may appear as induced subgraphs of larger graphlets. In this respect, edges (i.e. G_0 graphlets) are the building blocks of all graphlets. Therefore, the number of

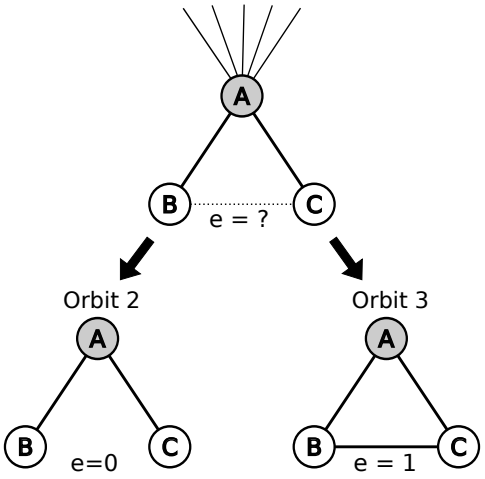


Figure 2.1: Illustration of graphlet degree redundancies among orbits $\{0, 2, 3\}$. When two edges $\{A, B\}$ and $\{A, C\}$ are combined at orbit 0, forming the induced subgraph of $\{A, B, C\}$, node A corresponds to orbit 2 if B and C are disconnected, and orbit 3 otherwise. Therefore, $\binom{C_0}{2}$ is equal to the sum of C_2 and C_3, where C_i represents the graphlet degree for orbit i.

edges in a network define an upper bound on the number of larger graphlets that can appear in the network. In a combinatorial perspective, larger graphlets are formed as combinations of smaller graphlets, and therefore their statistics are bounded by the statistics of the smaller graphlets. The descriptive power of larger graphlet statistics comes from the information provided about the distributions of larger graphlets in a network within the upper limit defined by the smaller graphlets.

This phenomena indicates the existence of redundancies in the 73 dimensional graphlet degree vectors (GDVs) of nodes: an orbit is *redundant* if its graphlet degree can be derived from the graphlet degrees of a set of other orbits. The simplest example of redundancies is observed among orbits 0, 2, and 3 when two edges (G_0) are "combined" at orbit 0 as illustrated in Figure 2.1. Given two adjacent edges, (A, B) and (A, C), the orbit touching A from the graphlet induced by $\{A, B, C\}$ is either orbit 3 if B and C are connected by an edge, or orbit 2 otherwise. Therefore, $\binom{C_0}{2}$ is equal to the sum of C_2 and C_3, where C_i represents the graphlet degree for orbit i.

When combining graphlets for producing larger graphlets, the same or-

bits may be produced by more than one graphlet combination. For example, combining a graphlet G_1 at orbit 2 with an edge (G_0 – orbit 0), the node at the combination point may correspond to orbits 7, 11, or 13. Let us consider the case where the combination point corresponds to orbit 7, the corresponding graphlet, G_4, being the subgraph of nodes $\{A, B, C, D\}$ and node A being the combination point (illustrated in Figure 2.2). This configuration can be obtained by three different graphlet combinations: (1) Combination of $\{B, A, C\}$ with $\{A, D\}$, (2) Combination of $\{C, A, D\}$ with $\{A, B\}$, and (3) Combination of $\{B, A, D\}$ with $\{A, C\}$. Therefore, in the corresponding redundancy equation, the C_7 count should be multiplied by 3. Similarly, for the case that combination point corresponds to orbit 11, the C_{11} count should be multiplied by 2, since there are two different G_1 that can make the combination point correspond to orbit 11.

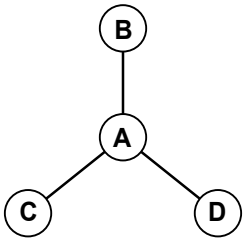


Figure 2.2: Example graphlet that is used for explaining the redundancy weighting. This graphlet can be formed combining a G_1 with an edge (i.e., G_0) at node A, where node A respectively corresponds to orbits 2 and 0.

We systematically test all combinations of 2-, 3-, and 4-node graphlets that produce graphlets of size ≤ 5 for producing the corresponding redundancy equations. We algorithmically identify 26 such combinations by implementing an automated procedure that systematically combines graphlets at different orbits, and identifies the graphlet orbits that can be produced as a result of these combinations. This procedure produces 26 redundancy equations. However, only 17 of these equations are independent from each other; i.e., they cannot be derived from the other equations. Different groups of 17 independent equations can be chosen from the complete set of 26 equations. A set of 17 independent equations is listed as follows:

1. $\binom{C_0}{2} = C_2 + \mathbf{C_3}$

2. $\binom{C_2}{1}\binom{C_0-2}{1} = 3C_7 + 2C_{11} + \mathbf{C_{13}}$

3. $\binom{C_1}{1}\binom{C_0-1}{1} = C_5 + 2C_8 + C_{10} + 2\mathbf{C_{12}}$

4. $\binom{C_3}{1}\binom{C_0-2}{1} = C_{11} + 2C_{13} + 3\mathbf{C_{14}}$

5. $\binom{C_4}{1}\binom{C_0-1}{1} = \mathbf{C_{16}} + C_{29} + 2C_{34} + 2C_{36} + 2C_{46} + C_{51} + 2C_{52} + C_{59}$

6. $\binom{C_5}{1}\binom{C_0-2}{1} = 2\mathbf{C_{21}} + C_{26} + 2C_{30} + 2C_{38} + 2C_{47} + C_{48} + C_{53} + C_{60}$

7. $\binom{C_6}{1}\binom{C_0-1}{1} = \mathbf{C_{20}} + C_{32} + C_{37} + C_{40} + 2C_{49} + 2C_{54}$

8. $\binom{C_7}{1}\binom{C_0-3}{1} = 4\mathbf{C_{23}} + 2C_{33} + C_{42} + C_{55}$

9. $\binom{C_8}{1}\binom{C_0-2}{1} = \mathbf{C_{38}} + 3C_{50} + C_{53} + 2C_{63} + C_{64} + C_{68}$

10. $\binom{C_9}{1}\binom{C_0-1}{1} = \mathbf{C_{28}} + C_{43} + C_{51} + C_{59} + 2C_{62} + 2C_{65}$

11. $\binom{C_{10}}{1}\binom{C_0-2}{1} = \mathbf{C_{26}} + 2C_{41} + C_{48} + C_{53} + 2C_{57} + C_{60} + 2C_{64} + 2C_{66}$

12. $\binom{C_{11}}{1}\binom{C_0-3}{1} = 2C_{33} + 2C_{42} + 4\mathbf{C_{44}} + 3C_{58} + 2C_{61} + C_{67}$

13. $\binom{C_{12}}{1}\binom{C_0-2}{1} = \mathbf{C_{47}} + C_{60} + C_{63} + C_{66} + 2C_{68} + 3C_{70}$

14. $\binom{C_{13}}{1}\binom{C_0-3}{1} = C_{42} + 3C_{55} + 2C_{61} + 2C_{67} + 4\mathbf{C_{69}} + 2C_{71}$

15. $\binom{C_1}{2} = C_6 + C_8 + C_9 + C_{12} + \mathbf{C_{17}} + C_{25} + C_{34} + C_{37} + C_{40} + 2C_{49} + C_{51} + C_{52} + 2C_{54} + C_{59} + 2C_{62} + 2C_{65}$

16. $\binom{C_3}{2} = C_{13} + 3C_{14} + C_{44} + C_{61} + C_{67} + 2C_{69} + 2C_{71} + 3\mathbf{C_{72}}$

17. $\binom{C_2}{1}\binom{C_3}{1} = 2C_{11} + 2C_{13} + C_{33} + 2C_{42} + 3C_{55} + 3C_{58} + C_{61} + 2C_{67} + \mathbf{C_{71}}$

The remaining 9 equations that can be derived from the 17 independent equations are listed below:

18. $\binom{C_0}{3} = C_7 + C_{11} + C_{13} + C_{14}$

19. $\binom{C_0}{4} = C_{23} + C_{33} + C_{42} + C_{44} + C_{55} + C_{58} + C_{61} + C_{67} + C_{69} + C_{71} + C_{72}$

20. $\binom{C_1}{1}\binom{C_0-1}{2} = C_{21} + C_{26} + C_{30} + 2C_{38} + C_{41} + 2C_{47} + C_{48} + 3C_{50} + 2C_{53} + C_{57} + 2C_{60} + 3C_{63} + 2C_{64} + 2C_{66} + 3C_{68} + 3C_{70}$

21. $\binom{C_2}{1}\binom{C_0-2}{2} = 6C_{23} + 5C_{33} + 4C_{42} + 4C_{44} + 3C_{55} + 3C_{58} + 3C_{61} + 2C_{67} + 2C_{69} + C_{71}$

22. $\binom{C_3}{1}\binom{C_0-2}{2} = C_{33} + 2C_{42} + 2C_{44} + 3C_{55} + 3C_{58} + 3C_{61} + 4C_{67} + 4C_{69} + 5C_{71} + 6C_{72}$

23. $\binom{C_{14}}{1}\binom{C_0-3}{1} = C_{58} + C_{67} + 2C_{71} + 4C_{72}$

24. $\binom{C_2}{2} = 3C_7 + C_{11} + 3C_{23} + 2C_{33} + C_{42} + 2C_{44} + C_{61} + C_{69}$

25. $\binom{C_1}{1}\binom{C_2}{1} = C_5 + 2C_8 + C_{21} + C_{26} + 2C_{38} + C_{41} + 2C_{47} + 3C_{50} + C_{53} + C_{60} + 2C_{63} + C_{68}$

26. $\binom{C_1}{1}\binom{C_3}{1} = C_{10} + 2C_{12} + C_{30} + C_{48} + C_{53} + C_{57} + C_{60} + C_{63} + 2C_{64} + 2C_{66} + 2C_{68} + 3C_{70}$

For example, $Eq.18$ is equivalent to $(Eq.2 + Eq.4)/3$, when C_3 is replaced by using $Eq.1$:

- $(Eq.2 + Eq.4)/3 : (C_2(C_0 - 2) + C_3(C_0 - 2))/3 = C_7 + C_{11} + C_{13} + C_{14}$
- From $Eq.1 : C_3 = \binom{C_0}{2} - C_2$
- Replacing C_3 by the term from Eq.1 in $(Eq.2 + Eq.4)/3$: $\frac{C_2(C_0-2)+(\binom{C_0}{2}-C_2)(C_0-2)}{3} = C_7 + C_{11} + C_{13} + C_{14}$
- Simplifies to: $\frac{\binom{C_0}{2}(C_0-2)}{3} = C_7 + C_{11} + C_{13} + C_{14}$
- Which is exactly $Eq.18 : \binom{C_0}{3} = C_7 + C_{11} + C_{13} + C_{14}$

Other equations from the above list, numbered 18-26, can be similarly derived from the 17 independent equations.

We use these equations to remove redundant orbits from graphlet degree vectors, so they will not contain redundant information. Since there are 17 independent equations, we can eliminate up to 17 orbits as redundant. The set of 17 independent equations, and the 17 corresponding redundant orbits are chosen arbitrarily based on the 26 redundancy equations. Therefore, one can eliminate a different set of 17 orbits based on these 26 equations. One arbitrary set of redundant orbits that can be eliminated from graphlet degree vectors is written in bold in the first 17 equations. Similarly, for 2- to 4-node graphlets, we can eliminate up to 4 orbits as redundant. We chose to eliminate orbits 3, 12, 13 and 14 using Equations 1, 2, 3, and 4. The remaining set of 11 non-redundant orbits are illustrated in Figure 2.3.

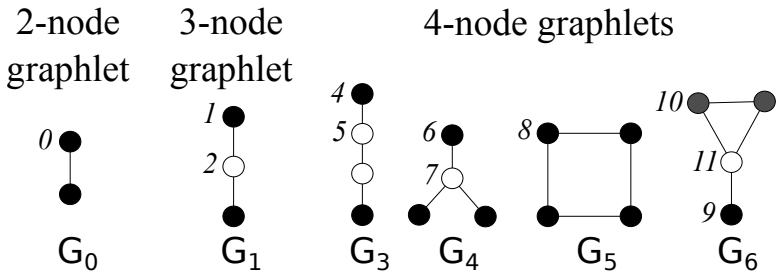


Figure 2.3: The list of 2- to 4-node non-redundant graphlet orbits. The non-redundant set of orbits are chosen based on redundancy Equations 1, 2, 3, and 4.

Eliminating the redundant orbits in graphlet degree vectors reduce the noise effect of these orbits on the graphlet degree vector based distance measures. However, there are dependencies among orbits even in the non-redundant orbit set, since the small graphlets that appear in the larger ones. If a small graphlet is an induced subgraph of a larger graphlet, and orbit j in the larger graphlet corresponds to orbit i when induced on the small graphlet, then orbits i and j are dependent; e.g., the dependencies for orbit 21 is illustrated in Figure 2.4. In this respect, the simplest dependency is between graphlet G_0 and all other graphlets. The number of graphlets that can appear in a network are all bounded by the number of edges in the network due to this dependency. The orbit dependencies for all orbits of 2- to 5-node graphlets are provided in Table 2.1. The level of dependency between two orbits, i and j, is quantified by computing the Spearman's Correlation Coefficient [179] among the i^{th} and j^{th} graphlet degrees of all nodes.

2.3 Graphlet Correlations

It is expected to observe a positive Spearman's Correlation between the graphlet degrees of two dependent orbits (Table 2.1). The interesting questions to investigate are: How do the independent orbits correlate with each other in a network? Are these correlation patterns consistent among networks from the same models? Can this information be used for identifying topological similarities among networks? We investigate the answers

Table 2.1: Complete list of orbit dependencies for all 2- to 5-node graphlet orbits.

Orbit	Dependent Orbits	Orbit	Dependent Orbits
1	0	**37**	0, 1, 2, 5, 6, 8
2	0	**38**	0, 1, 2, 5, 7, 8
3	0	**39**	0, 1, 2, 7, 9
4	0, 1	**40**	0, 1, 3, 6, 10, 12
5	0, 1, 2	**41**	0, 1, 2, 3, 10, 13
6	0, 1	**42**	0, 2, 3, 7, 11, 13
7	0, 2	**43**	0, 1, 3, 9, 10
8	0, 1, 2	**44**	0, 2, 3, 11
9	0, 1	**45**	0, 1, 4, 9
10	0, 1, 3	**46**	0, 1, 3, 4, 12
11	0, 2, 3	**47**	0, 1, 2, 3, 5, 11, 12
12	0, 1, 3	**48**	0, 1, 2, 3, 5, 10, 13
13	0, 2, 3	**49**	0, 1, 2, 6, 8
14	0, 3	**50**	0, 1, 2, 7, 8
15	0, 1, 4	**51**	0, 1, 2, 4, 5, 8, 9
16	0, 1, 2, 4, 5	**52**	0, 1, 3, 4, 10
17	0, 1, 2, 5	**53**	0, 1, 2, 3, 5, 8, 10, 11
18	0, 1, 4	**54**	0, 1, 3, 6, 12
19	0, 1, 4, 6	**55**	0, 2, 3, 7, 13
20	0, 1, 2, 5, 6	**56**	0, 1, 9
21	0, 1, 2, 5, 7	**57**	0, 1, 3, 10, 14
22	0, 1, 6	**58**	0, 2, 3, 11, 14
23	0, 2, 7	**59**	0, 1, 3, 4, 9, 10, 12
24	0, 1, 4, 9	**60**	0, 1, 2, 3, 5, 10, 12, 13
25	0, 1, 3, 10	**61**	0, 2, 3, 11, 13
26	0, 1, 2, 3, 10, 11	**62**	0, 1, 2, 8, 9
27	0, 1, 4	**63**	0, 1, 2, 3, 8, 11, 12
28	0, 1, 2, 5, 9	**64**	0, 1, 2, 3, 8, 10, 13
29	0, 1, 3, 4, 10	**65**	0, 1, 3, 9, 12
30	0, 1, 2, 3, 5, 11	**66**	0, 1, 3, 10, 12, 14
31	0, 1, 6, 9	**67**	0, 2, 3, 11, 13, 14
32	0, 1, 3, 6, 10	**68**	0, 1, 2, 3, 8, 12, 13
33	0, 2, 3, 7, 11	**69**	0, 2, 3, 13
34	0, 1, 2, 4, 5	**70**	0, 1, 3, 12, 14
35	0, 1, 4, 6	**71**	0, 2, 3, 13, 14
36	0, 1, 2, 4, 8	**72**	0, 3, 14

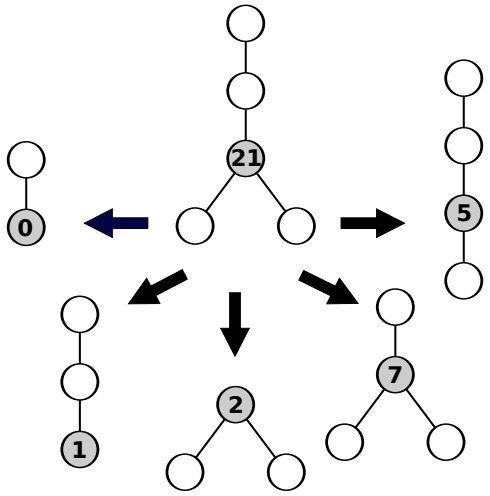


Figure 2.4: Graphlet orbit dependencies for orbit 21. The induced subgraphs of graphlet G_{10} are illustrated. Orbit 21 of graphlet G_{10} corresponds to orbits $\{0, 1, 2, 5, 7\}$ in its induced subgraphs, making orbit 21 dependent on these orbits.

to these questions by proposing a new network statistic called *Graphlet Correlation Matrix* and a new topological network distance measure called *Graphlet Correlation Distance*.

2.3.1 Graphlet Correlation Matrix

The Graphlet Correlation Matrix is a new network statistic that encodes the topology of a network using the Spearman's Correlation Coefficients among various node properties contained in graphlet degrees, over all nodes. Given a network $G(V, E)$, first we compute graphlet degree vectors of all nodes, $v \in V$, and construct a matrix where each row represents the graphlet degree vector of a node, $GDV(v)$. We exploit the existence of dependencies between orbits by computing the Spearman's correlation coefficient among all pairs of orbits (i.e., among all columns of the matrix of graphlet degree vectors) and present them in a $n \times n$ symmetric matrix that we name as the *Graphlet Correlation Matrix* of network, GCM_G. Graphlet correlation matrices can be defined using different sets of orbits. We focus on two particular orbit sets in our experiments: (1) 11 non-redundant orbits of 2- to 4-node graphlets (illustrated in Figure 2.3), (2) the complete set of 73 orbits of 2- to 5-node

graphlets (illustrated in Figure 1.4). In this way, we can encode the topology of a network of any size into an $n \times n$ symmetric matrix with values in the interval $[-1, 1]$, where n is the number of orbits that are used for computing the GCM. Graphlet Correlation Matrix computation is illustrated in Figure 2.5 on a random geometric graph with 500 nodes and 1% edge density.

Networks that have different topologies are expected to have different graphlet correlation matrices. For example, Figure 2.6 illustrate the graphlet correlation matrices of four different networks: a scale-free network that is generated by the preferential attachment (i.e., Barabàsi-Albert) model, a network generated by the geometric random network model, the world trade network of 2010, and the human metabolic network. In agreement with known properties of scale-free Barabási-Albert (SF-BA) networks, orbits 0, 2, 5, and 7, which are characteristic to existence of hubs, form a cluster of dependent orbits with their correlation coefficients being close to 1 (Figure 2.6–A). Orbits 10 and 11, which are characteristic to existence of clustering "near" hubs, also form a cluster of correlated orbits. Finally, orbits 1, 4, 6, and 9, which are characteristic to existence of a large number of degree 1 nodes, are dependent as well. The picture is quite different for geometric random graphs (GEO) of the same size, which have Poisson degree distributions, and hence the structure is not dominated by a large fraction of degree 1 nodes and a small number of hubs (Figure 2.6–B).

Uncovering orbit dependencies in real-world networks is much more interesting, since they can reveal currently unknown organizational principles of these networks. Indeed, the world-trade network of 2010 [34] contains two large clusters of dependent orbits, $\{0, 2, 5, 7, 8, 10, 11\}$ and $\{6, 9, 4, 1\}$, while there is anti-correlation between orbits $\{4, 6, 9\}$ and orbits $\{0, 2, 5, 7, 8, 10, 11\}$ (Figure 2.6-C). Investigating the implications of this, we notice that orbits $4, 6$ and 9 correspond to *peripheral*, degree 1 nodes that are "hanging" from graphlets G_3, G_4 and G_6 (Figure 2.3), while members of the large cluster of correlated orbits, $\{0, 2, 5, 7, 8, 10, 11\}$, correspond to higher degree, either clustered (in a dense neighbourhood), or *broker*-type (*mediators* between nodes that are not directly interacting) orbits. Since these two clusters are anti-correlated, we can conclude that countries are either clustered/brokers, or on the periphery of the world trade [44], but not both. Hence, GCM unveils a hidden structure of this network that can be further interpreted qualitatively: through further analysis presented below, we interpret this

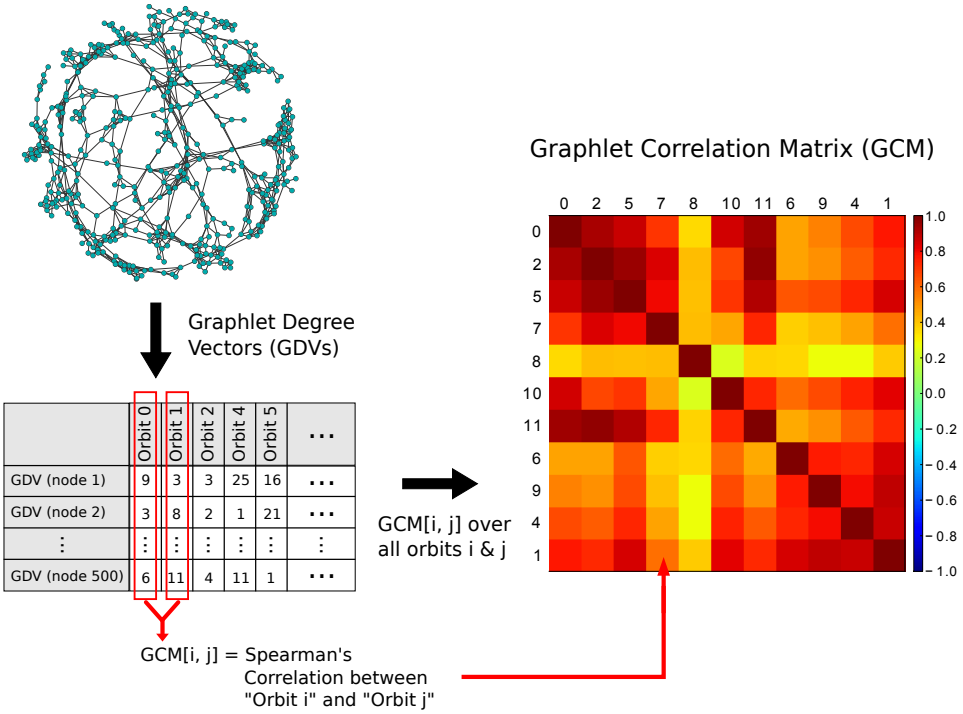


Figure 2.5: Graphlet Correlation Matrix computation is illustrated on a geometric network G with 500 nodes and 1% edge density (the network on the left). In the matrix of graphlet degree vectors (shown on the left), each row represents the graphlet degree vector of a node, and each column contains the graphlet degrees of all nodes for orbit i, d_G^i. The graphlet degrees of orbits 0 and 1, d_G^0 and d_G^1 are highlighted in red. The graphlet correlation between orbits i and j, $GCM_G[i,j]$, is the Spearman's correlation coefficient between d_G^i and d_G^j. Computing the $GCM_G[i,j]$ for all pairs of orbits, we obtain the symmetric graphlet correlation matrix of G, GCM_G. The rows and columns of the GCM_G are ordered based on the correlation similarities of orbits for visualising the orbit clustering patterns better.

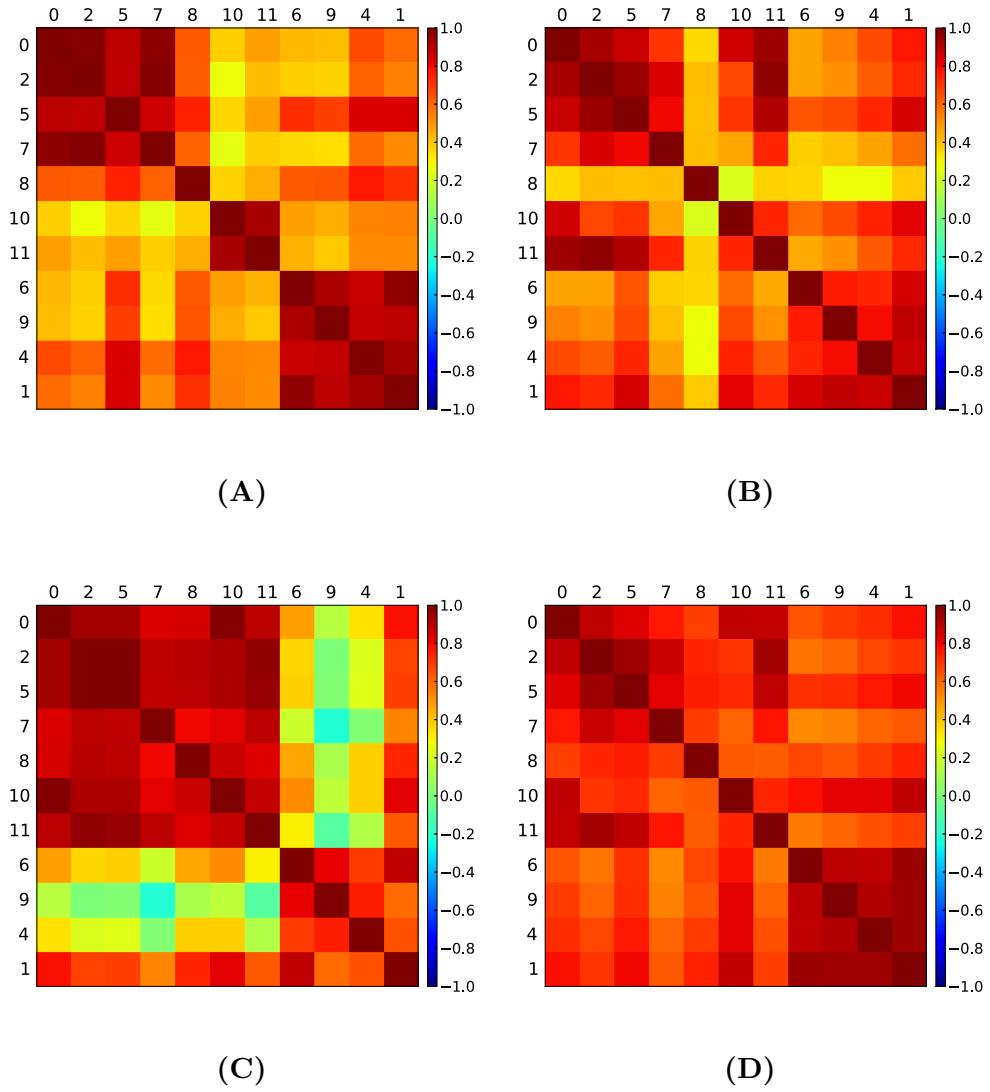


Figure 2.6: Graphlet Correlation Matrices (GCMs) of different types of networks: Panel A – a scale-free Barabàsi-Albert (SF-BA) network with 500 nodes and 1% edge-density; Panel B – a geometric random network (GEO) with 500 nodes and 1% edge-density; Panel C – the world trade network of 2010; and Panel D – the human metabolic network. The rows and columns of the GCMs are ordered based on the correlation similarities of orbits for visualising the orbit clustering patterns better.

observation on 49 world trade networks corresponding to trade data from 1962 to 2010. In contrast, the topology of the human metabolic network [98] is very different from the topology of world trade networks: the correlations between all orbits are high, indicating that constituent bio-molecules can be at the same time both peripheral and clustered/broker (Figure 2.6-D).

It is possible that a graphlet does not appear in a network. When this is the case, graphlet degrees of all nodes are equal to 0 for the corresponding orbits. Since the graphlet degrees are constant for all nodes, Spearman's Correlation coefficient cannot be computed for these orbits. To overcome this problem, we include a dummy graphlet degree vector, $[1, 1, ..., 1]$, into the matrix of graphlet degree vectors. This small amount of noise resolves the Spearman's correlation coefficient computation problem. As a result, the problematic orbits correlate perfectly (having Spearman's correlation coefficients of 1) while these orbits do not correlate with the rest of the non-zero orbits (having Spearman's correlation coefficients close to 0).

The graphlet degrees of different orbits do not scale within the same intervals, due to the differences in the search spaces of orbits. For example, graphlet degree of orbit 15 searches up to 4^{th} neighbourhood of a node, while graphlet degree for orbit 7 is only dependent on the 1^{st} neighbourhood, which causes the graphlet degrees of orbit 15 to span at a wider range. The graphlet degree ranges might even differ for orbits that search the same distance neighbourhoods, since the chances of each graphlet's appearance are not distributed evenly and depend on the density of the network. Due to the differences in the graphlet degree scales, a ranking based correlation coefficient that measures monotonic correlations between orbits (i.e., Spearman's Correlation Coefficient) is preferable over a correlation coefficient that measures the linear correlations among graphlet degrees (i.e., Pearson's Correlation Coefficient) for measuring the correlation between the graphlet degrees of different orbits. This is the reason for us to define the Graphlet Correlation Matrices based on Spearman's Correlation Coefficients rather than any other correlation coefficients.

2.3.2 Graphlet Correlation Distance

Apart from enabling in-depth examination of the topological organisation in a network, GCMs can also be used for quantifying the topological correspon-

dence between two networks. Being encouraged by the differences observed for the GCMs of different networks (Figure 2.6), we define a new network distance measure that we term *Graphlet Correlation Distance (GCD)*. The GCD between two networks, G_1 and G_2, is the Euclidean distance of the upper triangle values of their GCMs that are constructed based on d orbits:

$$GCD(G_1, G_2) = \sqrt{\sum_{i=1}^{d} \sum_{j=i+1}^{d} (GCM_{G_1}(i,j) - GCM_{G_2}(i,j))^2}. \quad (2.1)$$

In this dissertation, GCD-11 denotes the graphlet correlation distance that is computed from the 11×11 GCM of non-redundant 2- to 4-node graphlet orbits (orbits in Figure 2.3). Similarly, GCD-73 denotes the graphlet correlation distance that is computed from the 73×73 GCM of all 2- to 5-node graphlet orbits (all orbits in Figure 1.4). We aim to emphasize the larger differences rather than accounting for smaller differences between the correlations of orbit pairs, and have a robust distance measure by using the Euclidean distance (that is in ℓ_2 form) rather than Manhattan distance (that is in ℓ_1 form).

2.4 Validation Results

In this section, we evaluate the model clustering performance of graphlet correlation distance in comparison to the state-of-the-art network distance measures, assess its performance on classifying noisy networks, and also assess its performance on networks with sampled network properties.

2.4.1 Performance on Model Clustering

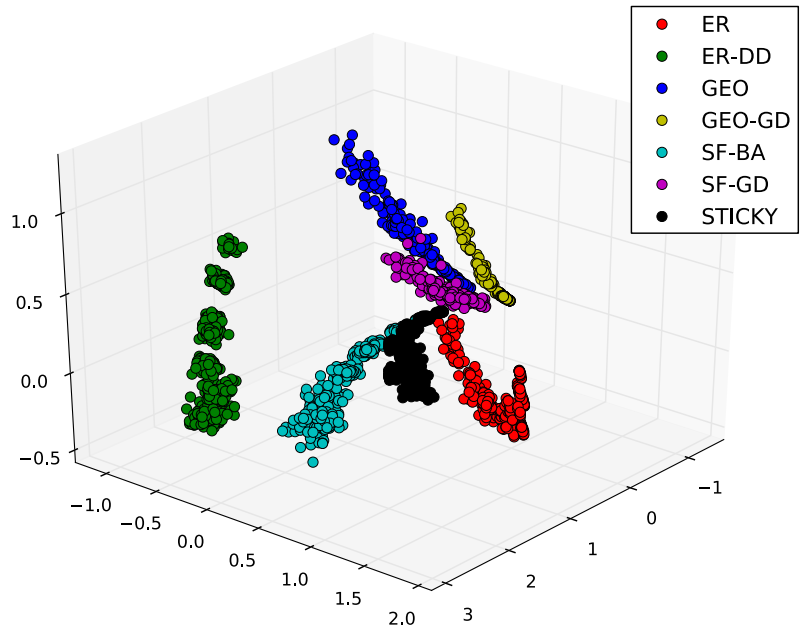
We use synthetic networks that are generated from the seven networks models (i.e., ER, ER-DD, SF, SF-GD, GEO, GEO-GD, and STICKY models – Section 1.5) for assessing the performance of GCD on clustering networks of the same type. It is infeasible to perform the model clustering experiments so as to cover the size and densities of all observed real-world networks. It is also known that networks from different models are better separated with increasing network sizes. The better separation of networks from different models simplify the model clustering tests, and make our experiments less stringent. Most of the real-world networks contain between 1,000 to

6,000 nodes, and have densities between 0.5% to 1%; e.g., the sizes and densities of the real-world networks that are analysed in the scope of this dissertation can be found in Table 4.1. For this reason, we chose the sizes and densities for the model clustering experiments so as to cover the most commonly observed sizes and densities of the real-world networks. In this respect, from each model, we generate 30 networks for each combination of the following node sizes and edge densities: $\{1000, 2000, 4000, 6000\}$ nodes and $\{0.5\%, 0.75\%, 1\%\}$ edge densities. Hence, the total number of synthetic networks that we generate is $7 \times 4 \times 3 \times 30 = 2,520$.

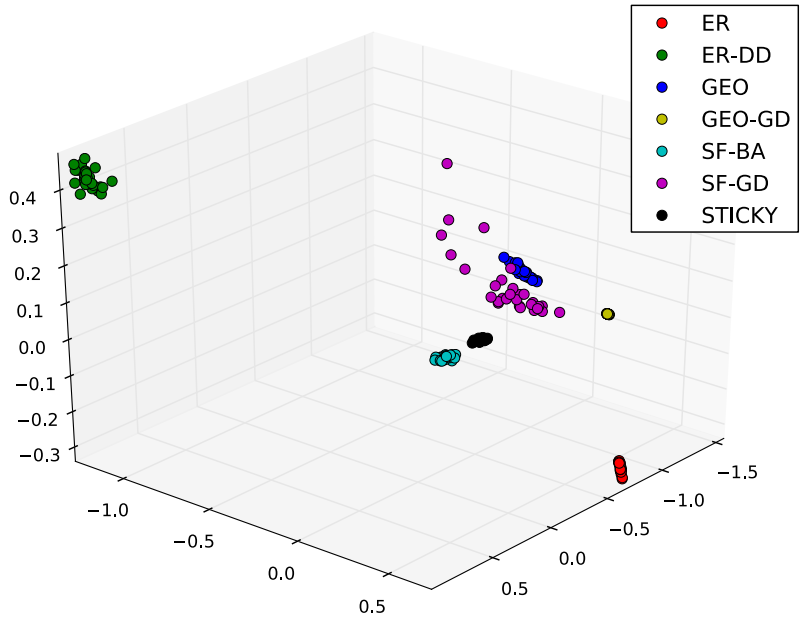
In order to assess whether GCD-11 is able to cluster networks that are generated from the same models together, we compute the GCD-11 distances between all pairs of the $2,520$ synthetic networks. For illustrating the clustering of these networks based on GCD-11, we use the standard method of multi-dimensional scaling (MDS) [37] using the squared metric stress criterion. We embed the $2,520$ networks as points into 3-dimensional space so that their GCD-11 distances are preserved as much as possible. As illustrated in Figure 2.7-A, networks belonging to the same model are grouped together in space regardless of size and edge-density. Model networks of the same size and density are grouped even better (Figure 2.7-B).

We illustrate GCD-11's performance on grouping real-world networks from the same domain by applying the same embedding methodology on $11,407$ real-world networks from five different domains: 733 autonomous networks of routers that form the Internet, Facebook networks of 98 universities, metabolic networks of enzymes of $2,301$ organisms, $8,226$ protein structure networks, and 49 world trade networks corresponding to years 1962 to 2010 (details are provided in Section 1.2). As in the case of model networks, MDS embedding of GCD-11 distances among the $11,407$ networks shows clear clustering among networks from the same domain (Figure 2.8).

We formally assess the model clustering performance of GCD by comparing its clustering quality with other state-of-the-art network distance measures. In particular, the model clustering performance of a network distance measure can be tested and quantified by using the standard Receiver Operator Characteristic (ROC) Curve [18]. Network pairs that are generated from the same model define the *True* set of the evaluation, while network pairs that are generated from different models define the *False* set. For small increments of parameter $\epsilon > 0$, four statistics are computed:



(A)



(B)

Figure 2.7: 3-Dimensional embedding of model networks based on GCD-11 distances: Panel A – 3D embedding of all 2,520 model networks that have 1000, 2000, 4000, 6000 nodes and 0.5%, 0.75%, 1% edge-density. Panel B – 3D embedding of 210 model networks that have 6000 nodes and 1% density.

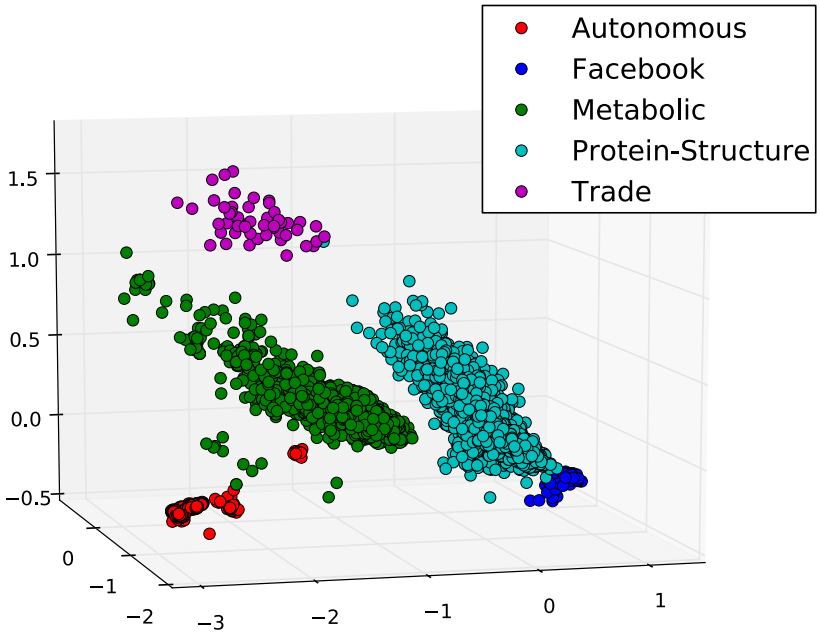


Figure 2.8: 3-Dimensional embedding of 11,407 real-world networks from different real-world domains (i.e., autonomous systems, Facebook, metabolic, protein structure, and world trade networks) based on GCD-11 distances.

1. TP – the number of True pairs having pairwise distances smaller than ϵ,
2. TN – the number of False pairs having pairwise distances greater or equal to ϵ,
3. FN – the number of True pairs having pairwise distances greater or equal to ϵ, and
4. FP – the number of False pairs having pairwise distances smaller than ϵ.

From these four statistics, we compute the *True Positive Rate (TPR)* that is the fraction of networks correctly grouped together and the *False Positive Rate (FPR)* that is the fraction of networks incorrectly grouped together as

follows:

$$\text{True Positive Rate} = \frac{TP}{TP + FN} \tag{2.2}$$

$$\text{False Positive Rate} = \frac{FP}{FP + TN} \tag{2.3}$$

ROC curve is obtained by plotting the TPR against FPR for all increments of ϵ. The Area Under the ROC curve (AUC) standardly measures the quality of the grouping by a given distance measure: for two randomly chosen pairs of elements, one pair from the True set and the other pair from False set, AUC represents the probability that the distance between the pair of elements from the True set will be smaller than the distance between the pair of elements from False set. An additional measure of quality is the *maximum accuracy* achieved over all values of ϵ:

$$Max.Accuracy = \arg\max_{\epsilon} \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.4}$$

Some studies [187, 205] argue that the early identification is more important than the overall class separation performance that is identified by the ROC curves and corresponding AUC scores. In these studies, distance measures that optimize the number of correctly clustered pairs of networks that are at the shortest distance, and hence are *retrieved first* by the distance measure, are accepted to perform better since most clustering algorithms aim to group objects that are at the smallest distances to each other. A standard measure to evaluate this "early identification" performance is truncated ROC (ROC_n) curves. ROC_n curves measure TPR against FPR up to a given false positive threshold n (i.e., n false positives are allowed) [205]. The average number of incorrectly clustered networks per query network is commonly called as *Errors Per Query (EPQ)*; i.e., $n = EPQ \times N$ where N is the number of networks in the comparison. Analogous to AUC scores of ROC curves, the area under the ROC_n curves are annotated with AUC_n, where n is the false positive threshold of the ROC_n computation.

Using these performance measures and the $2,520$ model networks that are illustrated in Figure 2.7-A, we evaluate the model clustering performances of different network distance measures. ROC and ROC_n curves are computed over two sets of distances: (1) the distances between all pairs of the $2,520$ model networks — $\binom{2,520}{2} = 3,173,940$ network pairs, and (2) the distances

between same size and edge-density model networks — $4 \times 3 \times \binom{7\times30}{2} = 263,340$ network pairs. The first set of distances test the model identification performance when the sizes and edge-densities of the model networks are different. The second set of distances define an easier test, and evaluate the model separation in the case of same size and edge-density model networks. The threshold for the number of allowed false positives for the computation of ROC_n curves are chosen such that the average number of incorrectly clustered networks per query network (EPQ) is 10; i.e., since there are $2,520$ model networks, $n = 2,520 \times 10 = 25,250$. We annotate the AUC_n score computed for 10 errors per query as $AUC_{EPQ=10}$.

With this performance evaluation technique, we first test the effect of removing redundant orbits from the graphlet degree vectors, and the effect of including 5-node graphlets into the network distance measure. In this respect, we systematically compare the model clustering performances of four different GCD variants: (1) GCD–11, computed by using non-redundant 2- to 4-node graphlet orbits (i.e., orbits 0, 1, 2, 4, 5, 6, 7, 8, 9, 10, and 11 in Figure 1.4), (2) GCD–15, computed by using all 2- to 4-node graphlet orbits (i.e., orbits 0–14 in Figure 1.4), (3) GCD–56, computed by using non-redundant 2- to 5-node graphlet orbits (i.e., orbits 0–72 except {3, 5, 7, 14, 16, 17, 20, 21, 23, 26, 28, 38, 44, 47, 69, 71, 72} in Figure 1.4), and (4) GCD–73, computed using all 2- to 5-node graphlet orbits (i.e., orbits 0–72 in Figure 1.4). We compute the ROC and ROC_n curves of these 4 distance measures using the pairwise distances among all pairs of the above described $2,520$ model networks. The resulting curves are presented in Figure 2.9 and the corresponding AUC, $\text{AUC}_{EPQ=10}$, and maximum accuracies are provided in Table 2.2. Note that, no deviation statistics are provided for these experiments, since the experiments are performed on a single set of $2,520$ model networks for which the pairwise GCD-11 distances are illustrated in Figure 2.7–A. In the most general setting, when comparing networks having different network sizes and edge densities, GCDs using redundant 2-to-4 node graphlet orbits (i.e., GCD-15) slightly outperforms its non-redundant counterparts (GCD-11) (Figure 2.9-A and -C, Table 2.2-A). However, when comparing networks with same sizes and edge densities, GCD-11 outperforms all the other GCDs (Figure 2.9-B and -D, Table 2.2-B). GCDs using up to 4-node graphlet orbits (i.e., GCD-11 and GCD-15) have slightly better performance over GCDs using 5-node graphlets (i.e.,

Distance	AUC	Max. Accuracy	$AUC_{EPQ=10}$
GCD-11	0.827	**0.892**	0.164
GCD-15	**0.840**	0.891	**0.200**
GCD-56	0.786	0.883	0.121
GCD-73	0.798	0.883	0.143

(A)

Distance	AUC	Max. Accuracy	$AUC_{EPQ=10}$
GCD-11	**0.997**	**0.978**	**0.945**
GCD-15	0.995	0.971	0.913
GCD-56	0.978	0.950	0.750
GCD-73	0.983	0.952	0.781

(B)

Table 2.2: AUC, Maximum Accuracy and $AUC_{EPQ=10}$ scores showing the model clustering performances of different GCD versions. Table A presents the scores of the experiments that are performed comparing all pairs of the $2,520$ networks independent of their size and edge-density. Similarly, Table B presents the scores obtained by comparing only the same size and edge-density networks.

GCD-56 and GCD-73). This should be due to the fewer orbit dependencies in GCD-11 and GCD-15 compared to the other GCD variants. Surprisingly, the performance of GCD-73 is slightly better than the performance of GCD-56. Since the real-world applications (e.g., finding the model that best fits a real world network, and analysing time series of world trade networks) involves comparing networks having similar node size and edge densities, we focus on the performances of GCD-11 and GCD-73 as representatives of GCDs defined from 2- to 4-node and 2- to 5-node graphlets.

Being encouraged by the observed model separations in Figures 2.7 and 2.8, we systematically compare the model clustering performance of different network distance measures. The six other commonly used, or sensitive

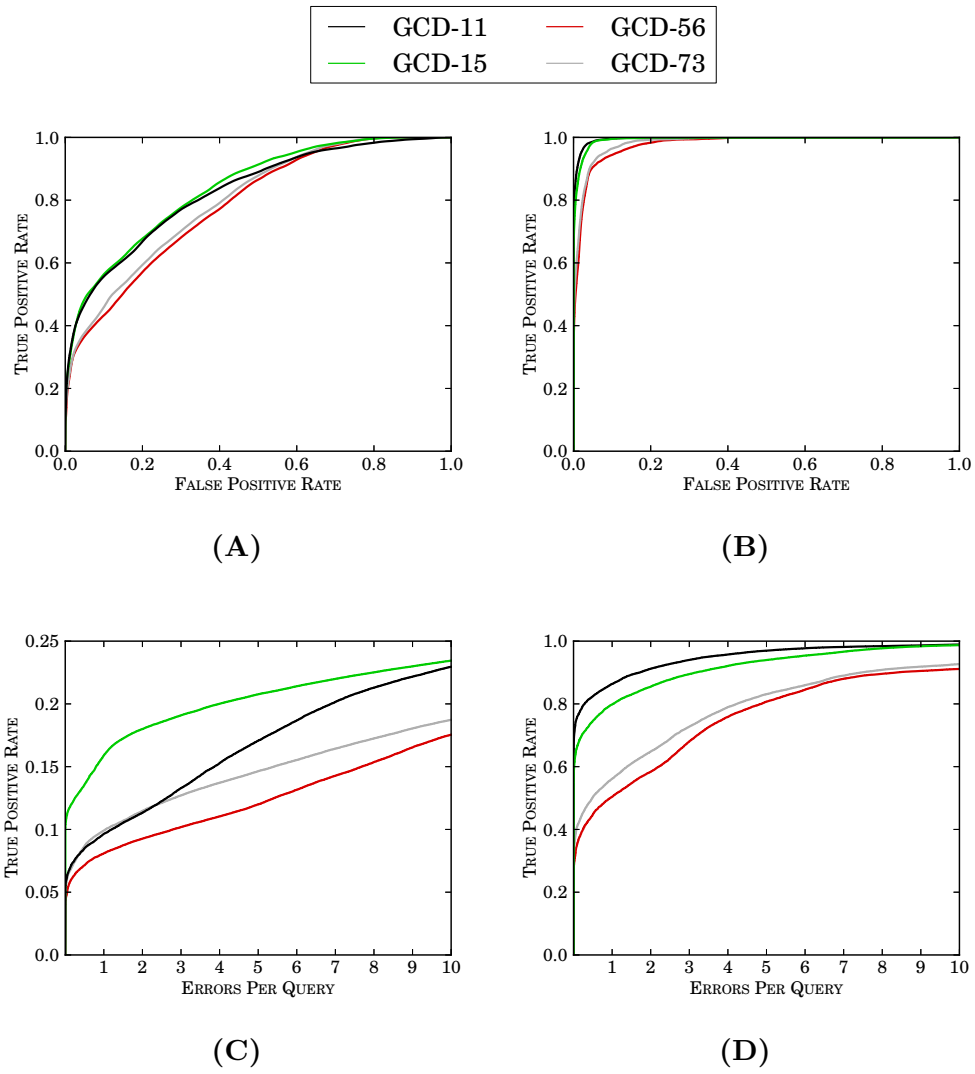


Figure 2.9: Model Clustering Performances of different GCD versions. Panels A and C present the model clustering performance in case of comparing different size and edge-density networks, while B and D present the performance in case of comparing only same size and edge-density networks. Panels A and B illustrates the ROC curves that evaluate the model clustering performances the four GCD versions. Similarly, Panels C and D illustrate the ROC_n curves up to 10 EPQ.

and robust network comparison measures that we use in this comparison are degree distribution [145], clustering coefficient [145], network diameter [145], spectral distance [201], Relative Graphlet Frequency Distribution [157], and Graphlet Degree Distribution Agreement [156] (See Section 1.4 for details about these network distance measures). As in the case of the comparison among different GCD versions, we compute the ROC and ROC_n curves of different network distance measures using the pairwise distances among the above described $2,520$ model networks. Figure 2.10 illustrates the resulting ROC curves and ROC_n curves, and Table 2.3 presents the corresponding AUCs, $\text{AUC}_{EPQ=10}$s, and maximum accuracies. Note that, no deviation statistics are provided for these experiments since the experiments are performed on a single set of $2,520$ networks for which the pairwise GCD-11 distances are illustrated in Figure 2.7–A. Even though ROC curves show slight outperformance of the clustering coefficient and RGF distance over GCD, with GCD-11 being the third best and GCD-73 being the fourth best, the best maximum accuracy and $\text{AUC}_{EPQ=10}$ (i.e., early retrieval) scores are achieved by GCD-11 for the comparison of all networks independent of their size and density (Figures 2.10-A and -C, Table 2.3-A), being followed by GCD-73 as the second best. When the comparison is made for the same size and density networks, GCD-11 outperforms all other network distance measures in all tests (Figure 2.10-B and -D, and Table 2.3-B). GCD-73 competes with RGF distance on being the second best – it is outperformed by RGF Distance in terms of AUC and maximum accuracy scores, but outperforms RGF distance in terms of $\text{AUC}_{EPQ=10}$ score.

Overall, GCD measures outperform all other measures: their ability of early retrieval clearly explains their superiority in clustering networks with similar topologies. Perhaps counter-intuitively, GCD-11 outperforms GCD-73. However, this is easily explained, since orbits in GCD-11 are not only non-redundant, but also "more independent" (since there are fewer of them) than the full set of 73 orbits that comprise GCD-73. Outperformance of GCD-11 is good news, since it is much faster to compute GCD-11 than GCD-73. The worst case time complexity of computing up to 4-node orbits is $O(N^4)$, while it is $O(N^5)$ for computing up to 5-node graphlets, where N is the size of the input network (detailed in Section 2.5). Note that, the computational complexity of graphlet counting is much lower in practice, due to sparsity of the network data. Hence, GCD-11 is a very efficient and

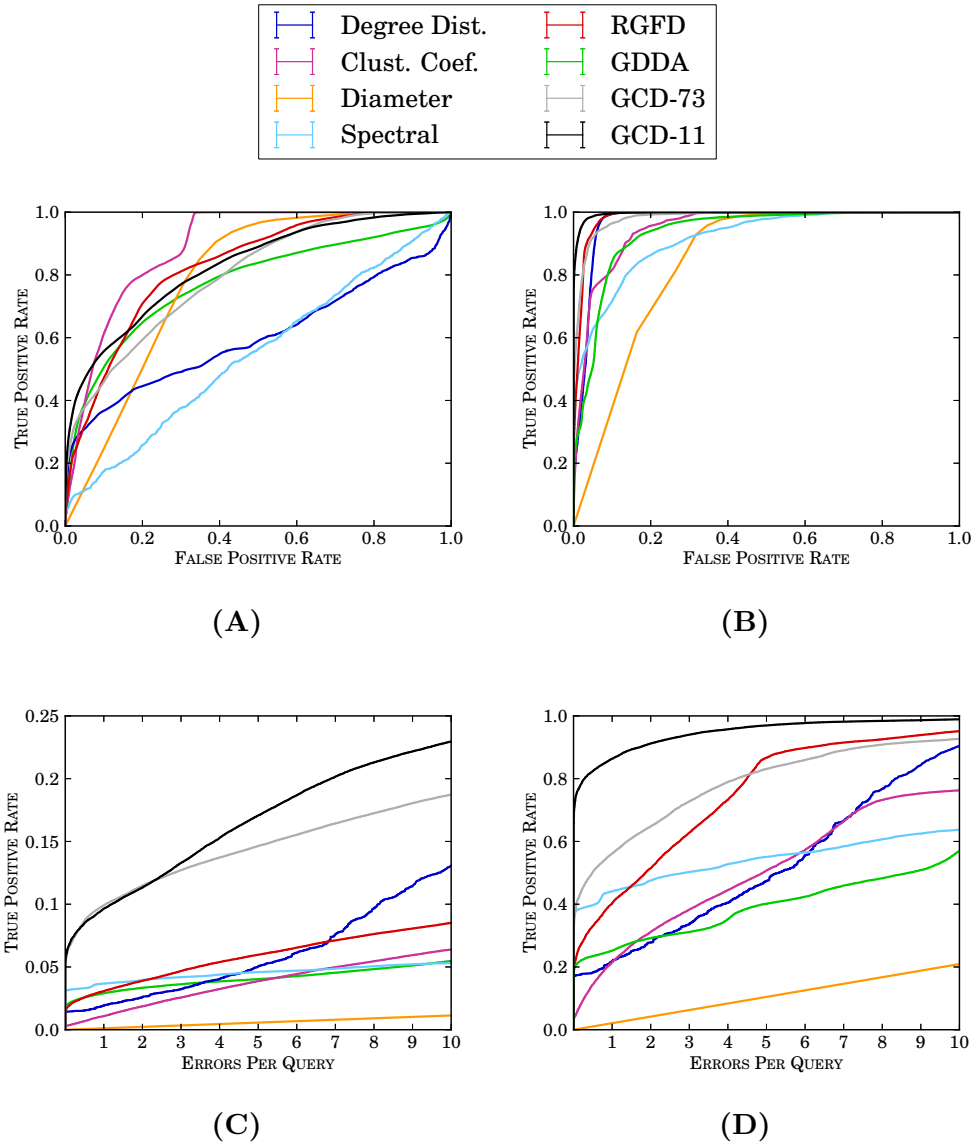


Figure 2.10: Model clustering performances of different network distance measures. All of these tests are performed on the 2,520 model networks . The illustrated curves are: Panel A – ROC curves that are obtained from all network pairs, Panel B – ROC curves that are obtained from same size and density network pairs, Panel C – ROC_n curves up to 10 EPQ that are obtained from all network pairs, and Panel D – ROC_n curves up to 10 EPQ that are obtained from same size and density network pairs.

Distance	AUC	Max. Accuracy	$\text{AUC}_{EPQ=10}$
GCD-11	0.827	0.892	0.164
GCD-73	0.798	0.883	0.143
Spectral Dist.	0.549	0.862	0.045
RGFD	0.829	0.872	0.058
GDDA	0.776	0.877	0.040
Degree Dist.	0.603	0.879	0.058
Clust. Coef.	0.890	0.870	0.032
Diameter	0.788	0.811	0.006

(A)

Distance	AUC	Max. Accuracy	$\text{AUC}_{EPQ=10}$
GCD-11	0.997	0.978	0.945
GCD-73	0.983	0.952	0.781
Spectral Dist.	0.918	0.916	0.538
RGFD	0.985	0.958	0.743
GDDA	0.936	0.898	0.387
Degree Dist.	0.971	0.940	0.508
Clust. Coef.	0.951	0.924	0.479
Diameter	0.796	0.805	0.105

(B)

Table 2.3: AUC, Maximum Accuracy and $\text{AUC}_{EPQ=10}$ scores showing the model clustering performances of different network distance measures. Table A presents the scores of the experiments that are performed comparing all pairs of the $2,520$ networks, independent of their size and edge-density. Similarly, Table B presents the scores obtained by comparing only the same size and edge-density networks.

powerful measure for clustering networks.

2.4.2 Performance on Noisy Networks

Since real networks are noisy and incomplete, we evaluate the clustering quality of the above distance measures in the presence of noise. We randomize each network 30 times for each tested noise type (i.e., false interactions, and missing interactions) and noise rate. If the randomization was performed on the entire set of $2,520$ networks, it would be computationally prohibitive. Hence, we use a subset of 280 out of the $2,520$ networks – for each model, we pick 10 networks from each combination of the following node sizes and edge densities: $\{1000, 2000\}$ nodes and $\{0.5\%, 1\%\}$ edge densities. We use these node sizes and edge densities because these networks are more difficult to cluster than larger networks, so if we can show the methodology to be robust under more stringent conditions, we can be confident that it will perform well on real-world problems.

We test the performance of different network distance measures on noisy networks that contain false interactions by randomly rewiring a percentage of edges. For a network that has $|E|$ edges, a "k% noisy network" is generated as follows: at each step, three nodes, a, b, c, are chosen such that, the edge (a, b) is in the network but not the edge (a, c). The edge (a, b) is rewired by removing it from the network and adding edge (a, c) into the network. This process is repeated $|(|E| \times k)/100|$ times, producing the noisy network that contains false interactions. We randomize each of the 280 model networks by rewiring $k\%$ of their edges. This results in 280 noisy model networks. We evaluate the clustering performance of a network distance measure on this set of noisy networks by measuring the early identification performance using $\text{AUC}_{EPQ=10}$. We perform these tests on the distances obtained by: (1) comparing all pairs of the 280 random networks (i.e., on $\binom{280}{2} = 39,600$ network pairs), and (2) comparing network pairs that are of the same size and edge density (i.e., on $2 \times 2 \times \binom{7\times10}{2} = 9,660$ network pairs). We repeat these tests 30 times for each noise level, k, and report the averages and standard deviations of the 30 experiments. Note that performing these tests amounts to a large number of computations, since for each of the 9 noise levels (in increments of 10%), we have $30 \times 280 = 8,400$ networks to count graphlets for. That is, we count graphlets for $9 \times 8,400 = 75,600$

networks, which takes a long time even if done in parallel on a decent computing cluster. Figure 2.11 summarizes the results of these experiments for both settings. When the network pairs of different sizes and edge densities are compared together, if we randomly rewire up to 80% of edges in the model networks, $\text{AUC}_{EPQ=10}$ shows that GCD-11 has the best early identification performance over all tested distance measures. Similarly, when the comparison is made only between same size and edge density network pairs, $\text{AUC}_{EPQ=10}$ shows that GCD-11 has the best early identification performance over all tested distance measures for all noise rates. Note that, the $\text{AUC}_{EPQ=10}$ scores on the vertical axis of Figure 2.11 are not the same as those in Figure 2.9 (Table 2.3), since they correspond to the 280 networks described above, while those in Figure 2.9 (Table 2.3) correspond to the full set of $2,520$ networks.

Apart from containing false interactions, many real-world networks are incomplete; i.e., they have missing interactions. For evaluating the performance of network distance measures on incomplete networks, we sample $k\%$ of edges from a model network and make a subgraph induced on the sampled edges. We do this sampling for each of the 280 above described model networks, for sampling rates of $\{10\%, 20\%, 30\%, \ldots, 90\%\}$. We repeat this random tests 30 times per sampling rate, resulting in $280 \times 9 \times 30 = 75,600$ networks to count graphlets for. We evaluate the early retrieval performances of different network distances based on the averages and standard deviations of obtained $\text{AUC}_{EPQ=10}$ scores. In addition, for testing the clustering performance of the distance measures for both noisy and incomplete data [71, 184], we perform the same edge sampling experiments, but this time using the 280 networks with 40% rewired edges. So in total, we count graphlets for $2 \times 75,600 = 151,200$ networks. To test the model identification performance separately for the comparison of different size networks and same size networks, we perform the experiments using two different sets of distances: (1) the distances obtained by comparing all pairs of 280 networks — $\binom{280}{2} = 39,060$ network pairs, and (2) the distances obtained from the comparison of only same size and density networks from the 280 networks — $\binom{10\times7}{2} \times 4 = 9,660$ network pairs. Figure 2.12 illustrates the results of the experiments in these 4 settings. Similar to the results obtained for rewired noisy networks, GCD-11 outperforms all other network distance measures for all the settings, up to 20% edge completeness (i.e., 80% missing

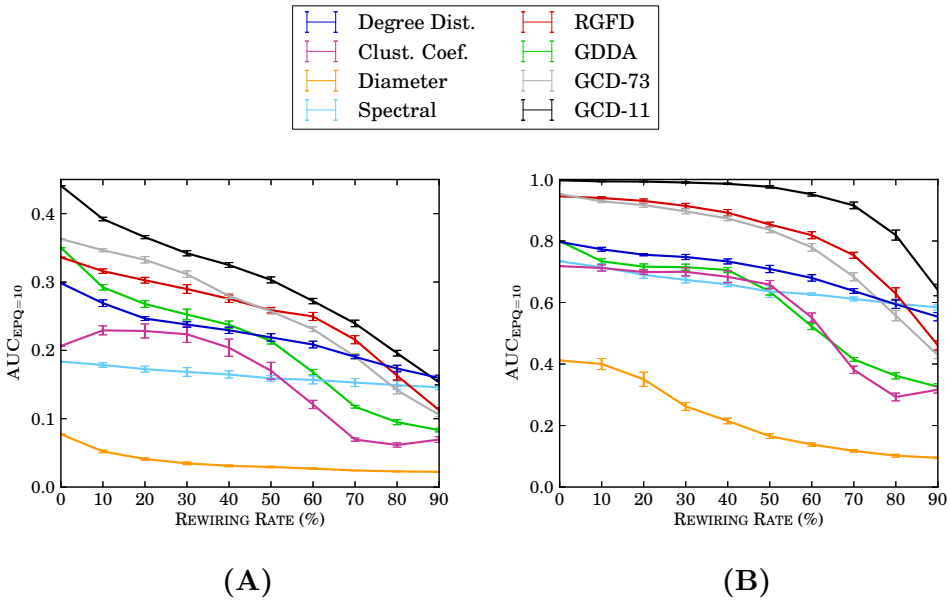


Figure 2.11: Effects of rewiring model networks on $\text{AUC}_{EPQ=10}$ scores of different network distance measures. The vertical axis represents the average $\text{AUC}_{EPQ=10}$ scores and their standard deviations for the 30 randomized experiments that are performed at each of the noise levels that are presented by the horizontal axis independently. Panel A – the ROC_n scores obtained by comparing all pairs of the 280 networks. Panel B – the ROC_n scores obtained by comparing only same size and density networks.

edges). Therefore, we conclude that GCD-11 is the best distance measure for model clustering and it is highly noise-tolerant for both false positive interactions and missing interactions.

2.4.3 Performance with Sampled Network Properties

We test the model identification performance of the network distance measures when only partial information about the network is available; i.e., when the network distance measures are computed from the properties of $k\%$ of the nodes. The distances are computed from node properties of the $k\%$ of the nodes as follows:

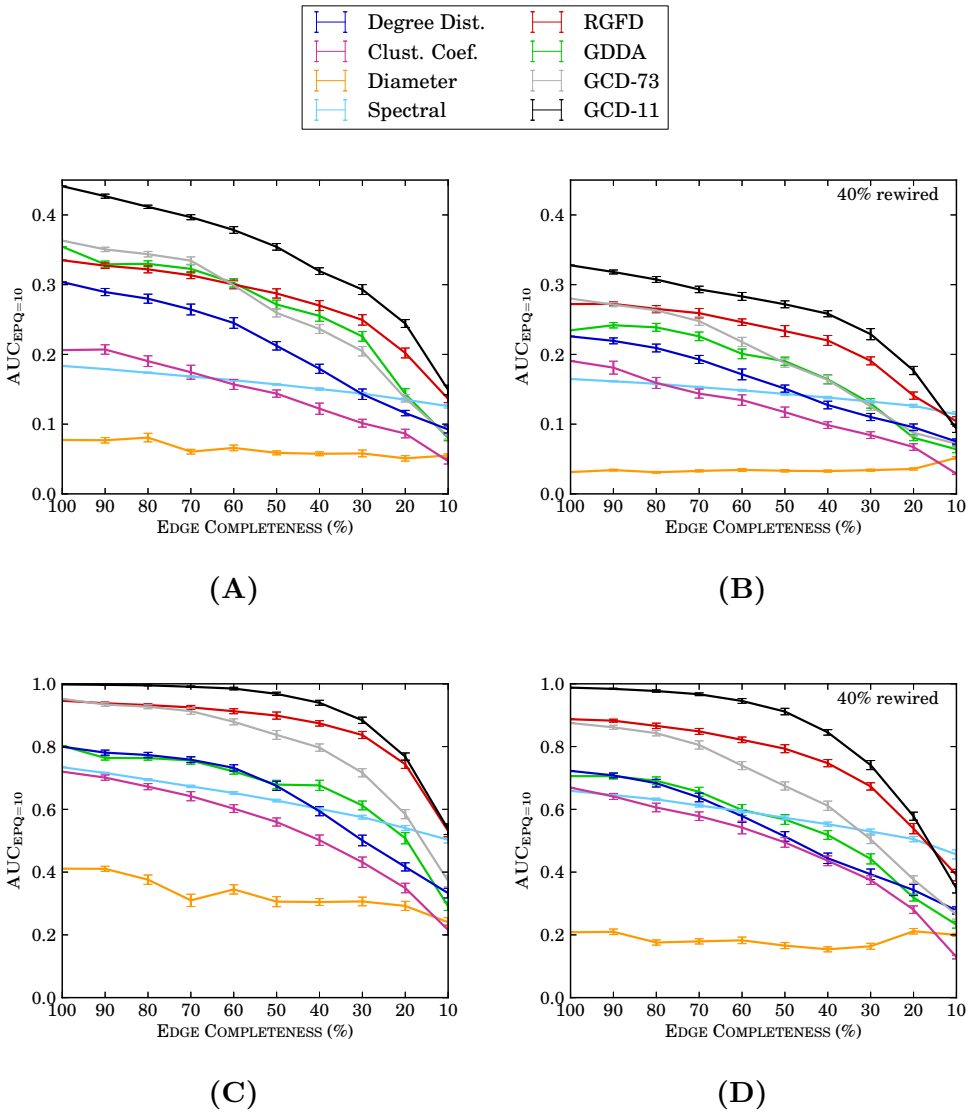


Figure 2.12: Model clustering performance comparison on incomplete networks. These experiments are performed on the reduced set of 280 model networks. The illustrated statistics are: Panel A – AUC$_{EPQ=10}$ scores for the comparison of all network pairs in the case of missing edges, Panel B – AUC$_{EPQ=10}$ scores for the comparison of all network pairs in the case of both missing and 40% rewired edges. Panel C – AUC$_{EPQ=10}$ scores for the comparison of only same size and density network pairs in the case of missing edges, Panel D – AUC$_{EPQ=10}$ scores for the comparison of only same size and density network pairs in the case of both missing and 40% rewired edges.

- **Degree Distribution:** We randomly choose k% of the nodes, and compute their degrees using the entire network. We then use the distributions defined by these degrees for the comparison of two networks.

- **Clustering Coefficient:** We randomly choose k% of the nodes, and compute their clustering coefficients using the entire network. We then average these clustering coefficients to obtain the clustering coefficient of the network.

- **Diameter:** We randomly choose k% of the nodes of a network and compute their eccentricities in the entire network. We choose the largest eccentricity over the $k\%$ sampled nodes and that is the diameter of the network.

- **Spectral Distance:** We compute the Laplacian matrix of the complete network, randomly choose k% of the nodes, and compute the spectrum from the submatrix formed by the rows and columns of the Laplacian matrix corresponding to these nodes.

- **Relative Graphlet Frequency Distance (RGFD):** We randomly choose k% of the nodes, and compute the graphlet degree vectors (GDVs) of these nodes using the entire network. We derive the average number of graphlets from the sampled graphlet degree vectors, by summing up all the graphlet degrees of an orbit corresponding to the graphlet and normalizing this sum by dividing to the number of nodes in the graphlet that correspond to the chosen automorphism orbit. We use the derived graphlet counts to compute the RGFD as explained in Section 1.4.

- **Graphlet Degree Distribution Agreement (GDDA):** As for RGFD, we randomly choose $k\%$ of the nodes, for which graphlet degree vectors (GDVs) are computed using the entire network. Then, Graphlet Degree Distributions (GDDs) are computed over these GDVs, and GDDA is computed by comparing these distributions.

- **Graphlet Correlation Distance (GCD):** As for RGFD and GDDA, we randomly choose $k\%$ of a network's nodes and compute GDVs for each of these nodes using the entire network. Then, GCM is computed

from the GDVs of the selected nodes and GCDs are computed as the Euclidean distances between the obtained GCMs.

We sample $\{10\%, 20\%, 30\%, \ldots, 90\%\}$ of the nodes from each of the 280 model networks, and compare the "early identification" performances of the distances computed from the sampled network properties using $AUC_{EPQ=10}$ scores. We repeat these experiments 30 times for each sampling rate, and present the averages and standard deviations of the obtained $AUC_{EPQ=10}$ scores. In addition, for testing the early identification performances of the sampled network distances on noisy networks, we repeat the same experiments, but this time using the 280 model networks that contain 40% rewired edges. We assess the "early identification" performance of the distance measures by: (1) comparing pairs of 280 model networks — from $\binom{280}{2} = 39,060$ network pairs, and (2) comparing only the same size and density networks — from $\binom{10\times 7}{2} \times 4 = 9,660$ network pairs. The obtained $AUC_{EPQ=10}$ results of the experiments for these 4 settings are provided in Figure 2.13. These results show that a surprising speed up in computational time of the GCD-11 can be obtained without loss in the clustering quality: by taking GDVs of as few as 30% of randomly chosen nodes in a network to form its GCM-11, $AUC_{EPQ=10}$ of GCD-11 only slightly decreases and it outperforms all other measures in all experimental settings. In addition, for the noisy networks described above, the clustering obtained by GCD-11 again outperforms those obtained by all other measures and does not deteriorate even if we randomly sample as few as 30% of GDVs to form GCD-11.

2.5 Computational Complexities of Network Distance Measures

The graphlet-based network comparison methods (i.e., RGF Distance, GDD Agreement, Graphlet Correlation Distance) are computationally more expensive than other network distance measures due to the necessity of induced subgraph identification. However, the high computational complexity of graphlet-based methods are worth the cost since they obviously have better model identification and clustering performance than other standard network distance measures as shown in Section 2.4.

Given a network $G(V, E)$ with $|V|$ nodes and $|E|$ edges, the worst-case

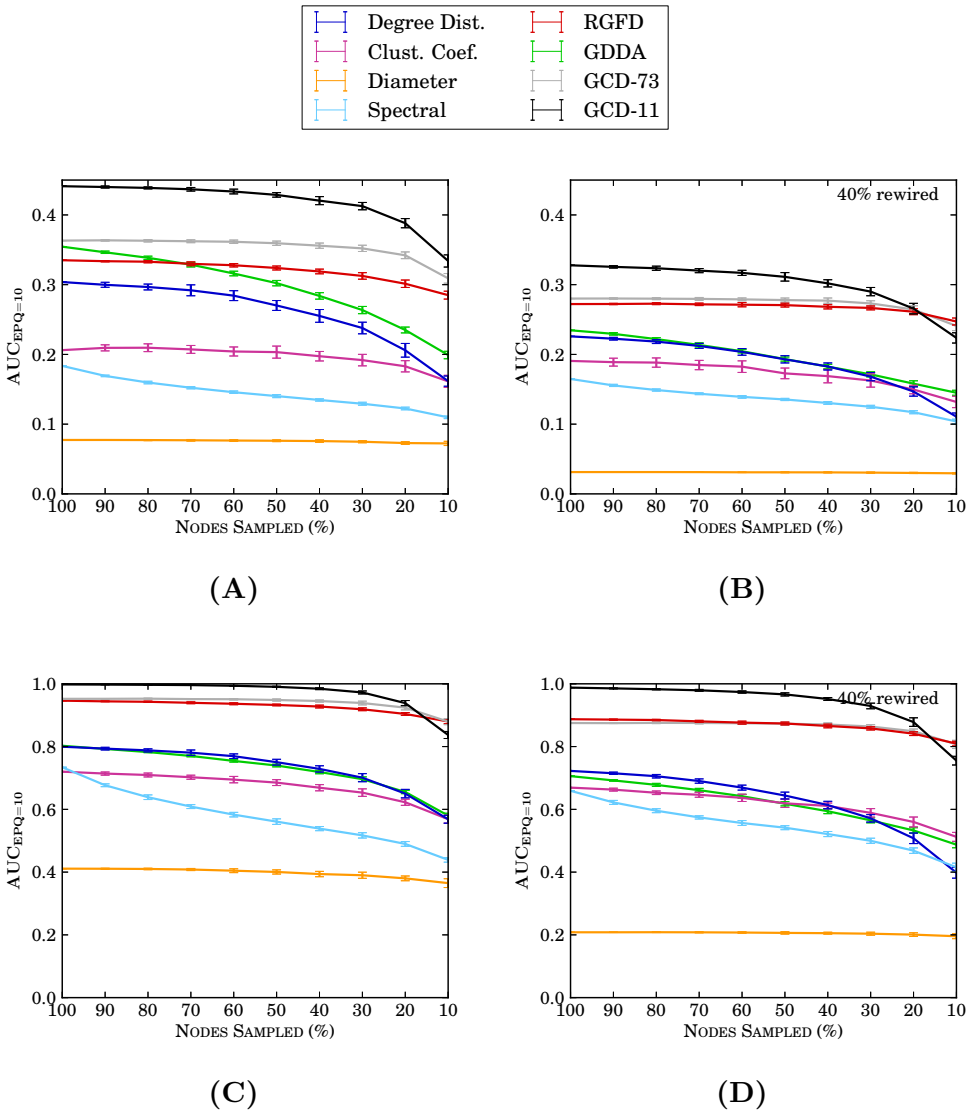


Figure 2.13: Model clustering performance comparison based on sampled network properties. These experiments are performed on the reduced set of 280 model networks. The illustrated statistics are: Panel A – AUC$_{EPQ=10}$ scores for sampling based comparison of all network pairs, Panel B – AUC$_{EPQ=10}$ scores for sampling based comparison of all network pairs in the existence of false interactions (40% rewired edges). Panel C – AUC$_{EPQ=10}$ scores for sampling based comparison of only same size and density network pairs, Panel D – AUC$_{EPQ=10}$ scores for sampling based comparison of only same size and density network pairs in the existence of false interactions (40% rewired edges).

computational complexity of degree-based properties are $O(|E|) = O(|V|^2)$, since identifying the degrees of all nodes requires a single pass over all edges in the network. For the clustering coefficient related network properties, the worst-case complexity is $O(|V|^3)$ as the links between each pair of a node's neighbours need to be checked. For the diameter and other shortest-path related properties, the worst-case complexity is $O(|V|^3)$ since all-pairs-shortest-paths problem can be best solved by the Floyd-Warshall algorithm [145]. Similarly, centrality measures are also bound to the problem of all pairs-shortest paths and their worst-case performance is bounded by $O(|V|^3)$. The spectral distance between two networks is strictly dependent on the eigenvalue decomposition of the $|V| \times |V|$ network matrix, which is $O(|V|^3)$.

The graphlet-based network distance measures (i.e., RGFD, GDDA, and GCD) requires counting the number of graphlets / graphlet degrees in the network. For a network with $|V|$ nodes, the worst-case complexity for counting all graphlets and graphlet degrees for 2- to k-node graphlets is $\mathrm{O}(|V|^k)$ and a tighter upper-bound is $\mathrm{O}(|V|d^{k-1})$, where $d \leq |V|$ is the maximum degree over all nodes in the network. In RGFD, computing the differences between the number of graphlets is done in O(1) time. In GDDA, computing the differences between the normalized distributions of graphlet degrees is done in $\mathrm{O}(|V|)$ time, since each graphlet degree distribution contains up to $|V|$ distinct values. The arithmetic average of these differences is then computed in O(1) time. For GCD, computing the Spearman's correlation coefficients between the orbits over $|V|$ nodes is done in $\mathrm{O}(|V| \ln(|V|))$ time, and the Euclidean distance between two GCMs is computed in O(1) time. Hence, the time complexities of graphlet-based distance measures are dominated by the complexity of counting graphlets. However, since GCD performs better when it uses up to 4-node graphlets rather than up to 5-node graphlets, it reduces the time complexity of GCD-based network comparison from $\mathrm{O}(nd^4)$ to $\mathrm{O}(nd^3)$. This is a big improvement for large networks. For example, for Facebook network of Berkeley University (which contains 22,937 nodes and 852,444 edges), counting all graphlets/graphlet degrees for up to 5-node graphlets takes $\sim$ 4 days, while it takes only $\sim$ 5 hours to count all of its up to 4-node node graphlets/graphlet degrees. This performance improvement makes GCD-based analysis feasible even for large networks.

2.6 Author's Contributions

Ömer Nebil Yaveroğlu collaborated with Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj for the work presented in this chapter.

In the identification of the graphlet degree vector redundancies and dependencies, Ömer Nebil Yaveroğlu designed and implemented the method for automatic retrieval of all graphlet degree vector redundancy equations, after being initiated by a previous discussion among Dr. Nataša Pržulj's former group's members.

In graphlet correlations related work, Ömer Nebil Yaveroğlu took part in the discussions on graphlet correlations, suggested quantifying the topological distance between two networks using graphlet correlation matrices, implemented and tested all the ideas generated from these discussions on graphlet correlations, and documented the corresponding results.

3 Analysis & Comparison of World Trade Networks

In this chapter, we analyse the dynamic system of world trade networks based on our new graphlet correlation based methodology. In particular, we first analyse the topology of world trade networks based on their graphlet correlation matrix patterns, track the changes in the topology of world trade networks, link these changes with the changes in crude oil price and link the economic wealth of a country with its network position. Furthermore, we propose two graphlet based scores that evaluate the strength of brokerage and peripheral position of a country, and analyse the effects of a country's network position on its crisis patterns based on these scores.

3.1 Motivation

The world economy has never been a stable and easy-to-predict system as it is composed of many components that affect each other with their individual actions. The recent global recession has once again shown that a local malfunctioning in these economic components may have uncontrollable consequences on the world economy on a global scale. Insights into the functioning of the world economy can be mined from the flow of money between countries, which is woven into their trade relations. Therefore, studies on world trade networks are recently gaining more and more attention [101, 170]. Due to the increasing interest in understanding the world trade relations, we first obtain the world trade networks from 1962 to 2010 from UN Comtrade database [34] (construction of the world trade networks is explained in Section 1.2.1). We analyse the obtained networks using our new graphlet correlations based methods (Chapter 2), aiming to: (1) gain insights into the organisational principles of the world trade network, (2) track the changes on the world trade network topology over the years and

relate these changes with economic changes in the world through the crude oil prices, (3) develop models of the world trade network and evaluate their fit on the observed world trade networks, and (4) analyse the effects of a country's network position on its wealth.

In order to relate our topological analysis of world trade networks with the economic facts, we need the statistics on some external economic wealth indicators. We obtain the crude oil prices for all years between 1962 and 2010 from UNCTADSTAT Reports [144] (downloaded in November 2012). As the crude oil prices in this dataset are provided on a monthly basis (and our world trade network data is on a yearly basis), we compute the crude oil price of a year as the average price of the corresponding 12 months. We obtain the economic wealth indicator statistics of countries from PENN World Table (PENN) [80] (version 7.1; downloaded in November 2011) and International Monetary Fund World Economic Outlook Database (WEO) [60] (downloaded in October 2012). All prices in these statistics were expressed in 2005 US Dollars. The list of economic wealth indicators and their definitions are as follows:

- **Gross Domestic Product - version 1 (RGDPL):** Purchasing Power Parity converted Gross Domestic Product Per Capita (Laspeyres), derived from the growth rates of consumption share, government consumption share, and investment share. This data is from PENN.
- **Gross Domestic Product - version 2 (RGDPL2):** Purchasing Power Parity converted Gross Domestic Product Per Capita (Laspeyres), derived from growth rates of domestic absorption. This data is from PENN.
- **Gross Domestic Product - version 3 (RGDPCH):** Purchasing Power Parity converted Gross Domestic Product Per Capita (Chain Series). This data is from PENN.
- **Consumption Share (KC):** Consumption Share of Purchasing Power Parity Converted Gross Domestic Product Per Capita at 2005 constant prices (RGDPL). This data is from PENN.
- **Government Consumption Share (KG):** Government Consumption Share of Purchasing Power Parity Converted Gross Domestic

Product Per Capita at 2005 constant prices (RGDPL). This data is from PENN.

- **Investment Share (KI):** Investment Share of Purchasing Power Parity Converted Gross Domestic Product Per Capita at 2005 constant prices (RGDPL). This data is from PENN.
- **Openness (OPENK):** Trade openness as a percent of 2005 constant prices. This data is from PENN.
- **Population (POP):** The total population of the country. This data is from WEO.
- **Level of Employment (LE):** The number of people who, during a specified brief period such as one week or one day, (a) performed some work for wage or salary in cash or in kind, (b) had a formal attachment to their job but were temporarily not at work during the reference period, (c) performed some work for profit or family gain in cash or in kind, (d) were with an enterprise such as a business, farm or service but who were temporarily not at work during the reference period for any specific reason. This data is from WEO.
- **Current Account Balance (BCA):** Current account is all transactions other than those in financial and capital items. The major classifications are goods and services, income and current transfers. The focus of the BCA is on transactions (between an economy and the rest of the world) in goods, services, and income. This data is from WEO.

KC, KI and KG are expressed in percentage of GDP per capita. We included copies of these indicators, converted into constant price per capita, i.e., multiplied by GDP per capita (e.g., KC $\times$ RGDPL). We also included copies of the indicators expressed in constant price per capita (also including RGDPL, RGDPL2, RGDPCH) but converted into raw constant price value – these are multiplied by the population (e.g., RGDPL $\times$ POP).

3.2 Topology of World Trade Networks

We first explore the topology of the world trade networks using their graphlet correlation matrices (explained in Section 2.3.1), which are constructed from all 2- to 4-node graphlet orbits; i.e., orbits 0-14 (Figure 1.4). After computing the graphlet correlation matrices of all total world trade networks, we cluster the orbits based on the similarities of their pairwise correlation patterns using single linkage clustering. We use the identified clusters to reorder the graphlet correlation matrices so as to highlight the similar correlation patterns among orbits. The orbit clusters that are obtained from the world trade networks of different years are mostly consistent. For this reason, we plotted the graphlet correlation matrices of all world trade networks based on the orbit order obtained for the world trade network of 1962.

Figure 3.1 represents the graphlet correlation matrices of the total world trade network for the networks of 1962, 1970, 1980, 1990, 2000, and 2010. As illustrated, the graphlet correlation matrices of world trade networks are more or less similar over the years. There are two consistent, strongly clustered orbit groups over all years: (1) group of orbits $\{2, 5, 11, 13\}$ – which correspond to broker positions (i.e., mediators between unconnected nodes) in 2- to 4-node graphlets, (2) group of orbits $\{0, 3, 10, 14\}$ – which correspond to densely connected positions (i.e., positioned over triangular patterns) in 2- to 4-node graphlets. Orbits 7 and 8, which also represent broker positions, cluster well with the group of broker orbits, but they still show slightly different patterns from the main broker group. The broker and densely connected orbit groups also positively correlate, even though their inter-group correlation is smaller than their intra-group correlations. This indicates that a broker country is also located in a densely connected region of the network, but not all densely connected nodes are located as brokers in the network.

Peripheral orbits (i.e., orbits 1, 4, 6, 9, 12) are not as strongly correlated as the two other orbit groups; i.e., broker orbits and densely connected orbits. However, their correlations with the two other orbit groups show similar patterns. The correlations among the peripheral orbits and the remaining orbits are very low, meaning that a node can be located in either a peripheral position or a densely connected / broker position, but not both.

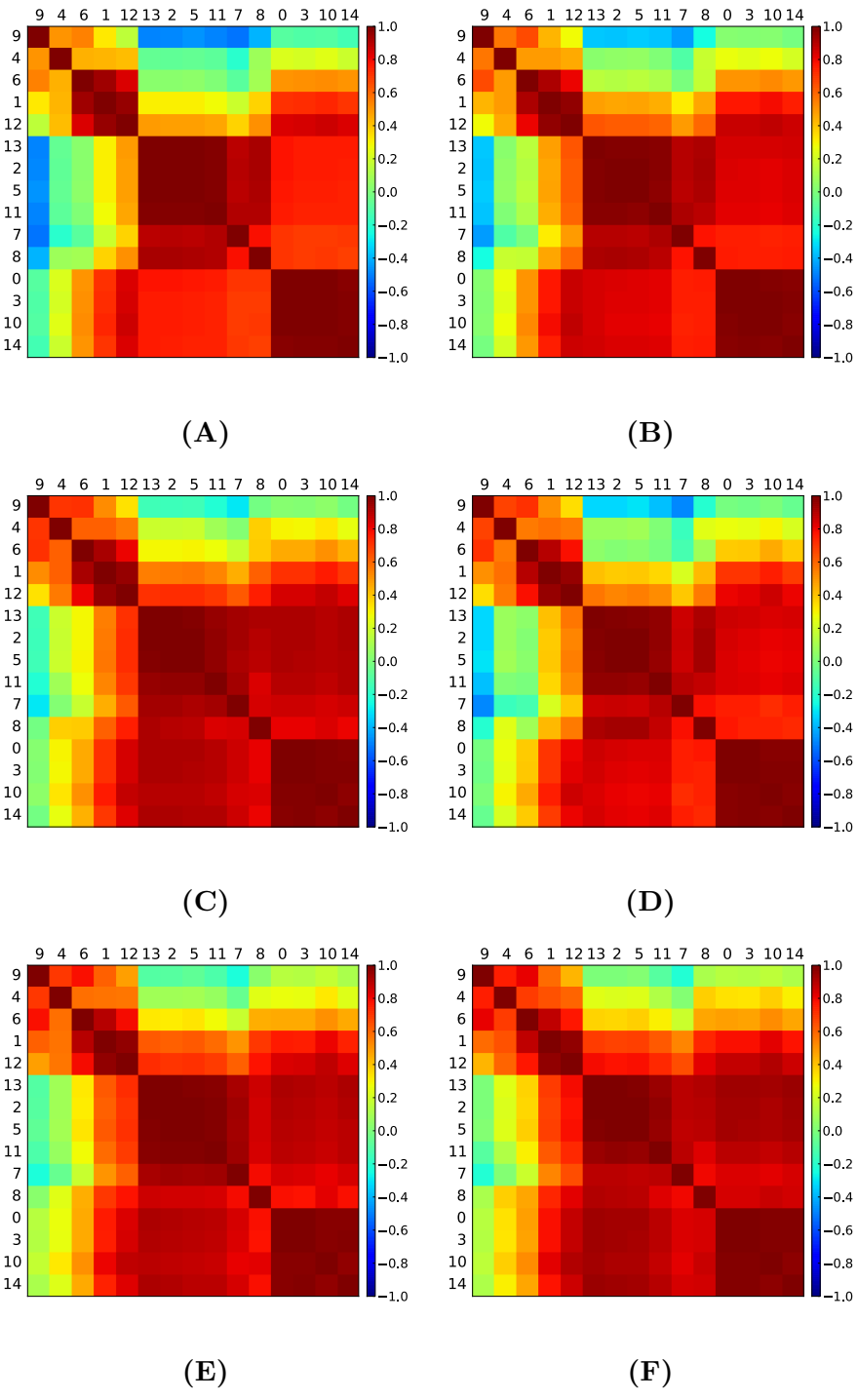


Figure 3.1: Graphlet correlation matrices of the world trade network at different years: Panel A – 1962, Panel B – 1970, Panel C – 1980, Panel D – 1990, Panel E – 2000, and Panel F – 2010.

He et al. [77] show that globalization reduces the hierarchical organisation in world trade network. Parallel to this observation, we trace the change on graphlet correlation matrices over time. We observe that the strong clustering pattern for broker orbits is gradually decreasing. Broker orbits are becoming more strongly correlated with the densely connected orbits, reducing the separation among these orbit sets. Furthermore, the correlations peripheral orbits and the two groups of orbits also increase. Nevertheless, even though the clustering patterns among the different orbit groups become less observable, the orbit groups are still clearly observable in the world trade network of 2010.

3.3 Effect of Crude Oil Price Changes on the World Trade Network

Crude oil (petrol) price is an indicator of global recessions in the world [41]. Most of the sudden changes in crude oil prices are associated with global recessions. To gain insights into the effects of economic crises on the topology of the world trade network, we analyse the topological changes in the world trade networks using GCD-11 from 1962 to 2010, and relate the identified topological changes with the changes in crude oil price.

We quantify the change in world trade network topology using our new network distance measure, GCD-11 (explained in Chapter 2); e.g., the topological change on 1990 is equal to the graphlet correlation distance between the networks of 1989 and 1990. GCD is an unsigned network distance measure by which only the amount of topological change is measured without any indication of a change direction. For this reason, we quantify the change in crude oil price by the absolute difference of crude oil prices for consecutive years; e.g., the change for 1990 is equal to $|Price_{1990} - Price_{1989}|$. We obtain two change distributions by computing the differences among all consecutive years over the period of 1962–2010: (1) the distribution of changes in crude oil price, and (2) the distribution of changes (measured by GCD) in world trade network topology. We test the relatedness of these two distributions using two different correlation measures: (1) Spearman's Correlation Coefficient, and (2) Phi Correlation Coefficient. The Spearman's Correlation Coefficient takes the size of the changes into account, while the

Phi correlation coefficient only evaluates the similarities of the upward and downward trends.

With these correlation tests, we aim to uncover possible effects of crude oil price and the world trade network topology on each other; i.e., whether the change in crude oil price follows the topological change in world trade network and vice versa. In order to test this, we shift the two time-series distributions forward and backward over each other by up to 3 years in time, and compute the corresponding correlations. Negative year shifts test the effects of topological changes in the world trade network on crude oil price, and positive shifts reflect the effects of the changes in crude oil price on the topology of the world trade network. On the other hand, comparing the changes on yearly basis may cause fluctuations in the two change distributions, hiding their general patterns and making the comparisons error-prone. In order to smooth the change distributions and cope with the yearly data variability, we apply a simple low-pass filter on the change distributions. In this respect, we compute the change distributions by performing comparisons in blocks of years. For a year, y, and block size of n, the change score is the arithmetic average of all pairwise comparisons among years blocks $\{p-(n-1),\ldots,p\}$ and $\{p+1,\ldots,p+n\}$. For example, the crude oil price change for 1990, $Change_{1990}$, is equal to:

$$Change_{1990} = \frac{1}{4} \quad (|Price_{1992} - Price_{1989}| + |Price_{1992} - Price_{1990}| + |Price_{1991} - Price_{1989}| + |Price_{1991} - Price_{1990}|),$$

when computed for a block size of 2. We test the trade networks from all 11 commodities for each block size in $\{1,2,3\}$ and year shifts of $\{-3,-2,-1,0,1,2,3\}$, resulting with $7 \times 3 \times 11 = 231$ tests. We apply Holm-Bonferroni correction on the p-values of the obtained correlations for correcting the bias of multiple hypothesis testing [87].

Table 3.1 lists all the significant (adjusted p-values < 0.05) positive correlations between the change distributions of crude oil price and network topology. Figure 3.2 illustrates the distributions of the changes in crude oil price and network topology that have significant Spearman's correlations. Similarly, Figure 3.3 illustrates the distributions of the changes in crude oil price and network topology that have significant Phi correlations.

We find that the changes in crude oil price are correlated with the changes

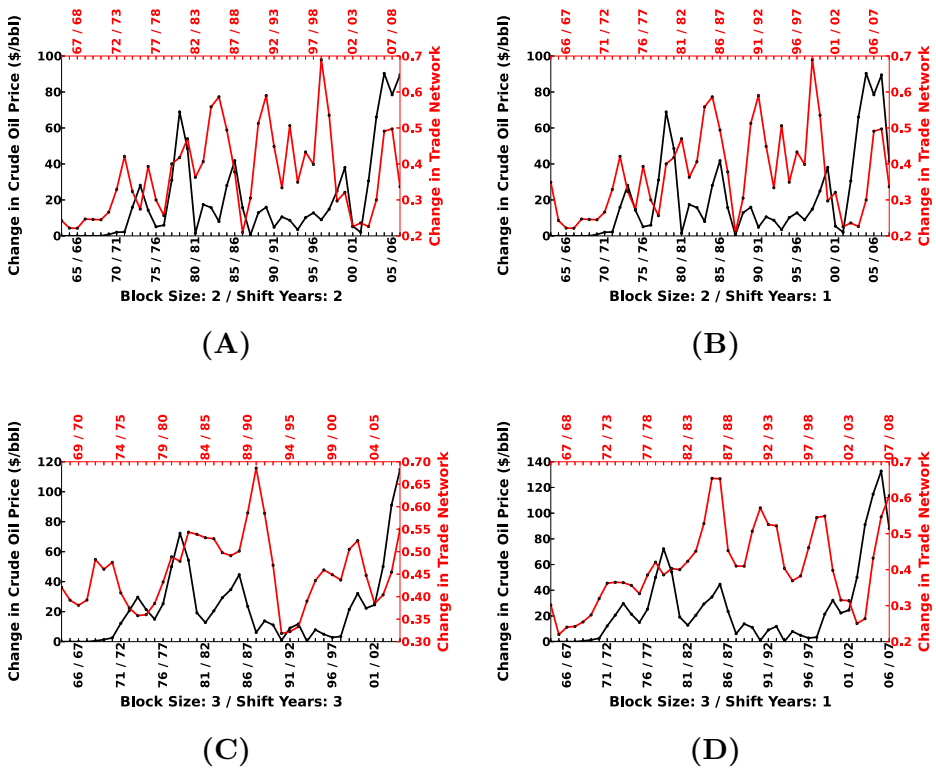


Figure 3.2: The crude oil price and network topology changes that are identified to be significantly correlated (adjusted $p - value < 0.05$) using Spearman's Correlation Coefficient (ordered by decreasing correlations). The topological trade network change patterns that are presented in the figures are: Panel A – "Total" Trade network with year shift = 2 and block size = 2 (corr. = 0.414; p-value = 0.005), Panel B – "Total" Trade network with year shift = 1 and block-size = 2 (corr. = 0.356; p-value = 0.016), Panel C – "Misc. Manufactured" network with year-shift = 3 and block-size = 3 (corr. = 0.347; p-value = 0.026), Panel D – "Total" trade network with year-shift = 1 and block-size = 3 (corr. = 0.316; p-value = 0.039).

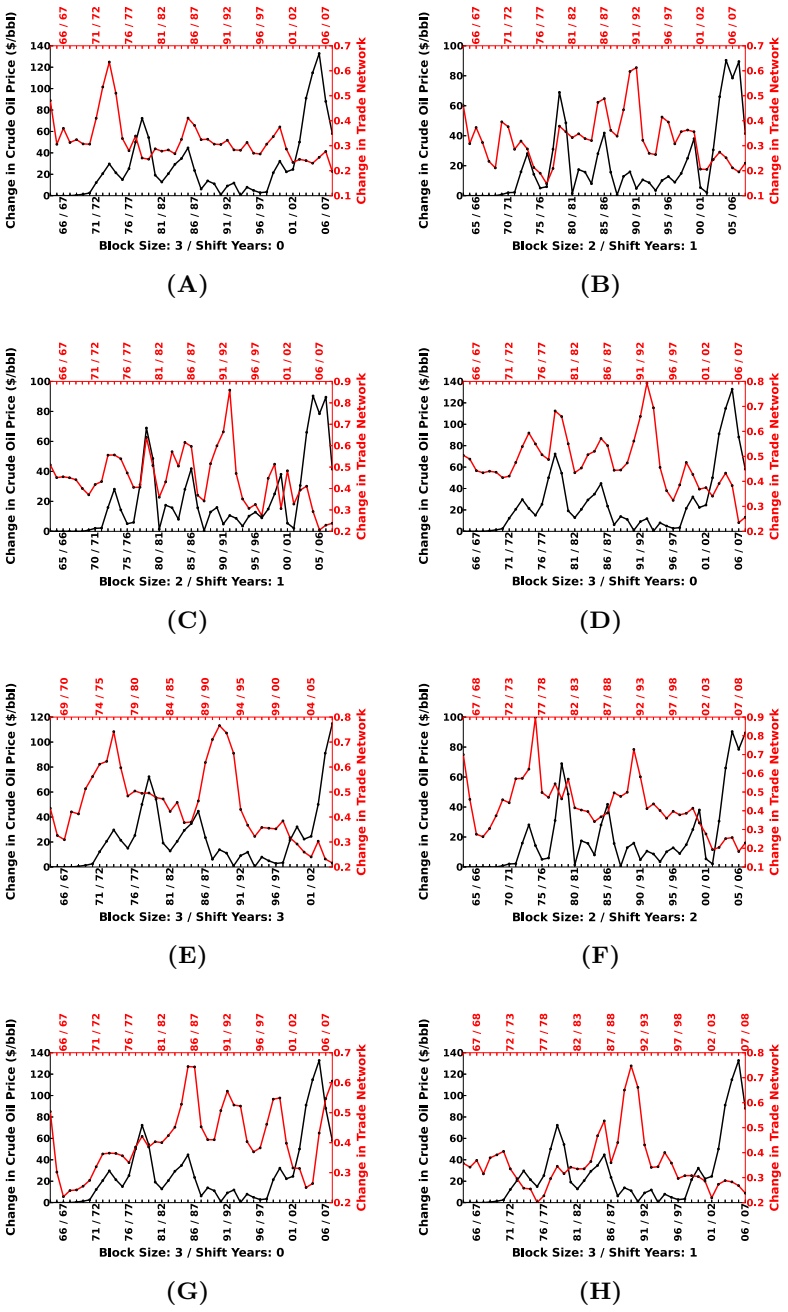


Figure 3.3: The crude oil price vs. network topology changes that are identified to be significantly correlated (adjusted $p - value < 0.05$) using Phi Correlation (ordered by decreasing correlations): A – "Food and Live Animals" network with year-shift = 0 and block-size = 3 (corr = 0.479; p-value = 0.001), B – "Crude Material (except Fuel)" network with year-shift = 1 and block-size = 2 (corr = 0.468; p-value = 0.001), C – "Chemicals" network with year-shift = 1 and block-size = 2 (corr = 0.465; p-value = 0.001), D – "Chemicals" network with year-shift = 0 and block-size = 3 (corr = 0.403; p-value = 0.007), E – "Mineral Fuels" network with year-shift = 3 and block-size = 3 (corr = 0.402; p-value = 0.001), F – "Mineral Fuels network with year-shift = 2 and block-size = 2 (corr = 0.399; p-value = 0.001), G – "Total" trade network with year-shift = 1 and block-size = 2 (corr = 0.371; p-value = 0.001), H – "Crude Material (except Fuel)" network with year-shift = 1 and block-size = 2 (corr = 0.334; p-value = 0.001).

Commodity	Block Size	Shift Years	Corr. / P-value (Spearman)	Corr. / P-value (Phi Coef.)
TOTAL	2	2	**0.414 / 0.005**	-0.055 / 0.725
TOTAL	2	1	0.356 / 0.016	-0.025 / 0.875
MISC. MANUFACTURED	3	3	0.347 / 0.026	0.012 / 0.940
TOTAL	3	1	0.316 / 0.039	0.089 / 0.575
FOOD & LIVE ANIMALS	3	0	-0.321 / 0.033	**0.479 / 0.001**
CRUDE MAT. (exc. FUEL)	2	1	-0.022 / 0.885	0.468 / 0.001
CHEMICALS	2	1	-0.021 / 0.893	0.465 / 0.001
CHEMICALS	3	0	-0.084 / 0.589	0.403 / 0.007
MINERAL FUELS	3	3	-0.087 / 0.588	0.402 / 0.010
MINERAL FUELS	2	2	-0.114 / 0.461	0.399 / 0.008
TOTAL	3	0	0.212 / 0.166	0.371 / 0.014
CRUDE MAT. (exc. FUEL)	3	1	-0.469 / 0.001	0.334 / 0.031

Table 3.1: All significantly correlated changes in Crude Oil Price and Trade Network Topology (adjusted p-value < 0.05) when tested for block sizes of [1, 3] and shift years of [-3 , 3]. The presented p-values of correlations are adjusted using Holm-Bonferroni method [87].

in "TOTAL" trade network topology that occur one and two years later. The strongest correlation is observed two years later, with a Spearman's correlation coefficient of 0.414 and p-value of 0.005 (Figure 3.2-A and -B). These correlations are expected [40], since petroleum is critical for moving goods. Freight transportation consumes about 35% of all transport energy that is used worldwide, which is virtually based only on petroleum [40]. The increases in crude oil price raise the transportation costs, and thus erode the advantages of the long-distance supply chains. Similarly, the significant positive correlation observed with the "Crude Material (except Fuel)" and "Miscellaneous Manufactured" commodities can also be explained with the effects of the increase (or decrease) in the transportation costs, since the products in these categories are highly transportation dependent.

Since WTN consists of trades in many commodities, different commodities are affected differently by the oil price (Figures 3.2 and 3.3). The strongest and immediate effect (in the same year in which oil price changes) is on the trade of "Food & Live Animals": Phi correlation coefficient of

0.479 and p-value of 0.001 (Figure 3.3-A). This may be explained by agriculture needing oil, as well as by increase in demand for corn and soy that are used for production of bio-ethanol and bio-diesel as oil price increases [78, 141]. We further confirm this by observing that the correlation between oil price and the structure of the network of trade in "Food & Live Animals" increases over time, as agriculture becomes more oil dependent: Phi correlation coefficient rises from 0.31 in years 1962 to 1986, to 0.51 in years 1986 to 2007. The significant positive correlations observed for the "Mineral Fuels" and "Chemicals" commodities are also no surprise, as crude oil and its products form the "Mineral Fuels" commodity and the "Chemicals" category includes many different types of petroleum products.

3.4 Graphlet Change Profile of Global Recessions

After observing the correlation between the change in network topology and oil prices, we ask how exactly the network structure changes when there is a global recession in the world. A global downturn is an economic crisis that satisfies the following criterion [56, 128]: (i) a world GDP growth below 2%, (ii) a drop of more than 1.5% in the world GDP growth from previous 5 years' average to current rate, and (iii) a GDP growth that is at a minimum with respect to the two previous and two following years. Based on this definition of global downturn, four global downturns are identified in [57]: 1975, 1982, 1991, and 2001. We investigate the changes in the world trade network during these downturns based on their graphlet counts.

All downturns are characterized by the same deteriorate-then-recover pattern of the graphlets, as illustrated on the global recession of 1991 in Figure 3.4. During a downturn, the counts of weakly connected graphlets (e.g., G_5, G_{15}, G_{16}, G_{20} in Figure 1.4) strongly decrease, while the counts of graphlets representing brokerage relations (e.g., G_{11}, G_{14}), densely connected graphlets (e.g., G_8, G_{29}), and degree (i.e., G_0) remain stable. The deteriorated graphlets are recovered immediately after the downturn.

The deteriorate-then-recover patterns for all global downturns defined by [57] are illustrated in Figure 3.5. The most obvious of these patterns is the 1991 crisis (Figure 3.5-D). The 1982's downturn is almost identical (except for the magnitude) to the downturn patterns of 1991 (Figure 3.5-C), with a deterioration pattern between 1981 and 1982, followed by a recovery

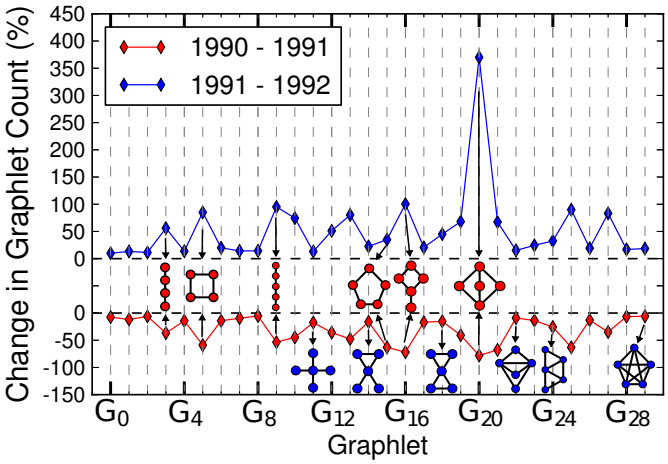
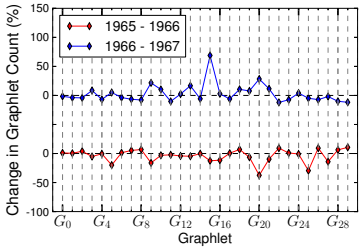


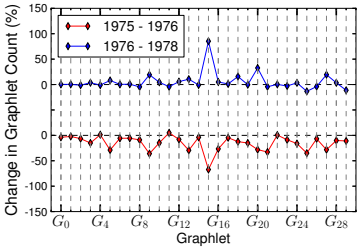
Figure 3.4: Graphlet Change Pattern during the 1991 Global Downturn: The red distribution represents the graphlet count change percentages while entering a crisis, and the blue distribution represents the graphlet count change percentages when getting out of a crisis.

pattern between 1982 and 1983. The 1975's downturn is slightly different (Figure 3.5-B) since the deterioration pattern is observed between 1975 and 1976 (one year later than expected), and is followed by a recovery pattern over two years from 1976 to 1978. The 2001's downturn is also slightly different (Figure 3.5-E) since both the deterioration and the recovery patterns are observed over two years: deterioration between 2000 to 2002, and recovery between 2002 to 2004.

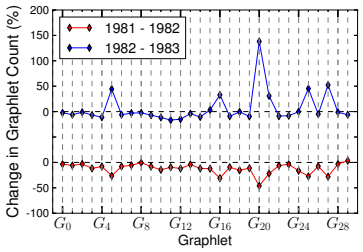
We search for similar deteriorate-then-recover patterns based on the Pearson's Correlation Coefficients of the graphlet change distributions. In particular, we obtain the change distributions of all years by comparing the graphlet count changes within 1 year (e.g., the graphlet count change from 1990 to 1991), and 2 years (e.g., the graphlet count change from 1990 to 1992). This produces 48 (from 1 year change) + 47 (from 2 years change) = 95 change distributions. We pair these 95 distributions in such a way that two change distributions that follow each other (e.g., the change distribution of 2000-2002 and 2002-2003) are combined. When we compute the Pearson's Correlation Coefficients between all pairs of combined graphlet change distributions, interestingly, most of the highest positive correlations



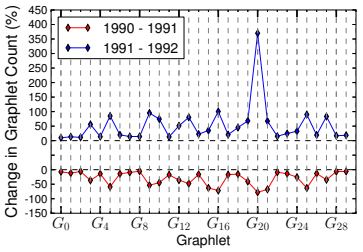
(A) 1966 – Credit Crunch Crisis



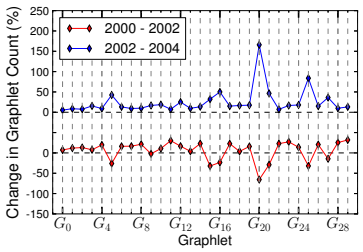
(B) 1975 – Global Downturn



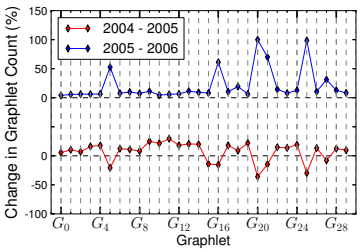
(C) 1982 – Global Downturn



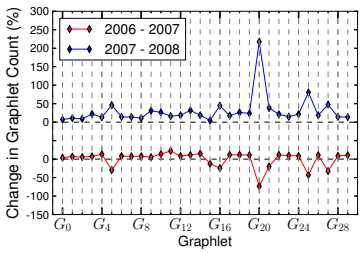
(D) 1991 – Global Downturn



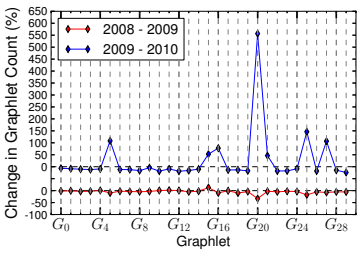
(E) 2001 – Global Downturn



(F) 2005 – Correlation Crisis



(G) 2007 – Sub-Prime Crisis



(H) 2009 – Global Financial Crisis

Figure 3.5: The graphlet count change patterns during crisis years. The red distributions represent the graphlet change distributions while entering a crisis, and the blue lines represent the graphlet change distributions when getting out of the crisis. The presented crisis years are: A – 1966, B – 1976, C – 1982, D – 1991, E – 2001, F – 2005, G – 2007, H – 2009.

are observed among the distributions with the deteriorate-then-recover patterns, that is observed for the four global downturn years. We identify four additional years that have the deteriorate-then-recover pattern and, therefore, being positively correlated with the four downturn years. We validate that these years also correspond to financial crisis years that are not defined by [57]: (1) 1966 - the credit crunch crisis, (2) 2005 - the correlation crisis, (3) 2007 - the sub-prime crisis, and (4) 2009 - the global financial crisis. The graphlet change patterns of these four additional crises are also presented in Figure 3.5.

Despite the studies that focus on the degrees and the density of the trade networks during crisis years, we do not observe any obvious changes on the number of edges (i.e., count of graphlet G_0) during any crises. Therefore, we show that the changes in the topology are not reflected in the number of edges in the network, but in more detailed descriptors that are characterized by graphlets.

3.5 Graphlets and Economic Wealth Indicators

Canonical Correlation Analysis [69] is a technique that identifies combinations of random variables that correlate well with each other. Given two column vectors $X = (x_1, x_2, \ldots, x_m)'$ and $Y = (y_1, y_2, \ldots, y_n)'$ of random variables, canonical correlation analysis seeks weighting vectors a and b such that the random variables $a'X$ and $b'Y$ maximise the correlation $\phi = corr(a'X, b'Y)$. The weighting vectors a and b that maximise the correlation ϕ, reveal the variable subsets in X and Y that are correlated and anti-correlated with each other. After identifying the first set of such weighting vectors a and b, further weighting vectors can be sought, subject to the constraint that they are supposed to be anti-correlated with the first pair of canonical variables; this gives the second pair of canonical variables. This procedure can be repeated up to $min\{m, n\}$ times, each time obtaining weighting vectors for less obvious correlations.

In Section 3.2, our analysis on the graphlet correlation matrices of the world trade network showed that a country in the world trade network is located either peripheral or densely-connected/broker, but not both. We perform further analysis on this observation in order to offer a qualitative explanation of this observation, and relate it with the economic wealth indi-

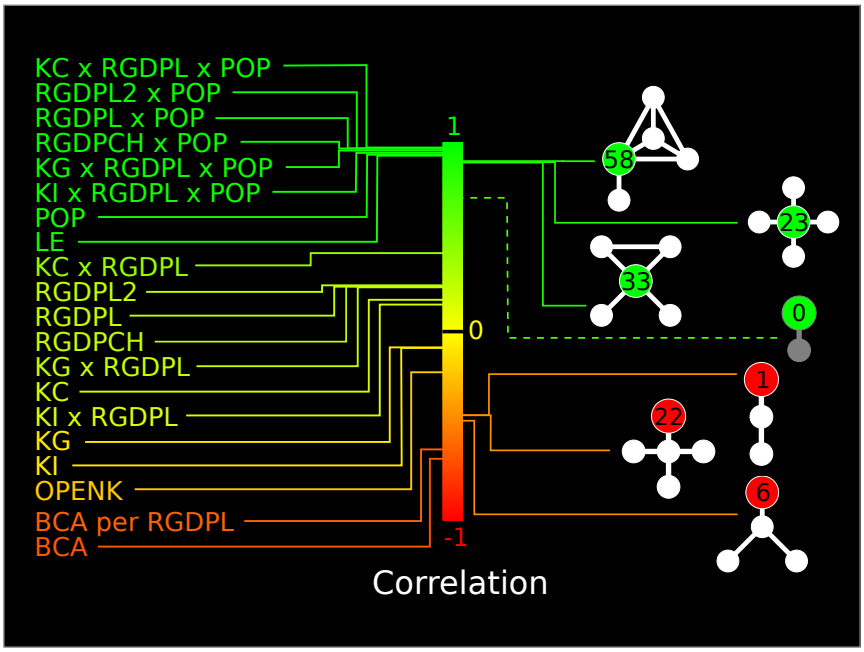


Figure 3.6: Canonical Correlation Analysis Results on Economic Wealth Indicators and Graphlet Degrees of Countries: Broker orbits (e.g., orbits 23, 33, 58) are positively linked with the wealth of a country, while peripheral orbits (e.g., orbits 1, 6, 22) are linked with indicators of economic poverty.

cators of countries. In particular, we use the canonical correlation analysis to correlate economic wealth indicators of a country [60, 80] with its position in the world trade network. In other words, the first set of random variables of canonical correlation analysis, X, corresponds to the economic wealth indicator statistics of countries (explained in Section 3.1) for different years, and the second set of random variables, Y, corresponds to the graphlet degrees of countries in the world trade networks of different years. Due to the limited availability of level of employment (LE) and current account balance (BCA) statistics in the WEO database before 1980s, we perform the canonical correlation analysis on datasets of graphlet degrees and economic wealth indicators for the years after 1980. Figure 3.6 represents the correlation coefficients that are computed from the first set of estimated weighting vectors, a and b.

Interestingly, the indicators of economic wealth such as gross domestic

product (i.e., RGDPL × POP, RGDPL2 × POP, RGDPCH × POP), level of employment (i.e., LE), consumption share of purchasing power parity (i.e., KC × RGDPL × POP), and investment share of purchasing power parity (i.e., KI × RGDPL × POP) strongly correlate with a country being in a brokerage relationship (i.e., a mediator between two unconnected countries), or within a cluster of densely connected countries, while the indicators of economic poverty such as current account balance (i.e., BCA) correlate with a country being peripheral in the network (i.e., linked only to one other country by a trade relationship). Orbit 0 is presented in Figure 3.6 only to illustrate that these results could not have been obtained from node degree. Since a country is either peripheral or clustered/broker, this may indicate that one of the factors that contribute to the wealth of a country could be its brokerage/clustered position in the world trade network.

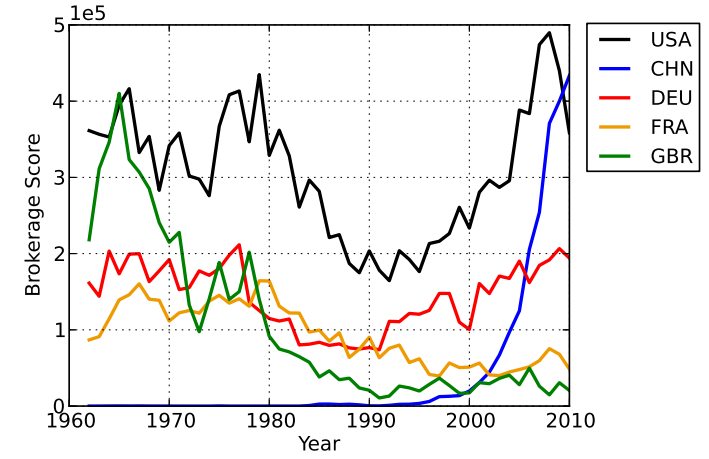
3.6 Effects of a Country's Network Position on its Crisis Patterns

After observing the effects of brokerage position in economic wealth and peripheral position in economic poverty with the canonical correlation analysis, we quantify the strength of a country's broker / periphery position in the world trade network at each year, and track the changes in this positioning over the years. We define the *brokerage score* of a country at a particular year as the weighted linear combination of the graphlet degrees for broker orbits; in particular C_{23}, C_{33}, C_{44}, and C_{58}. We specifically choose these brokerage orbits to compute the brokerage scores as they appear more frequently in the world trade network than the other brokerage related orbits, and they express the brokerage relation more strongly as they appear in sparser graphlets. These orbits were also observed to be better correlated with the economic wealth of the countries in the canonical correlation analysis. Similarly, we define the *peripheral score* of a country at a particular year as the weighted linear combination of graphlet degrees for peripheral orbits; in particular C_{15}, C_{18}, and C_{27}. Apart from appearing in sparser graphlets which appear more frequently in the networks, these orbits are all at distance 2 to the center of the graphlet they reside in; therefore, expresses peripheral positioning more strongly. The weighting coefficients

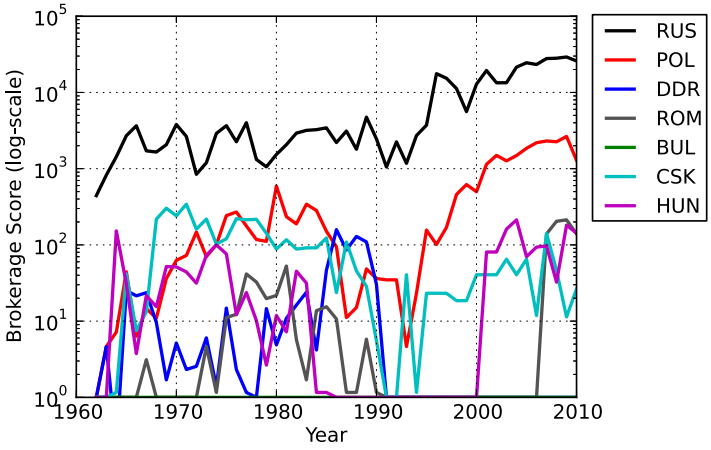
are obtained from the canonical correlation analysis: they are the values of the weighting vectors a and b that maximize the correlation between the economic wealth indicators and the graphlet degrees. With this weighting, while the brokerage/peripheral scores are defined based only on the topological properties of the countries, the obtained scores better correlate with their economic wealth indicators. We track how the network position of a country changes from 1962 to 2010 using the brokerage and peripheral scores, and analyse if these changes coincide with economic crises and other events impacting the economy of the country.

Indeed, we find that during 1980s, brokerage scores of the world's highest brokers fall (Figure 3.7-A), for which we find support in the economic literature. For example, in the USA during the first Reagan administration, a mix of monetary policy and loose budgets sky-rocketed the dollar and sent international balances into the wrong direction. The merchandise trade deficit rose above $100 billion in 1984, there to remain through the decade. The ratio of the USA imports to exports during the eighties peaked at 1.64, a disproportion not seen since the War between the States. Such a drop in the export power of the USA, and thus the change of the trade network, had no precedent in modern USA history [45]. Another example is that of Great Britain. We can see a huge drop in its brokerage score as it loses the Empire in the 1960s, seeing a small improvement in 1973 when the Conservative Prime Minister, Edward Heath, led it into the European Union (EU). However, the downward trend governed by the dissolution of the colonial superpower has continued [104]. On the other hand, the reunification of Germany moved it from being in the shadow of the Second World War a peripheral economy of Western Europe, with most of the decisions in Europe having been made by France and the UK, to being the central economy of Europe [142]. Among the countries of the former Eastern Block, USSR has been the most dominant broker, with both Russia and Poland sharply gaining in brokerage scores after the fall of communism (Figure 3.7-B; y-axis is in logarithmic scale).

Similarly, peripheral scores (Figure 3.8) are consistent with economic reality. China's peripheral score dropped sharply in the early 1970's, which coincides with President Nixon's international legitimization of China [39]. This was a turning point that changed China from a closed economy to an economy deeply integrated into global financial markets [154], as evident



(A)



(B)

Figure 3.7: Brokerage Score Changes between 1962-2010: Panel A – Brokerage scores of the United States (USA), China (CHN), Germany (DEU), France (FRA), and United Kingdom (GBR). Panel B – Brokerage scores of the Eastern Bloc countries: the Soviet Union until 1991 replaced by Russia afterwards (RUS), Poland (POL), Eastern Germany (DDR), Romania (ROM), Bulgaria (BUL), Czechoslovakia until 1991 replaced by the sum of Czech Republic and Slovakia afterwards (CSK), and Hungary (HUN).

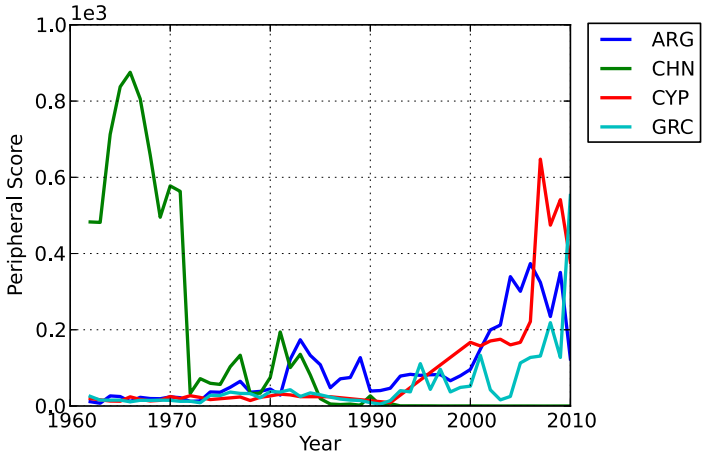
by its fallen peripheral score (Figure 3.8-A) and increased brokerage score that has surpassed that of the USA in 2009 (Figure 3.7-A). Also, raised peripheral scores of Argentina, Cyprus and Greece coincide with their recent economic crises. By year 2001, poor management in great part led to Argentina's real GDP shrinkage, unemployment sky-rocketed, and the international trade plunged, so Argentina turned into a peripheral economy [6]. Less than a decade later, Cyprus and Greece have gone the "South American way": the similarities, starting with the fixed exchange regime followed by the bank runs, were striking [13].

Interestingly, accession of countries into the EU makes them more peripheral in the world trade network, as evident by increases in their peripheral scores before and after accession (Figure 3.8-B). Even though all trade within the EU is exempt of all import taxes, at the time of accession countries are required to leave other advantageous free trade associations (e.g., BAFTA, CEFTA, CISFTA, EFTA). The fact that a country has to leave free trade agreements with other non-EU member countries leads to the destruction of trade connections while the positive effects of EU accession on trade need time to materialize. In other words, since trade connections are easy to break, but much more difficult to build, EU accession increases the peripheral score of a country and whether and when the country will recover remains an open question.

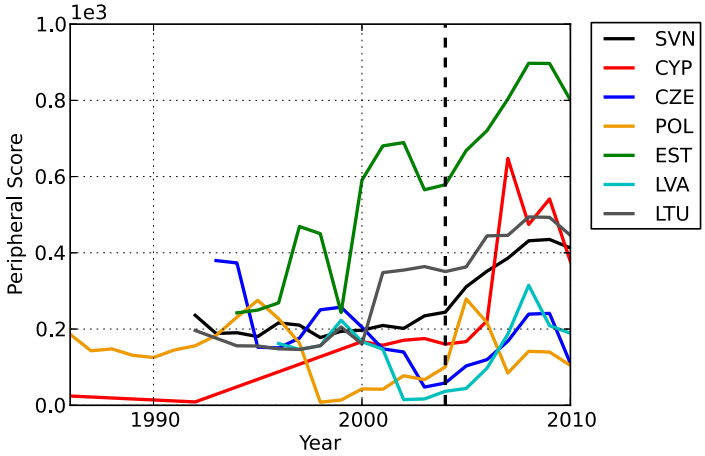
3.7 Author's Contributions

Ömer Nebil Yaveroğlu collaborated with Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj for the work presented in this chapter.

In this collaboration, Ömer Nebil Yaveroğlu designed the experiments linking the change in trade network topology and crude oil price, defined brokerage / peripheral scores of countries, performed all the experiments except the canonical correlation analysis, and documented the results of these experiments.



(A)



(B)

Figure 3.8: Peripheral Score Changes between 1962-2010: Panel A – Peripheral scores of Argentina (ARG), China (CHN), Cyprus (CYP), and Greece (GRC). Panel B – Peripheral scores of countries that joined EU in 2004 and show an increase in their peripheral scores right before and after joining the EU: Slovenia (SVN), Cyprus (CYP), Czech Republic (CZE), Poland (POL), Estonia (EST), Latvia (LVA), and Lithuania (LTU).

4 Models of World Trade Networks

In this chapter, we first evaluate the goodness-of-fit of the seven random network models (which are described Section 1.5.1) on autonomous, Facebook, metabolic, protein structure, and world trade networks (Section 4.2). Since all of the seven network models fail to fit the world trade networks, we propose two random network models that fit well on these networks: (1) the gravitational random network model, and (2) the brokerage model (Section 4.3). We extend our analysis on the world trade networks in Chapter 3 based on the properties of brokerage network model (Section 4.3.3).

4.1 Motivation

Identifying the models that fit real-world networks sheds light onto the rules that govern the formation and evolution of these networks. Model-fitting tests require successful network comparison techniques for assessing the topological correspondence between the networks described by the models and the input networks. For example, the seven network models that are explained in Section 1.5.1 (i.e., ER, ER-DD, GEO, GEO-GD, SF-BA, SF-GD, and STICKY) were compared for their successes in explaining the topology of protein interaction networks using relative graphlet frequency distance (RGFD) and graphlet degree distribution agreement (GDDA) [76, 156, 157] (detailed in Section 1.5.3), giving insights into understanding the organizational principles of these networks. In Chapter 2, we have shown that our new network distance measure, the Graphlet Correlation Distance (GCD), performs better than RGFD and GDDA in network classification. This raises the need for re-evaluating the network models that best fit the topology of real-world networks. On the other hand, the graphlet-based model fitting experiments are mostly applied for identifying well-fitting models of

protein-protein interaction networks [156, 157], showing that the topology of protein-protein interaction networks are best modelled by SF-GD, GEO-GD, and STICKY models [76, 158, 159]. Identification of the models that fit other types of networks from different real-world domains such as technology (e.g., autonomous networks), sociology (e.g., Facebook networks), finance (e.g., world trade networks), and biology (e.g., metabolic networks, protein structure networks) remains an open problem.

In this chapter of the dissertation, we analyse the fit of network models on the "unmodelled" real-world network types using the model fitting procedures that are explained in Section 1.5.3. In these tests, we use the accurate and sensitive GCD to measure the distance between the model networks and the real-world networks. We analyse the fit of the seven network models that are explained in Section 1.5.1 on five different types of real-world networks from different domains: (1) autonomous systems networks, (2) Facebook networks, (3) metabolic interaction networks, (4) protein structure networks, and (5) world trade networks. These networks are obtained from public datasets as explained in Section 1.2. The sizes and densities of these real-world networks are summarized in Table 4.1. Among the analysed real-world network types, world trade networks were not fit by any of the seven network models. To understand the distinct topology of the world trade network better, we introduce two new models of the world trade networks, test their fit on the observed topology of these networks, and analyse the world trade networks based on the main characteristics of these new models.

	Number of	Number of Nodes			Edge Densities (%)		
Network Type	Networks	Min.	Med.	Max.	Min.	Med.	Max.
Autonomous Systems	733	103	4180	6474	0.06	0.09	4.55
Facebook	98	769	9949	41554	0.16	0.78	5.70
Metabolic	2301	100	366	705	0.74	1.17	3.39
Protein Structure	8226	100	178	1419	0.47	3.75	8.31
World Trade	49	86	103	125	8.72	11.64	13.53

Table 4.1: Summary of the sizes and densities of the real-world networks from different domains.

4.2 Model-fitting on Real-world Networks

We first evaluate the fit of the seven network models (i.e., ER, ER-DD, SF-BA, SF-GD, GEO, GEO-GD and STICKY) on the five following real-world network types: (1) autonomous system networks, (2) Facebook networks, (3) metabolic (enzyme – enzyme) networks, (4) protein structure networks, and (5) world trade networks (explained in Section 1.2). We use GCD-11 to compute the topological distances between the model networks and the real-world networks, and the state-of-the-art non-parametric test of Rito et al. [163] for evaluating the model fit (the method is explained in Section 1.5.3). For each input network, we first generate 30 networks from each of the seven models ($7 \times 30 = 210$ model networks per input network) having the same size and density with the input network. For each model, we compute the two following GCD distributions: (1) the distribution of data-vs-model distances corresponding to the distances between the input network and the 30 model networks, (2) the distribution of model-vs-model distances corresponding to the GCDs between $\binom{30}{2} = 435$ model network pairs. If these two distributions intersect, then the input network is in the set of topologies that the network model can generate. Therefore, the model fits the network.

When performed as described above, the non-parametric model fitting test evaluates the fit of a model on a single network. In order to extend this approach to evaluate the fit of a model to a set of networks from the same domain, we combine the data-vs-model and model-vs-model distances from each individual test, producing two distributions that test the overall fit of a model to a network domain.

Figures 4.1 and 4.2 presents the results of model-fitting experiments on the autonomous, Facebook, metabolic, and protein structure networks. For autonomous networks, ER-DD is the best fitting model, as identified by the observed intersection of the two distributions and the smallest data-vs-model distances. Surprisingly, the three other network types (i.e., Facebook, metabolic, and protein structure) are all best modelled by three network models that are geometric model (GEO), geometric model with gene duplications and mutations (GEO-GD), and scale-free model with gene duplications and mutations (SF-GD). While it is not difficult to explain why biological networks are best fit by networks that model evolutionary pro-

cesses, it may be surprising that Facebook networks seem to be organized by the same principles. A possible explanation is that Facebook grows as follows: when a person joins Facebook, he links to a group of his friends, which mimics a gene duplication, but he hardly ever has exactly the same friends as another person, which mimics the evolutionary process of divergence, or mutation. The fit of GEO to both Facebook and biological networks is perhaps more straightforward to explain, since all biological and social entities are subject to spatial constraints [159]. To our knowledge, this is the first time that such a parallel between online social networks and bio-molecular networks has been uncovered. It opens questions about networks from very different domains following the same evolutionary and organizational principles that may lead to explaining various societal processes.

When we perform the same model-fitting test on the world trade networks, surprisingly, no intersections of data-vs-model and model-vs-model distance distributions are observed for the seven random network models except the ER-DD model (Figure 4.3). For the ER-DD model, an intersection is observed, however the ER-DD model is unstable for the size and edge-density of world trade networks [76]. This instability is clearly observable with the widespread model-vs-model distances of this model in the range between 0 and 6. For this reason, it does not describe a well-defined network structure and cannot be accepted as a well-fitting model for the world trade networks. Interestingly, this result goes in parallel with what we observed in Figure 2.8, where the world trade networks are clearly separated from the networks from the four other real-world network types with high GCD distances. Therefore, all of the seven random network models fail to fit the topology of world trade networks, and better models of world trade networks are needed.

4.3 Models of World Trade Networks

The problem of understanding the rules governing the world trade is gaining interest, especially due to the recent global recession. Our analysis in Section 4.2 shows that the seven standard random network models (i.e., ER, ER-DD, SF, SF-GD, GEO, GEO-GD and STICKY) fail to fit the world trade networks and new models that correctly describe the topology of the world trade networks are needed. Gravity Model of Trade [4] and Core-

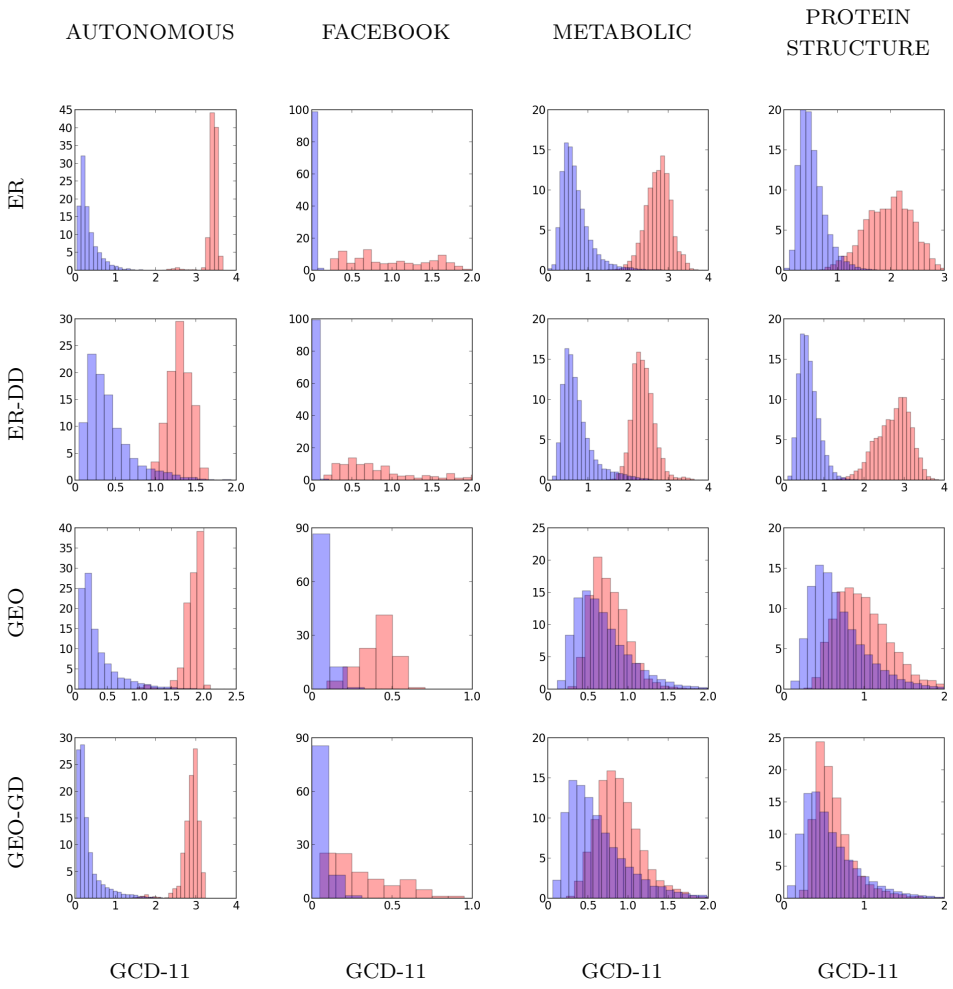


Figure 4.1: Results of comparing the seven random network models with the autonomous, Facebook, metabolic, and protein structure networks. The horizontal axis represents the GCDs, and the vertical axis represents the percentage of distances with the corresponding GCD. The blue distributions represent the model-vs-model distances, while the red distributions represent the data-vs-model distances.

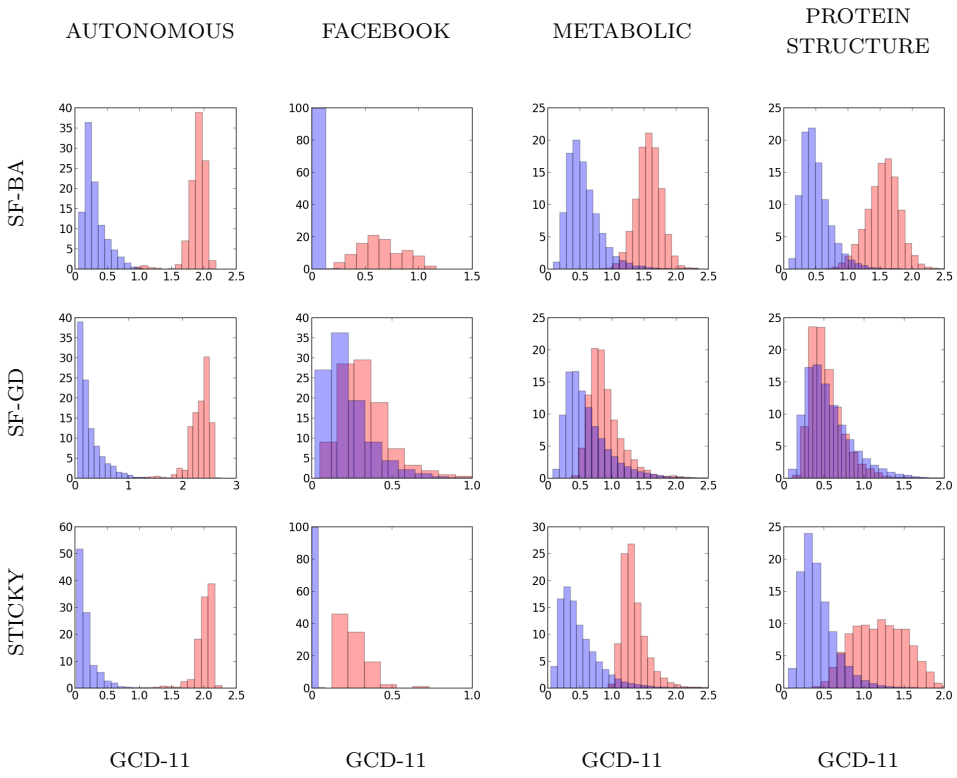


Figure 4.2: Results of comparing the seven random network models with the autonomous, Facebook, metabolic, and protein structure networks (continued). The horizontal axis represents the GCDs, and the vertical axis represents the percentage of distances with the corresponding GCD. The blue distributions represent the model-vs-model distances, while the red distributions represent the data-vs-model distances.

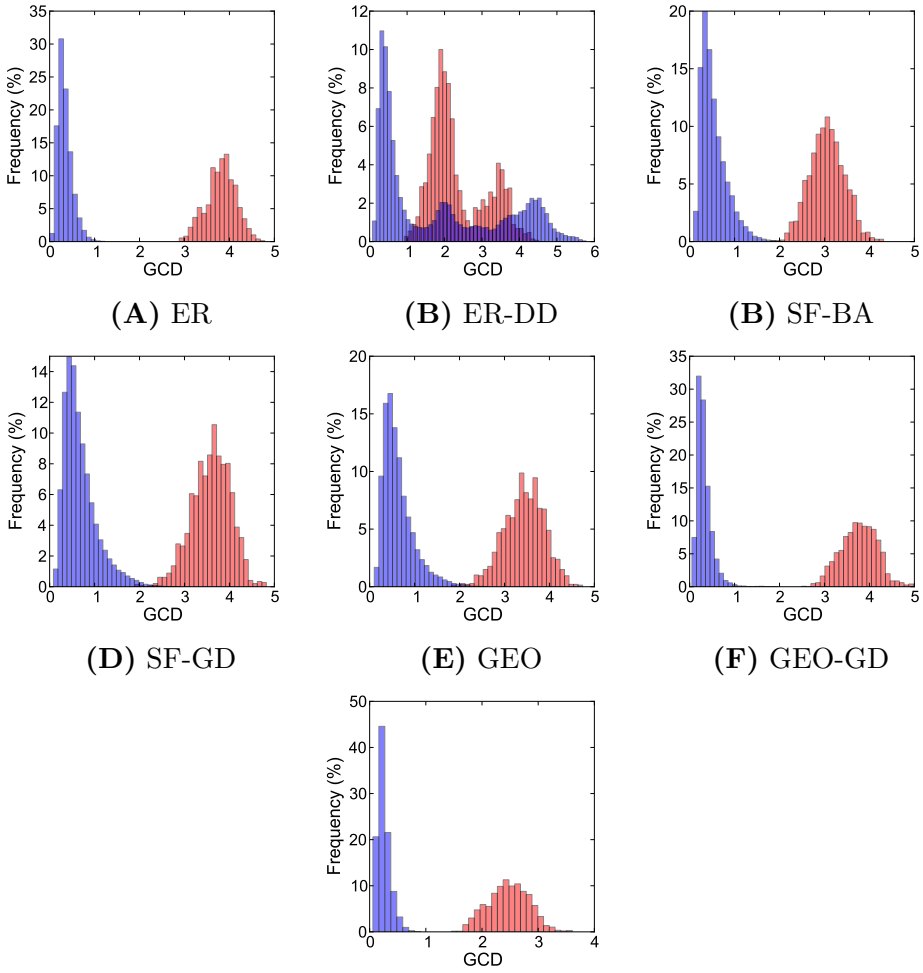


Figure 4.3: Comparison of the seven random network models with the world trade network of different years. The panels show the model-fitting test results of the following models on the world trade network: A – ER , B – ER-DD, C – SF-BA, D – SF-GD, E – GEO, F – GEO-GD, and G – STICKY. The blue distributions represent the model-vs-model distances, while the red distributions represent the data-vs-model distances.

Periphery model [32, 44, 77, 84, 153, 176] are widely accepted models of trade, though there is an ongoing debate about their suitability in explaining the observed topology of world trade networks (Section 1.6). Although the fit of these models on the world trade networks is evaluated individually, no systematic comparison of their topological goodness-of-fit have been made so far. Furthermore, these models have been mostly studied as descriptive models, not as generative models that can produce random graphs based on their principles. Moreover, no core-periphery modelling studies highlight the importance of the broker position in this organisation.

In this section of the dissertation, we contribute to the debate on the models of the world trade networks by proposing two new random network models; (1) Gravitational Random model, and (2) Brokerage model. We systematically evaluate the goodness-of-fit of these two models on the world trade networks by applying the non-parametric model-fitting test of Rito et al. [163] with the graphlet correlation distance (GCD-11).

4.3.1 Gravitational Random Model

We design a new generative network model, called Gravitational Random model (GR), that follows the principles of the Nobel Prize winning descriptive network model, called the Gravity Model of Trade [4]. Analogous to the Newton's Law of Universal Gravitation, the Gravity Model of Trade suggests that the trade volume (i.e., attraction) between two countries a and b, denoted by $F(a, b)$, is proportional to the product of their economic masses M_a and M_b (e.g., their Gross Domestic Products) and inversely proportional to their geodesic distance d_{ab}, as in Equation 4.1.

$$F(a, b) = \alpha \times \frac{M_a \times M_b}{d_{ab}}. \tag{4.1}$$

Given a trade network, we generate an instance from GR model as follows. First, we compute three values: (1) the matrix of pairwise geodesic distances between countries, D; (2) the empirical distribution of countries' GDP values, M; and (3) the number of edges in that trade network, e. Next, from a piecewise affine approximation of the distribution M, we generate a new set of random GDP values, M', which we then associate to the nodes of the model network. M and M' are verified to follow the same distribution using the Mann-Whitney rank-sum test [200]. Finally, we compute the edge

weights of all country pairs using Equation 4.1 with the distances in D, the GDPs in M', and $\alpha = 1$. The resulting model network is defined by the e highest weighted edges.

To evaluate the effectiveness of GR model in reproducing the topology of the world trade networks, we apply the non-parametric model fit test of Rito et al. [163] to networks generated from the GR model (as in Section 4.2). The results show that the GR model can reproduce the topology of world trade networks (Figure 4.4). This was concluded by observing the intersection between data-vs-model and model-vs-model distance distributions: the blue distribution represents the model-vs-model distances between GR models generated based on the properties of the world trade networks, while the red distribution represents the data-vs-model distances; i.e., the distances between the world trade networks and their corresponding GR models.

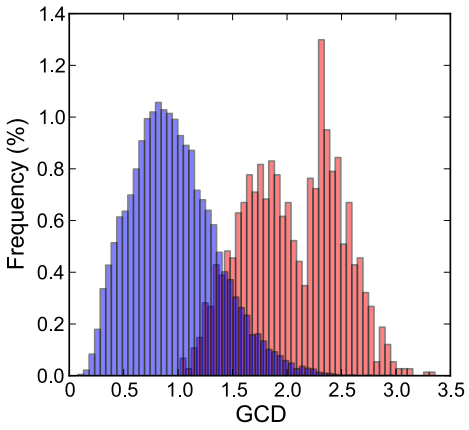


Figure 4.4: Comparison of Gravitational Random Model with World Trade Networks. The blue distribution represents the model-vs-model distances, while the red distribution represents the data-vs-model distances.

In contrast to standard random network models that do not fit the world trade networks (Section 4.2), the GR model shows a significant intersection between the two distributions (Mann-Whitney-U test, p-value ≤ 0.05). But still, half of the data-vs-model distances do not intersect with the model-vs-model distances. This means that GR model does not capture all the topological features of the world trade networks, and there is still room for

improvement.

4.3.2 Brokerage Model

Core-Periphery network models highlight the hierarchical organisation of the world trade network topology [32, 44, 77, 84, 153, 176]. Our analysis on the world trade networks (Chapter 3) highlights the importance of broker position of countries in their wealth; countries that mediate the trade between core and peripheral countries tend to be richer. Based on these observations, we propose a new generative random network model, the Brokerage model, that imposes a three-layer organisation for modelling the world trade networks. These three layers are formed by the densely connected nodes, broker nodes that mediate the trade between disconnected nodes, and peripheral nodes that are weakly connected to the rest of the network.

The brokerage model aims to maximize the number of G_{23} graphlets (Figure 1.4) in a random network with a defined number of nodes and edges. Given a network G, the brokerage model first generates a random ER network that contains the same number of nodes and edges as G. At each step of the G_{23} count optimization, an edge is randomly chosen and rewired; i.e., removed from the network, and one of its nodes is connected to another node in the network. If the rewiring increases the number of G_{23} in the network, the change is accepted and the algorithm iterates keeping the rewired edge. Otherwise, the rewiring is rejected and the iterations continue with the network before rewiring. If this iterative procedure fails to identify an accepted rewiring for a predefined number of steps, the algorithm returns the resulting network. For the size and density of the world trade networks, we observed that a sufficient threshold for convergence is $5,000$ states without any changes. With this optimization procedure, we do not aim to generate the network at global maximum (i.e., the network containing the maximum possible number of G_{23} for the size and density of the generated network), but we would rather generate networks at local maximums (i.e., networks containing high numbers of G_{23} that do not necessarily have to be the maximum possible count). Identifying local maximums produces a wider-range of random networks that have high numbers of G_{23} graphlets but with more diverse topological configurations.

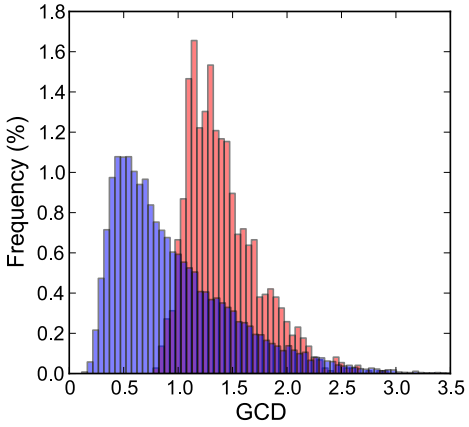


Figure 4.5: Comparison of Brokerage Model with World Trade Networks. The blue distribution represents the model-vs-model distances, while the red distribution represents the data-vs-model distances.

Again, we assess the effectiveness of brokerage model in reproducing the topology of world trade networks by applying the non-parametric model-fitting test of Rito et al. [163] (as in Section 4.2). Similar to the GR model, we observe a significant intersection between the data-vs-model and model-vs-model distributions (Mann-Whitney-U test, p-value ≤ 0.05), meaning the brokerage model fits the world trade networks (Figure 4.5). Interestingly, the Mann-Whitney score for the brokerage model is $6,505,259$ while the score for the GR model is $1,106,388$. This means that the data-vs-model and model-vs-model distributions are more likely to be generated by the same distribution. In other words, the intersection between the data-vs-model and model-vs-model distance distributions are statistically larger for the brokerage model than for the GR model.

Additionally, the data-vs-model distances are much smaller for the brokerage model than for the GR model: for the brokerage model, the mean of data-vs-model distances is 1.414 and their median is 1.341, while for the GR model the mean is 2.044 and the median is 2.036. All of the above suggests that the brokerage model fits the world trade networks substantially better than the GR model, even without using country-specific attributes (i.e., countries' GDP and longitude/latitude information are needed for the

GR model, but not for the brokerage model).

When data and model networks are compared on a per year basis (for the 49 years between 1962 - 2010), we observe that the brokerage model consistently has lower average GCD values than the GR model, approximating the topology of world trade networks better (Figure 4.6). This highlights an important topological characteristic of the world trade network system: it tends to maximise the core-broker-periphery organisation by obtaining the highest possible number of G_{23} graphlets for a given network size and edge-density.

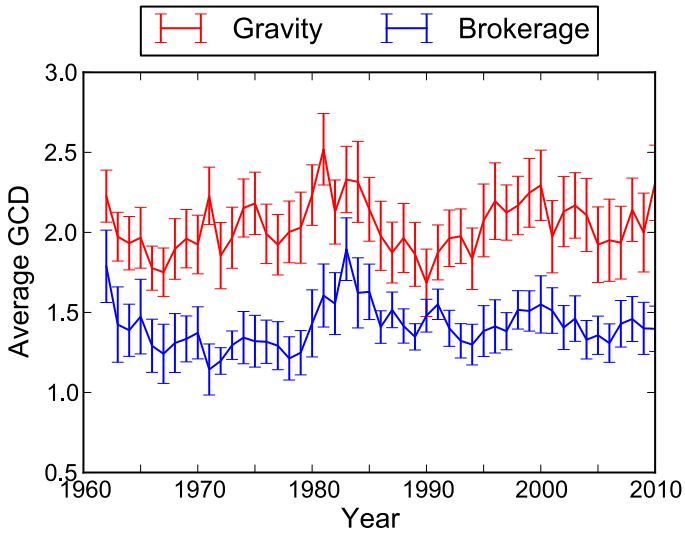


Figure 4.6: Per year data-vs-model distances between Brokerage and Gravity Random Models and world trade networks. Averages and standard deviations of data-vs-model graphlet correlation distances of Brokerage and Gravity Models from world trade networks.

The good fit of the brokerage model expose two interesting observations about world trade networks. He et al. [77] show that the hierarchical organisation of the world trade network decreased in recent years due to the effect of globalisation. In Figure 4.6, we observe that the average GCDs of the brokerage models during 1970's are not that different from the ones after 1990's, except the slight increase (but still predominantly constant) in average GCD values for the period after 1990's. It is this increase which re-

flects the minor loss of hierarchical organisation shown by [77]. Nevertheless, the brokerage model shows that, in spite of globalisation, the core-broker-periphery topology is still dominant in the networks of world trade all the time.

The second observation relates the Bretton Woods era (1945–1971) with the fit of brokerage model. With the foundation of the International Monetary Fund (IMF) and the International Bank for Reconstruction and Development (the World Bank) in 1944, the system of international relations that emerged after 1945 divided the world into three parts [160]: (1) the Capitalists that are well-connected among themselves; (2) the Eastern bloc of countries which are under Communist rule and largely isolated; and (3) the developing Third World countries that were produced by the decolonisation which was completed by 1970. Promoting the core-broker-periphery organisation during its time, the Bretton Woods system collapsed in 1971, causing deregularisation of international capital markets. This pattern is consistent with the change in the average GCDs captured by brokerage models of trade networks from that period: in Figure 4.6, we observe a gradual decrease in the GCDs of 1962–1971 brokerage models, indicating an increasing core-broker-periphery organisation during that time period. The core-broker-periphery organisation is most prominently visible in the world trade network of 1971 as it has the minimal GCD value among all the 49 modelled years. We hypothesize that these observations could be a consequence of the effects of the Bretton Woods era (1945–1971) of globalisation [160].

4.3.3 Analysing World Trade Network Organisation using the Brokerage Model

We have seen that the brokerage model, which is based on graphlet G_{23}, can be used to rather accurately describe the topology of world trade networks. Next, we want to know whether any one of the three positions (core, broker or periphery) in graphlet G_{23} is advantageous for the wealth of a country, as well as whether the wealth of a country can be predictive of its future topological position within the trade network.

To test whether there is a correlation between the wealth of a country and its topological position in the trade network, we compute the Pearson's

Correlation Coefficient (PCC) between the graphlet degrees of orbits 56 (peripheral position), 57 (core position), and 58 (broker position) in graphlet G_{23} (Figure 1.4) and the economic wealth indicators of a country; i.e., Gross Domestic Product, Consumption Share, Investment Share, Government Consumption Share and Level of Employment (more information on economic wealth indicators is available in Section 3.1). As expected, we find that the core and brokerage positions correlate positively, and the peripheral position correlates negatively with the countries' wealth indicators (Table 4.2). The brokerage position (i.e., orbit 58) is highly correlated with all five above-listed economic indicators of wealth (Pearson's Correlation Coefficient ≥ 0.8; shown in Figure 4.7).

	Periphery (Orbit 56)	Core (Orbit 57)	Broker (Orbit 58)
Gross Domestic Product	-0.2749	0.4350	0.8688
Consumption Share	-0.2620	0.4067	0.8489
Investment Share	-0.2708	0.3978	0.8390
Government Consumption Share	-0.2560	0.4419	0.8067
Level of Employment	-0.3205	0.2696	0.8751

Table 4.2: Pearson's Correlation Coefficients of economic wealth indicators and graphlet degrees of core-broker-periphery orbits in graphlet G_{23}.

The Pearson's Correlation Coefficients among the graphlet degrees of orbit 58 and the economic wealth indicators are all ≥ 0.8, showing that the graphlet degrees can be predictive of the economic wealth indicators and vice versa. In this respect, we identify the affine transformations using the least squares method. The identified transformations among graphlet degrees of orbit 58 (C_{58}) and economic wealth indicators are listed as follows:

- Gross Domestic Product $= 117,882,909.79 \times C_{58} + 115,362.201$
- Consumption Share $= 74,875,286.357 \times C_{58} + 78,501.651$
- Investment Share $= 28,287,531.946 \times C_{58} + 28,323.132$
- Government Consumption Share $= 9,558,625.724 \times C_{58} + 9,852.750$
- Level of Employment $= 3.901 \times C_{58} + 0.002$

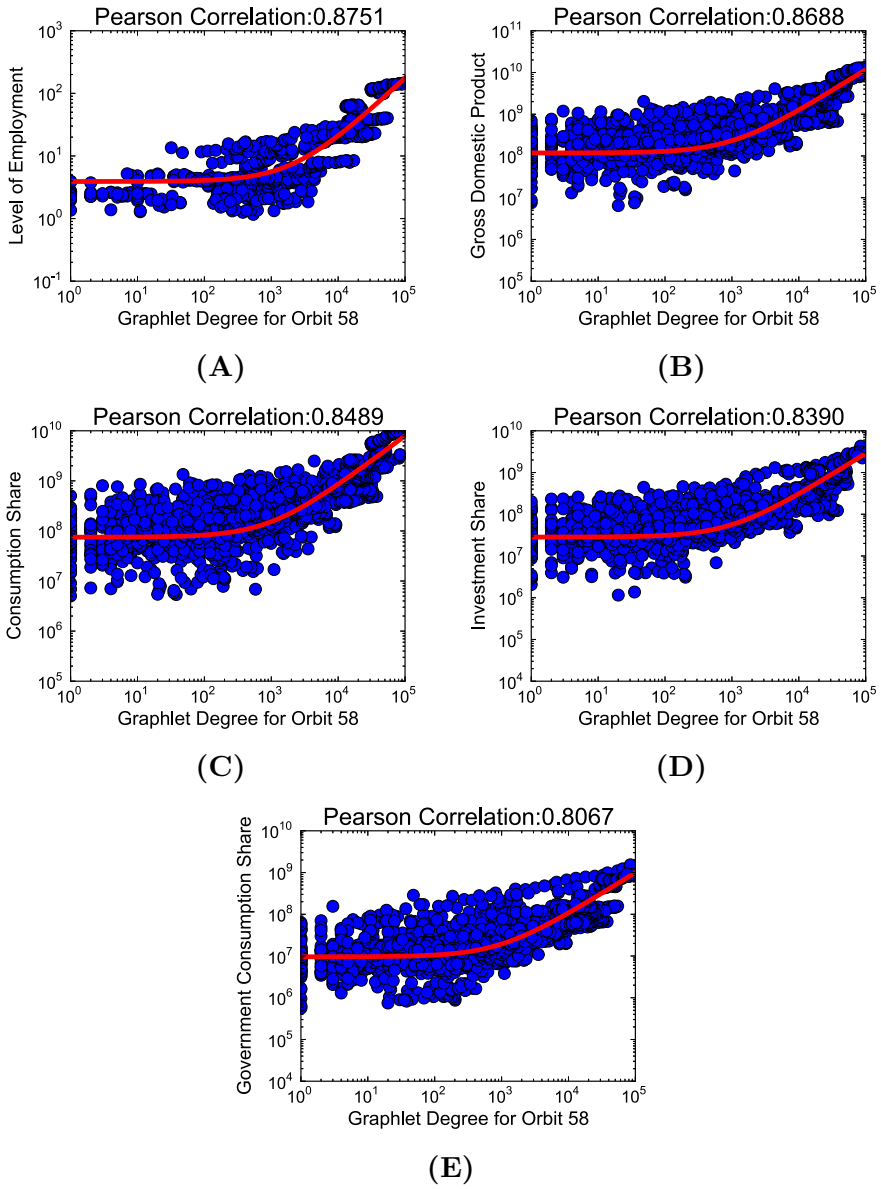


Figure 4.7: Scatter plots of economic wealth indicators vs. graphlet degrees of orbit 58 and the corresponding Pearson's Correlation Coefficients. The economic indicators that are illustrated are: A – Level of Employment, B – Gross Domestic Product, C – Consumption Share, D – Investment Share, and E – Government Consumption Share. The affine least squares fits are plotted with red lines. All five panels are in log-log scale, causing the fitted lines to be visualized as curves.

The high correlation between orbit 58 and a country's wealth yields two similar questions — can the current economic wealth indicators of a country be predictive of its brokerage position in the short-, mid- and long-term?; and, conversely, can its past brokerage position be predictive of its short-, mid- and long-term wealth indicators? To answer these two questions, we compute the Pearson's Correlation Coefficient between the economic indicators of year n and graphlet degrees of orbit 58 at year $n + year\_shift$. A positive $year\_shift$ (i.e., $+5, +10, +20$) tests the predictive power of current wealth indicators on a country's future brokerage position, and a negative $year\_shift$ (i.e., $-5, -10, -20$) tests the predictive power of past brokerage position on future wealth indicators. A zero $year\_shift$ indicates that the correlation is computed for the same year.

We find that Gross Domestic Product and Consumption Share values best correlate with same-year broker position (Table 4.3). This highlights the direct relation between these two economic wealth indicators and the country's current brokerage position (the correlation gradually drops over the following 20-year period). On the other hand, Investment Share, Government Consumption Share, and Level of Employment are predictive of a country's short-, mid- and long-term brokerage position, respectively (Table 4.3). This means that: (1) investments made at a particular year have short-term effects on the broker position of the country (highest correlation for a 5-year shift); (2) government consumption share, which involves infrastructure expenditures such as investments in education, transport, health and military services, has observable effects on the brokerage position of the country over a 10-year period; and (3) level of employment, which indicates the current size of the country's economy is predictive of that country's broker position over the long run.

4.4 Author's Contributions

Ömer Nebil Yaveroğlu collaborated with Noël Malod-Dognin, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj for the work presented in this chapter.

In this collaboration, Ömer Nebil Yaveroğlu collected all the analysed real-world networks except world trade networks, performed all the modelling experiments, implemented the Gravitational Random Model, designed

Year Shifts	-20	-10	-5	0	5	10	20
GDP	0.7677	0.7855	0.8270	**0.8688**	0.8647	0.8571	0.8430
Consumption	0.8010	0.8104	0.8376	**0.8489**	0.8466	0.8425	0.8358
Investment	0.6589	0.6817	0.7484	0.8390	**0.8513**	0.8385	0.8172
Gov. Consumption	0.5973	0.6488	0.7127	0.8067	0.8405	**0.8565**	0.8520
Employment	0.8520	0.8749	0.8768	0.8751	0.8735	0.8780	**0.8928**

Table 4.3: Pearson's Correlation Coefficients of economic wealth indicators and graphlet degrees of broker orbit (i.e., orbit 58) for different year shifts. The Pearson's Correlation Coefficients for Gross Domestic Product (GDP), Consumption Share (Consumption), Investment Share (Investment), Government Consumption Share (Gov. Consumption), and Level of Employment(Employment) are presented. The highest correlation values of each row are written in bold.

and implemented the Brokerage Model, designed and performed the experiments investigating the predictive power of economic attributes in the future brokerage position, and wrote the first version of the paper that describes these new models of world trade.

5 Exponential-family Random Graph Modelling using Graphlet Terms

In this chapter, we propose a generic framework that is based on exponential-family random graph models, and that enables network modelling based on any graphlet property (Section 5.2). We explain the algorithmic details about our framework (Section 5.3), and demonstrate the application of this framework by modelling two networks – one from the social domain and the other from biological domain (Section 5.4). We finalize this chapter by summarizing the effects of the current limitations of exponential-family random graph models on our framework, and make suggestions on handling these limitations (Section 5.5).

5.1 Motivation

To our knowledge, we defined the first generative random network model that is based solely on the graphlet properties, the Brokerage Model (Section 4.3.2). The superior fit of the brokerage model on the world trade networks motivates us to consider network modelling based on graphlet properties more extensively. Since the topological characteristics of each network model are different, we aim to develop a generic network modelling framework that allows defining and exploring network models based on the statistics of any combinations of graphlets. Exponential-family Random Graph Models (ERGMs) define an environment that is suitable for implementing this framework. ERGMs are probabilistic network models that are parametrized by sufficient statistics based on structural network properties (detailed description is provided in Section 1.5.2). Using the graphlet statistics as the model terms of the ERGMs, networks can be modelled based on

any of the graphlet properties.

The *ergm* package [92] for *R* statistical computing system is a collection of tools for network analysis within an ERGM framework. This package contains a wide variety of modelling terms that enable defining ERGMs based different network properties; e.g., the degree distribution, the number of triangles, the correspondence between the node attributes and node degrees, the number of cycles, the number of stars. Though some of these built-in model terms correspond to subgraph properties to some extent, they do not exactly match with what we want to achieve by graphlet based modelling since: (1) the subgraph statistics of these built-in terms are not based on induced subgraph properties, but partial subgraph properties, (2) these built-in terms do not cover all possible patterns that may appear among subgraphs with 4 and 5 nodes, and (3) these built-in terms are not sufficient for relating numerical and categorical node attributes with the subgraph statistics. Luckily, the set of available modelling terms in *ergm* package is extendible using the *ergm.userterms* package [73], and any user-defined network statistics can be embedded into ERG modelling process (see Section 1.5.2 for details).

We exploit the *ergm.userterms* package to embed graphlet statistics into *ergm* package as new modelling terms. In this respect, we implemented the *ergm.graphlets* package that contains four new modelling terms that are defined based on graphlet properties of networks. The *ergm.graphlets* package resolves the above listed issues with the built-in terms of the *ergm* package. In this section of the dissertation, we first introduce these four graphlet based modelling terms, and then, we provide the algorithmic details on their implementation and validation. We continue by applying the new terms of *ergm.graphlets* package on modelling two different networks; one from the social and the other from the biological domain. We conclude this section with a discussion on the possible weaknesses of this modelling methodology.

5.2 Graphlet Terms for Exponential-family Random Graph Modelling

Graphlets are local network properties that successfully capture the topological characteristics of a network. We exploit these powerful topological descriptors for exponential-family random graph modelling by implementing the *ergm.graphlets* package. The package introduces four new ERGM terms based on graphlet properties: (1) *graphletCount* - graphlet counts, (2) *grorbitCov* - graphlet orbit covariance, (3) *grorbitFactor* - graphlet orbit factor, and (4) *grorbitDist* - graphlet orbit distribution. Detailed descriptions of these ERGM terms are as follows:

1. **Graphlet Counts – *graphletCount(g)*:**
   The statistics of the number of times that a graphlet appears in a network can be included into an ERGM by using the *graphletCount* term. The question answered by the change score function of this term is: "How do the number of graphlets of type G_i change when an edge is flipped in the network?". This term has an optional argument, g. g is a vector of distinct integers representing the list of graphlets to be evaluated during the estimation of model coefficients (the complete list of graphlets are illustrated in Figure 1.4). When this argument is not provided, all graphlets are evaluated by default; i.e., in R notation $g = c(0:29)$. The term adds one network statistic to the model for each element in g. This term is defined for all 30 graphlets containing 2 to 5 nodes. Therefore, g accepts values between 0 and 29. Values outside this range are ignored.

   The *graphletCount* term shows similarity with some terms of the *ergm* package; e.g., *cycle*, *edges*, *kstar*, *threepath*, *triangle*, *twopath*. There is a major difference between these terms and the *graphletCount* term. Graphlets are defined as induced subgraphs. Therefore, *graphletCount* does not count a subgraph as a two-path if the subgraph actually forms a triangle when induced on the nodes of the graph. In contrast, the above listed terms of *ergm* package do not require subgraphs to be induced. For this reason, a three node subgraph that forms a *triangle* is also counted as three *twopath* subgraphs. A closer parallel is the *triadcensus* term, which counts induced subgraphs on three nodes;

note, however, that the triad census includes all isomorphism classes of order 3, while the order 3 graphlets consist only of the classes corresponding to connected graphs. Thus, while there is overlap between some quantities computed by *graphletCount* and some built-in terms of *ergm* package, the two are on the whole distinct.

2. **Graphlet Orbit Covariance – *grorbitCov(attrname, grorbit)*:** The covariance of a node's graphlet degree and a numeric node attribute value can be included into the ERGM by using the *grorbitCov* term. The *grorbitCov* term quantifies the covariance between node attributes and graphlet degrees using a network statistic that is defined as the sum of the multiplication of node attribute values with the graphlet degrees of the corresponding nodes. The question answered by the change score function of this term is: "How does the value of the node attribute relate with the change in the graphlet degree?". This term has two arguments: (1) *attrname*, and (2) *grorbit*. The *attrname* is a character vector providing the name of a numeric attribute in the network's node attribute list to the function. The optional *grorbit* argument is a vector of distinct integers representing the list of graphlet orbits to include into the ERGM model (the complete set of graphlet orbits are illustrated in Figure 1.4). When *grorbit* is not provided, all graphlet orbits are evaluated by default; i.e., in R notation $grorbit = c(0, 72)$. The term adds one network statistic to the model for each element in *grorbit*. Each term is equal to the sum:

$$grorbitCov(G, i, X) = \sum_{v \in V} C_i(G, v) \times X_v, \tag{5.1}$$

where X is the vector of attribute values, i is the queried graphlet orbit and $C_i(G, v)$ is the number of graphlets in network G that touch node v at orbit i. This term is defined for the 73 orbits corresponding to graphlets with up to 5 nodes. Therefore, *grorbit* accepts values between 0 and 72. Values outside this range are ignored. *grorbitCov* term extends the *nodecov* term in the *ergm* package. In fact, *nodecov* term is a special case of *grorbitCov* where $grorbit = 0$.

3. **Graphlet Orbit Factor – *grorbitFactor(attrname, grorbit, base)*:** The *grorbitFactor* term includes the relation between the graphlet de-

grees and a categorical node attribute into the ERGM. The *grorbitFactor* term quantifies the link between a node category and graphlet degrees using a network statistic that is equal to the sum of the graphlet degrees of all nodes that are annotated with the corresponding category. The question answered by the change score function of this term is: "How does the category of a node relate with the change in the graphlet degrees?". This term has three arguments: (1) *attrname*, (2) *grorbit*, and (3) *base*. The *attrname* is a character vector giving the name of a categorical attribute in the network's node attribute list. The optional *grorbit* argument is a vector of distinct integers representing the list of graphlet orbits to include into the model (the complete list of graphlet orbits are illustrated in Figure 1.4). When *grorbit* is not provided, all graphlet orbits are evaluated by default; i.e., in *R* notation $grorbit = c(0, 72)$. The optional *base* argument is a vector of distinct integers representing the list of categories in *attrname* that are going to be omitted. When this argument is set to 0, all categories are evaluated. When this argument is set to 1, the category having the lowest value (or lexicographically first name) is eliminated. The term sorts all values of the categorical attribute lexicographically and *base* term defines the indexes of the categories to be omitted in this sorted list. For example, if the "fruit" attribute has values "orange", "apple", "banana" and "pear", *grorbitFactor("fruit" , 0 , 2:3)* will ignore the "banana" and "orange" factors and evaluate the "apple" and "pear" factors. When the *base* argument is not provided, the argument is set to 1 by default; i.e., the first category is omitted. The *grorbitFactor* term adds $a \times |grorbit|$ terms into the model where a represents the number of categories and $|grorbit|$ is the number of graphlet orbits to be evaluated in the model. Each term is equal to the sum:

$$grorbitFactor(G, i, X_c) = \sum_{v \in V, category(v) = X_c} C_i(G, v), \qquad (5.2)$$

where X_c is the category of the term, i is the queried graphlet orbit, $category(v)$ is the category that node v belongs to, and $C_i(G, v)$ is the number of graphlets that touch node v at graphlet orbit i. This term is defined for the 73 graphlet orbits corresponding to graphlets with

up to 5 nodes. Therefore, *grorbit* accepts values between 0 and 72. The values outside this range are ignored. *grorbitFactor* term extends the *nodefactor* term in the *ergm* package. In fact, *nodefactor* term is a special case of *grorbitFactor* where $grorbit = 0$.

4. **Graphlet Degree Distribution - *grorbitDist(grorbit, d)*:**
   The graphlet degree distributions of different graphlet orbits can be included into the ERGM by using the *grorbitDist* term. The question that the change score function of this term answers is: "How do the number of nodes having graphlet degree n for orbit i change when an edge is flipped?". This term has two arguments: (1) $grorbit$, and (2) d. The $grorbit$ argument is a vector of distinct integers representing the list of graphlet orbits to include into the model (the complete list of graphlet orbits are illustrated in Figure 1.4). The d argument is a vector of distinct integers, defining the graphlet degree values to take into consideration as model terms. This term adds one network statistic to the model for each pairwise combination of the arguments in $grorbit$ and d vectors. The statistic for the combination of (i, j) is equal to the number of nodes in the network that have graphlet degree j for orbit i. This term is defined for the 15 graphlet orbits corresponding to graphlets with up to 4 nodes. Therefore, $grorbit$ accepts values between 0 and 14. Graphlets of size 5 are omitted for this term due to the high computational demand of the change score computation of the term for 5-node graphlets. The *grorbitDist* term extend the *degree* term in the *ergm* package. In fact, *degree* term is a special case of the *grorbitDist* where $grorbit = 0$. However, the *grorbitDist* function does not support the filtering functionalities of the *degree* term that are defined with the *by* and *homophily* arguments.

5.3 Implementation

In this section, we explain the algorithmic details about the graphlet based ERGM terms in order to provide a deeper understanding on their properties and limitations. Since testing the correctness of the implementation for the new model terms is computationally challenging due to the integrated development with the *ergm* package of R, we summarize the tests that we

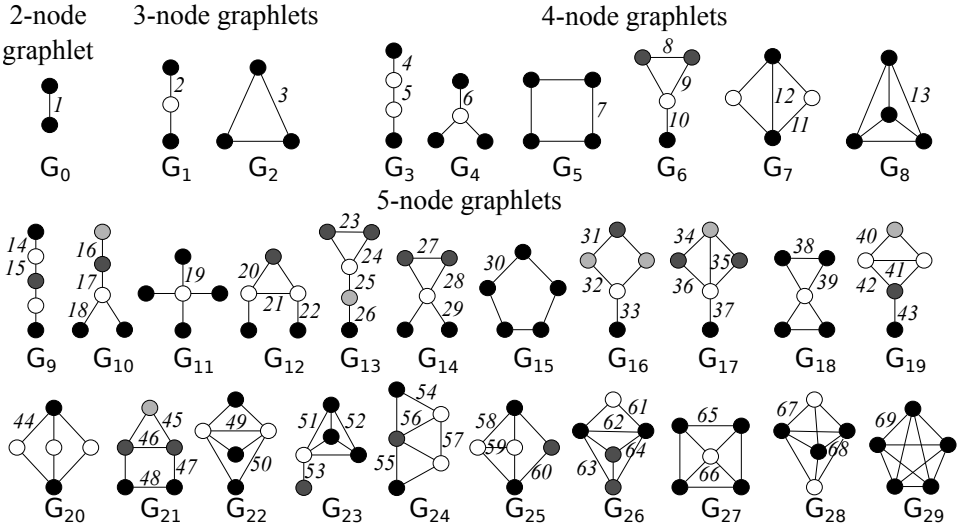


Figure 5.1: The 69 edge automorphisms of all 2- to 5-node graphlets.

used for validating the correctness of the new terms of our *ergm.graphlets* package.

5.3.1 Algorithmic Details

The four ERGM terms defined by the *ergm.graphlets* package are implemented using the *ergm.userterms* package [73]. The *ergm.userterms* package enables users to embed new modelling terms into the *ergm* package by implementing *C* code that calculates the change statistics of the new term. For the *ergm.graphlets* package, the main question that the change score function should answer is: "How do the graphlet counts in the network and graphlet degrees of the nodes change when an edge is flipped in the network?". This question can be answered efficiently by "touching" the graphlets on a flipped edge and counting only the graphlets that are going to be affected from the edge flip. This computation can be done by using the *edge automorphism orbits* in graphlets with 2, 3, 4 and 5 nodes (Figure 5.1) [178]. For clarity, we use the term *node orbits* for graphlet orbits that are provided in Figure 1.4 and *edge orbits* for edge automorphism orbits in Figure 5.1 throughout this section.

We apply a brute-force search algorithm for computing the change score for graphlet terms. For each flipped edge during the MCMC process of

ERGM parameter estimation, the edge orbits that are related with the queried graphlet structure are mapped on the flipped edge and the neighbourhood of that edge is searched for nodes that complete the graphlet structure. For each induced subgraph (i.e., node combination) that completes the graphlet structure, the count of the affected graphlets is incremented by one. The induced subgraphs that are of the same type with the queried graphlet, but turned into another graphlet by the edge flip are also identified and the count of the affected graphlets is decremented by one for each of these subgraphs. For identifying the change in the count of a specific graphlet, the computation is performed only for relevant edge orbits. The relations among graphlets and edge orbits are summarized in Table 5.1. For example, the change score for the counts of graphlet G_3 and G_5 can be calculated by counting edge orbits $\{4, 5, 7, 9, 12\}$. Let CE_i represent the number of graphlets counted by "touching" edge orbit i on the flipped edge. After counting the number of touched graphlets (i.e., CE_i) for all relevant edge orbits, the change score for graphlet G_3 is equal to $(CE_4 + CE_5 - CE_7 - CE_9)$ and the change score for G_5 is equal to $(CE_7 - CE_{12})$. By counting the graphlet change scores based on edge orbits, we both restrict the counting process to graphlets that are affected from the edge flip, and also avoid repeated counting of the same edge orbit for different graphlet counts. For instance, edge orbit 7 affects the count of G_3 negatively and the count of G_5 positively. With our implementation, the number of graphlets affected by edge orbit 7 are counted only once, and this change score is used for computing the changes in the counts of both G_3 and G_5.

The four ERGM terms of *ergm.graphlets* package are all implemented using graphlet counting based on edge orbits. The computation of the change scores differ slightly from each other depending on how the graphlet counts contribute to change statistics with these terms. The computation of the four terms of *ergm.graphlets* package are explained as follows:

1. ***graphletCount(g):*** For *graphletCount* term, the change score function directly reflects the change in the number of graphlets. For this reason, each identified graphlet directly increments (or decrements) the change score for the related graphlet by 1. The change score is computed by counting the graphlets for all edge orbits that are related with the graphlets provided in argument g. When all graphlets

Graphlet	Edge Automorphism		Graphlet	Edge Automorphism	
	Positive	Negative		Positive	Negative
$\mathbf{G_0}$	1	-	$\mathbf{G_{15}}$	30	46
$\mathbf{G_1}$	2	3	$\mathbf{G_{16}}$	31, 32, 33	35, 41, 44, 45
$\mathbf{G_2}$	3	-	$\mathbf{G_{17}}$	34, 35, 36, 37	49, 52, 54
$\mathbf{G_3}$	4, 5	7, 9	$\mathbf{G_{18}}$	38, 39	57
$\mathbf{G_4}$	6	8	$\mathbf{G_{19}}$	40, 41, 42, 43	51, 55, 60
$\mathbf{G_5}$	7	12	$\mathbf{G_{20}}$	44	50, 59
$\mathbf{G_6}$	8, 9, 10	11	$\mathbf{G_{21}}$	45, 46, 47, 48	56, 58
$\mathbf{G_7}$	11, 12	13	$\mathbf{G_{22}}$	49, 50	64
$\mathbf{G_8}$	13	-	$\mathbf{G_{23}}$	51, 52, 53	61
$\mathbf{G_9}$	14, 15	21, 24, 30, 32	$\mathbf{G_{24}}$	54, 55, 56, 57	63, 65
$\mathbf{G_{10}}$	16, 17, 18	20, 23, 28, 31	$\mathbf{G_{25}}$	58, 59, 60	62, 66
$\mathbf{G_{11}}$	19	27	$\mathbf{G_{26}}$	61, 62, 63, 64	67
$\mathbf{G_{12}}$	20, 21, 22	36, 40, 48	$\mathbf{G_{27}}$	65, 66	68
$\mathbf{G_{13}}$	23, 24, 25, 26	39, 42, 47	$\mathbf{G_{28}}$	67, 68	69
$\mathbf{G_{14}}$	27, 28, 29	34, 38	$\mathbf{G_{29}}$	69	-

Table 5.1: The complete list of edge orbit - graphlet associations. When evaluating the addition of an edge, positive associations increase the graphlet count since the graphlet is completed with the edge addition, while negative associations decrease the graphlet count since the considered graphlet turns into a different type of graphlet with the edge addition. This relation is reversed in the case of edge removal.

with the relevant edge orbits are counted, these counts are summed to get the overall change in the number of graphlets. For example, the change score for graphlet G_3 is equal to the summation of $(CE_4 + CE_5 - CE_7 - CE_9)$ where CE_i represents the number of graphlets that touch the flipped edge on edge orbit i.

2. ***grorbitCov(attrname, grorbit):*** The *grorbitCov* term relates a numeric node attribute with the graphlet degrees of the nodes according to Equation 5.1 as explained in Section 5.2. The change score function of this term is dependent on the change in graphlet degrees of nodes. Therefore, the nodes of each identified graphlet are linked to the node orbits that they correspond to. For example, when an edge (A, B) is added into network during the MCMC process and

a graphlet of type G_4 is formed by the induced subgraph on nodes $\{A, B, C, D\}$ (Figure 5.2), the change score for node orbit 6 is incremented by $Attr_A + Attr_C + Attr_D$, and the change score for node orbit 7 is incremented by $Attr_B$, where $Attr_v$ is the node attribute value for node v. The total change score of an edge flip is obtained by summing these attribute value changes from each graphlet identified by relevant edge orbits.

3. ***grorbitFactor(attrname, grorbit, base):*** The *grorbitFactor* term relates a categorical attribute with the graphlet degrees of nodes according to Equation 5.2 as explained in Section 5.2. As for *grorbitCov* term, the change score function of this term is dependent on the change in graphlet degrees of nodes and the nodes of each identified graphlet are linked to the node orbits. When the flip of an edge affects a node orbit, the change score that relates the category of the affected node with the node orbit is incremented (or decremented) by 1. For example, let an edge (A, B) be added into a network during the MCMC process and a graphlet of type G_4 is formed by the induced subgraph on $\{A, B, C, D\}$ (Figure 5.2). If node A and B belong to "Category 1", C and D belong to "Category 2", then change score for "Node Orbit 6, Category 1" and "Node Orbit 7, Category 1" will increase by 1 with the contribution of nodes A and B. The change score for "Node Orbit 6, Category 2" will increase by 2 with the increase in the graphlet degrees of nodes C and D. The final change score of an edge flip is obtained by summing these category - orbit pair score changes from each graphlet identified by relevant edge orbits.

4. ***grorbitDist(grorbit, d):*** The *grorbitDist* term identifies the change in the graphlet degree distribution of a node orbit when an edge is flipped during the MCMC process, as explained in Section 5.2. The change score computation of this term is slightly different from the three other *ergm.graphlets* terms: graphlet degrees of all nodes are needed at all steps of MCMC process of ERGM parameter estimation because the calculated change statistics is defined by the number of nodes that have a specific graphlet degree. In order to keep the computational complexity low, we compute the graphlet degree vectors (GDVs) of all nodes once at the beginning of the MCMC process. At

each step of the MCMC process, we update these GDVs using the change scores of edge flips. The changes in graphlet degrees of the nodes are identified similarly to the other terms: for each formed (or destroyed) graphlet with the edge flip, we identify the correspondence of the graphlet nodes to the node orbits, and update the GDVs of these nodes by increasing (or decreasing) the relevant graphlet degrees by 1. Since graphlets convert into each other with the edge flips during the MCMC procedure and in order to keep the graphlet degree vectors correct at all steps, the counting process should be applied to all edge orbits; i.e., it is not possible to restrict the counting procedure to edge orbits that are related with the query node orbits. This increases the computational complexity of *grorbitDist*, making the time required for the change score computation of 5-node graphlets prohibitive. Therefore, we implemented *grorbitDist* term only for 2-, 3-, and 4-node graphlets.

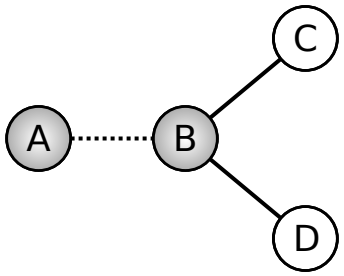


Figure 5.2: A small subgraph for illustrating the computation of *ergm.graphlets* terms. When edge (A, B) is added into the network, the subgraph forms a graphlet G_4 pattern. Nodes A, C, and D is linked to node orbit 6, and Node B is linked to node orbit 7 in this case.

Computational complexity of the change score computation based on edge orbits is dependent on the average degree (and therefore the density) of the modelled network. In average case, the computational complexity of the change score computation procedure is $O(d^2)$ where d represents the average degree of a node. The worst case scenario occurs when searching for graphlet G_9 in a clique. In this case, the computational complexity of the function is $O(n^3)$ where n is the number of nodes in the network. But

this situation occurs very rarely as most real-world networks are sparse.

5.3.2 Validation of the Implementation

Testing the correctness of the model terms that are implemented in the *ergm.graphlets* package is a challenging task, since the implementation is performed in an integrated manner to *ergm* package. Furthermore, this is the first graphlet identification implementation that relates the node attributes with the graphlet structures and there are no previous implementations that can be used for cross-checking the obtained statistics. For this reason, we developed some validation strategies for testing the correctness of the change score functions of *ergm.graphlets* package.

The first validation test for the correctness of the implemented change score functions uses the *summary* function of the *ergm* package. The *summary* function starts with an empty network, and adds the edges of the input network to the new network one-by-one, adding up the resulting change scores at each step. When all edges of the input network are added into the new network, the sum of all computed change scores should correspond to the exact model term statistics of the input network; e.g., the number of graphlets. We compare the *summary* function statistics of the new model terms of *ergm.graphlets* package with the graphlet statistics produced by the graphlet counting implementation of Pržulj et al. [156]. In this test, there are two indicators of a problem in the *ergm.graphlets* implementation: (1) a mismatch between the statistics obtained by the two implementations, and (2) inconsistent results over different runs of the *summary* function. The statistics of *graphletCount* and *grorbitDist* terms are directly comparable to the graphlet counts and GDVs produced by the implementation in [156]. Evaluating the correctness of the *grorbitCov* and *grorbitFactor* terms are slightly different as they are dependent on node attributes. In order to test the correctness of *grorbitCov* term, we first create a dummy node attribute, "dummy", that is equal to 1 for all nodes. By running the *summary* function of the *grorbitCov* term over the "dummy" node attribute, we obtain the sum of graphlet degrees of all nodes. We compare this sum with the sum of the graphlet degrees from the GDVs produced by the implementation in [156]. We repeat this test with weighted attribute values (e.g., when all values of "dummy" are set to 2) and confirm that the produced statistics are

scaled with the given weight. The validation for the *grorbitFactor* is similar to *grorbitCov* term: we create a categorical node attribute, "dummy", that assign the same category to all nodes and compute the change statistics of this category to obtain the sum of graphlet degrees of all nodes. When the category value is changed to another value, the output of the *summary* does not change for the *grorbitFactor* term.

A second test for validating the correctness of the *ergm.graplets* implementation is performed by running simulations on ERGMs that contain graphlet terms. In these tests, we define ERGMs containing an *edges* term and one graphlet term. We manually set the model coefficient for the graphlet term to various positive and negative values. We simulated 30 networks from each of these ERGM models; i.e., generated models that carry the properties defined by the model coefficients. With these simulations, we confirm that positive ERGM coefficients of graphlet terms promote the count of the related graphlet in the simulated networks. The count of related graphlet increase up to a certain coefficient value, until it reaches the maximum possible number of graphlets in the network. Similarly, negative coefficients have an effect of suppressing the appearance of the graphlet in the simulated networks. As the coefficient value gets closer to 0, the effect of the model term disappears. The range that the graphlet counts increase with the changing coefficient depends on the coefficients of the other terms in the ERGM model.

5.4 Case Studies

In this section, we illustrate the application of modelling terms from the *ergm.graphlets* package for the analysis of two different networks, one from the social sciences domain and one from the biological sciences domain.

5.4.1 Lake Pomona Emergent Multi-Organizational Network

Our first example comes from Thomas Drabek's [47] set of inter-organizational communication networks in the context of search and rescue operations. The setting for our example is the immediate aftermath of the capsizing of the Showboat Whippoorwill following its contact with a tornado near the southern shore of Lake Pomona, due south of Topeka, Kansas [47]. Sixty

passengers and crew were stranded in the lake, prompting the immediate response of the twenty organizations whose communication ties compose our network.

We use the graphlet terms to analyse patterns of brokerage in the organizational search and rescue network. Brokerage relations require (at least) three actors, one of whom bridges the connection between the two otherwise disconnected nodes (or sets of nodes, in extended brokerage structures) [67, 129]. The broker has the opportunity to mediate and facilitate exchanges between two parties, where the units exchanged may be goods, services, information, or any other transferable entities. Occupation of brokerage roles has been related to greater power in exchange networks [22] and control of information in inter-organizational disaster response networks [127]. Not all organizations are fit to occupy such roles, however, either by design or by happenstance [122, 127]. Previous studies of brokerage have been limited to the use of marginal tests to determine whether levels of brokerage exceed what we would expect by some baseline [67, 122, 127, 180]. The *ergm.graphlets* package enables us to examine brokerage using conditional tests in which we can identify entities' propensities to occupy brokerage roles independent of confounding factors such as degree. The *grOrbitFactor* and *grOrbitCov* terms allow us to determine whether occupation of local positions within graphlets is associated with particular covariates. These graphlet terms allow us to answer questions related to entities' local automorphism orbits (e.g., brokerage) in a model-based framework.

Drabek's Lake Pomona Emergent Multi-Organizational Network (EMON) dataset is found in the *network* package [23], which is automatically loaded alongside the *ergm* package [92]. Although the EMON network is represented as a digraph, the edge relations in the network are inherently undirected, as informants report on the existence of communication ties. We symmetrize the network via a union rule [106] to account for the undirected relations being measured. The nodes of the EMON network are associated with three node attributes. *Command rank score* is each organization's rating of how strong of a role it has in the network's chain of command, as reported by other organizations participating in the search and rescue effort. The *location* of each group's headquarters was also recorded; organizations were situated locally or non-locally in the Lake Pomona response. Finally, we include the *sponsorship level* of each organization: city, county, state,

federal, or private. When ranking those with the strongest role in the chain of command, informants were limited to the six organizations present from the early phase of the response. As a result, some organizations were not ranked and have been coded "NA" in the EMON data. For our example, we assume those who were not ranked have the lowest possible command rank score (being more marginal to the unfolding response) and assign them a score of 0. The resulting EMON network is illustrated in Figure 5.3.

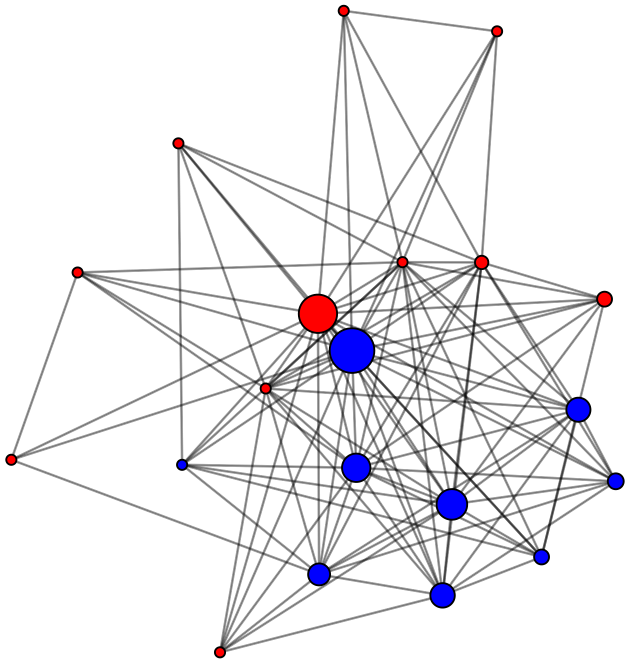


Figure 5.3: Lake Pomona emergent multi-organizational network (EMON) tasked with a search and rescue operation. Node size is scaled to command rank score and nodes are coloured by whether they had permanent headquarters situated locally (red) or non-locally (blue).

We model the EMON network using a combination of different model terms. We use the *edge* term to model the most fundamental network property: the number of edges in the network. One might expect that organizations at different sponsorship levels can be involved with more or fewer communication partnerships than organizations from a different sponsorship. Likewise, an organization's command rank score may be associated with its propensity to be involved in more communication partnerships.

We include these properties into the ERGM by the dyadic independence terms, *nodefactor* and *nodecov*, for the sponsorship level and command rank score attributes of nodes. Finally, we model the core-periphery structure of the network using the graphlet-based terms of our *ergm.graphlets* package. Graphlet G_6, which involves brokerage between cliques and individual nodes, is a natural choice for modelling the core-periphery structure of the EMON network, and we include all its automorphism orbits—9, 10, and 11—into our model. We incorporate the location covariate (i.e., node attribute) into the term to evaluate whether an organization's location is associated with its propensity to occupy these specific automorphism orbits. The modelling results are expected to demonstrate whether the location of an organization in this type of subgraph is associated with its role as a pendant (orbit 9), member of a dyad with ties to a broker (orbit 10), or broker between the pendant and the dyad (orbit 11).

We estimate the model parameters of the ERGM defined by these terms for the EMON network, and validate that the Monte Carlo Maximum Likelihood Estimation procedure for model parameters converge properly as described in [90]. The estimated ERGM for the EMON network is summarized in Model 1.

The results show significant effects for our *edge* term, command rank score, and non-local organizations' occupation of orbit 11. The results show a strong, positive association between an organization's command rank score and its odds of forming a tie. Most relevant to our interests, we find that one of the automorphism orbit terms is significant. We find a positive, significant association between non-local (NL) organizations and their propensity to occupy a brokerage role between a pendant and a dyad (automorphism 11). Substantively, this demonstrates that non-local organizations tend to occupy this specific structure of extended brokerage in which an organization occupies a brokerage position between one organization and a pair of connected organizations.

As explained in Section 1.5.3, AIC and BIC scores of models can be used for assessing the trade-off between model complexity and goodness-of-fit. When we compare the AIC and BIC scores of Model 1 with the baseline model (i.e., the model that only contains the *edges* term), we observe substantial improvements – we obtain lower AIC and BIC scores, although Model 1 contains more parameters. We further assess the goodness-of-fit for

```
R > summary(emon.ergm)

==========================
Summary of model fit
==========================

Formula:   emon.3 ~ edges + nodefactor("Sponsorship") +
nodecov("Command.Rank.Score") + grorbitFactor("Location", c(9:11))

Iterations:  20

Monte Carlo MLE Results:
                               Estimate Std. Error MCMC %  p-value
edges                         -2.450670   0.688351      9 0.000473 ***
nodefactor.Sponsorship.County -0.437354   0.319080      3 0.172175
nodefactor.Sponsorship.Federal -0.581708   0.606596      5 0.338852
nodefactor.Sponsorship.Private -0.041876   0.188267      1 0.824230
nodefactor.Sponsorship.State  -1.326516   0.785447      1 0.092967 .
nodecov.Command.Rank.Score     0.333315   0.075229      5  < 1e-04 ***
grorbitFactor.orb_9.attr_NL    0.009319   0.020540      0 0.650596
grorbitFactor.orb_10.attr_NL  -0.018051   0.014288      2 0.208081
grorbitFactor.orb_11.attr_NL   0.158800   0.031310      7  < 1e-04 ***
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 263.4  on 190  degrees of freedom
 Residual Deviance: 144.8  on 181  degrees of freedom

AIC: 162.8    BIC: 192    (Smaller is better.)
```

Model 1: ERGM model that is estimated for the EMON dataset based on terms, including the *grorbitFactor*.

the estimated ERGM by generating networks from the ERGM using maximum likelihood estimation and comparing them with the EMON network based on four different network properties: degree distribution, shortest-path length (geodesic) distance distribution, edge-wise shared partner distribution (i.e., the distribution of ep_k values for all $k \leq |V|$, where ep_k is the number of unordered, connected node pairs that have exactly k common neighbours), and the triad census (i.e., the distribution of 3-node subgraphs formed by all node triples in the network). Figure 5.4 illustrates the fit of the estimated ERGM on the EMON network based on these four network properties. As there are no clear discrepancies between the model-simulated networks and the original network, we find the ERGM to be an adequate fit.

The graphlet orbit terms enable us to link local position to covariates in a model-based framework. As demonstrated, this is a useful tool for modelling brokerage as we are able to link an entity's covariates to its propensity to occupy a specific brokerage role, whether it is a traditional (i.e., two-path) brokerage role or an extended brokerage role (e.g., orbit 11 in our ERGM). Beyond brokerage, these techniques can extend to any particular automorphism orbit contained within a graphlet: pendants, clique members, or other nodes whose position may be linked to some categorical or continuous variable. Being able to incorporate these covariate-driven graphlet terms into a model-based framework will enhance our ability to understand which factors are associated with nodes' occupation of local positions within graphlets.

5.4.2 Protein Secondary Structure Network

The past decade has seen a surge of interest in identifying network motifs whose size often ranges three to five nodes. Applications span a wide variety of networks including transcription networks [3, 139, 140, 171, 204], neuron synaptic connection networks [100, 139, 140], protein-protein interaction networks [2, 3, 139, 204], circuitry networks [100, 139, 140], worldwide web networks [139, 140], language networks [139], and social networks [139]. Typically scholars have used marginal tests to identify how frequently these subgraphs occur relative to some baseline. In these types of tests the observed network is compared a set of randomized networks that hold constant some statistic of the original network, often the degree distribution. While

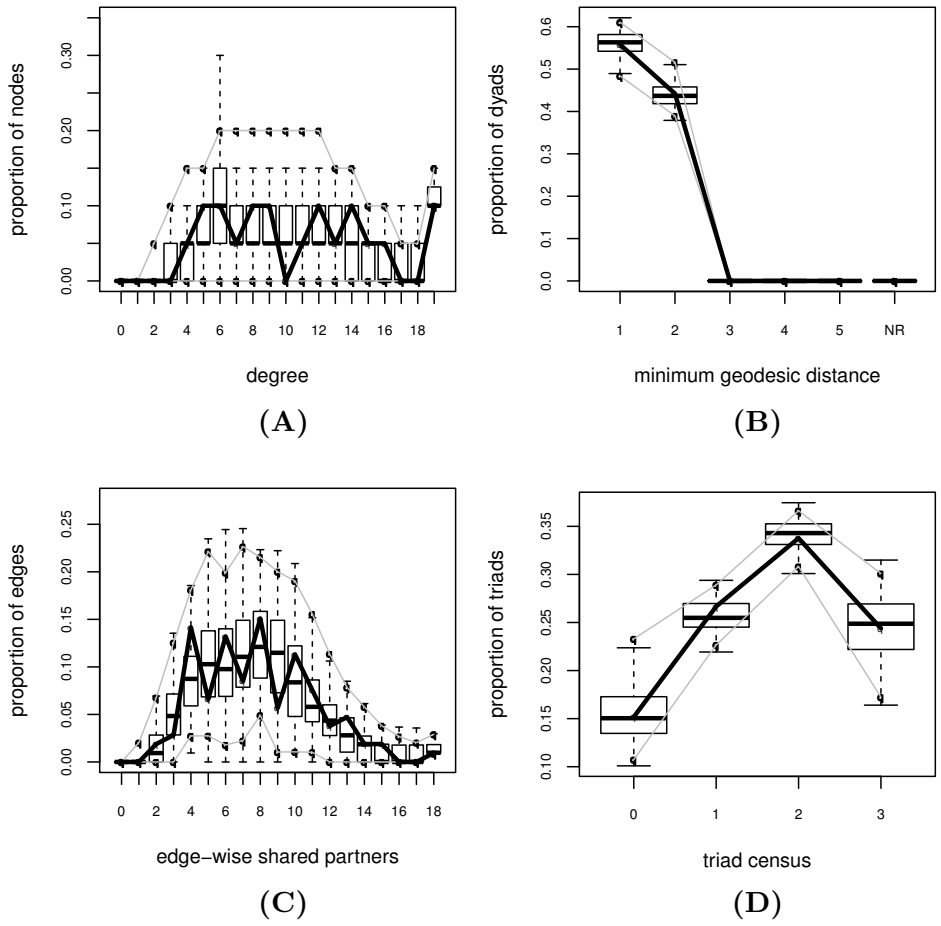


Figure 5.4: The goodness-of-fit test results of the ERGM estimated for the EMON data. The panels illustrate the results for the tests of: A – degree distribution, B – shortest path length distribution, C – edge-wise shared partner distribution, and D – triad census. The solid black line in each plot represents the EMON's observed statistics. The box plots illustrate the statistics for our simulated networks, as produced by the MLE.

these types of marginal tests have been employed by networks scholars for decades [24, 198], a model-based approach allows us to examine the likelihood of observing these graphlets, conditioned on a variety of parameters; e.g., degree, triadic closure, covariates. This is particularly important where the method of data collection itself may bias structure in particular ways; failure to account for these effects may result in spurious findings. In this section, we use the *graphletCount* terms to examine patterns of biological network motifs in an ERGM framework, while controlling for artefacts of the data collection process.

We model a protein structure network whose nodes are secondary structure elements (specifically, α helices and β sheets) which are connected if the distance between them is smaller than 10 Angstroms (Å) [139]. This network represents the proximity structure of a matriptase-aprotinin complex (PDB ID:1eaw) [59] as determined by x-ray crystallography (resolution 2.93Å). Milo et al. [139] examine the overrepresentation and underrepresentation of subgraphs in this network, by comparison to uniform random graphs conditional on the degree distribution. They find that subgraphs in the form of graphlets G_3 and G_4 are underrepresented while subgraphs in the form of graphlets G_6, G_7, and G_8 are overrepresented (see Figure 1.4). We will determine whether these results hold in a model-based framework that allows us to account for potentially confounding degree, transitivity, and mixing effects, some of which represent artefacts of the data collection process.

The structure of the matriptase-aprotinin complex contains two assemblies, each of which is a complex of two proteins (the catalytic domain of matriptase/MT-SP1 and a bovine pancreatic trypsin inhibitor/BPTI) [59]. The presence of multiple copies of a biologically relevant complex within a crystal structure is a common artefact of the crystallization process, and indeed the same system could potentially have been observed with more or fewer complexes in the asymmetric unit. This is of considerable importance for modelling the resulting network, as we would typically expect far more adjacencies within complexes than between them; failure to control for this effect may lead to very misleading conclusions. Indeed, as shown in Figure 5.5, the network is dominated by two dense subgraphs corresponding to the two complexes, with very few ties spanning these subgraphs. To account for this, we create node attributes based on biological assembly member-

ship as reconstructed from information in the Protein Data Bank [59], with polypeptide chains A and B of the structure belonging to assembly 1, and chains C and D belonging to assembly 2. By incorporating these attributes into the model, we are much better able to account for the patterns of clustering in the network than we would be if we neglected the data collection process.

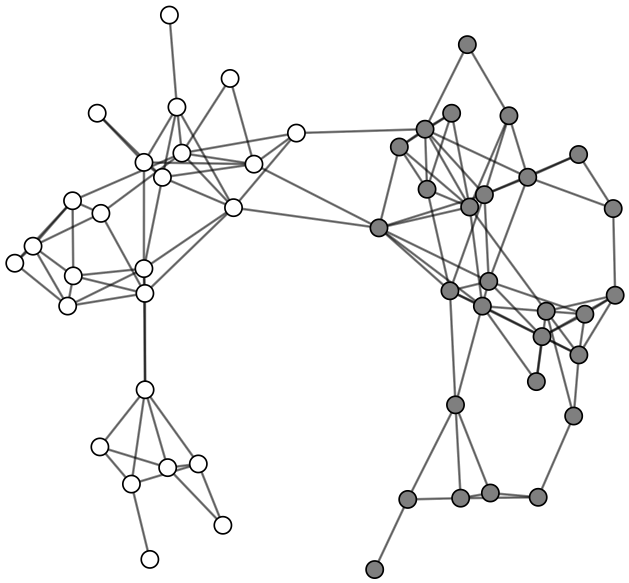


Figure 5.5: Network representation of the protein structure of the two matriptase-BPTI complexes. Secondary structure elements are shaded by the complex to which they belong.

We start modelling by first setting up our model with an *edge* term, a dyadic independence term, and several dyadic dependence terms, including our graphlet terms. As we observe very little tie formation across the sets of chains associated with each complex, we include a homophily term (i.e., *nodeMatch*) for protein assembly in our model. Additionally, we include a within-assembly triadic closure term (i.e., closure of triads where all members belong to the same assembly – *triangle*). We also include a degree term (i.e., *gwdegree*) as [139] was concerned with graphlet counts net of the degree distribution. Of principal interest is our *graphletCount* term, which includes graphlets G_3, G_4, G_6, G_7, and G_8, the same set [139] finds to occur at greater or lesser levels than chance. We estimate the model pa-

rameters of the ERGM defined by all of these terms, and validate that the Monte Carlo Maximum Likelihood Estimation procedure for model parameters converge properly as described in [90]. The estimated ERGM for the protein structure network of matriptase-aprotinin complex is summarized as in Model 2.

```
R> summary(spi.ergm.34678)

==========================
Summary of model fit
==========================

Formula:   spi ~ edges + nodematch("Assembly") + triangle("Assembly") +
gwdegree(0.5, fixed = T) + graphletCount(c(3, 4, 6, 7, 8))

Iterations:  20

Monte Carlo MLE Results:
                   Estimate Std. Error MCMC %  p-value
edges              -6.42760    1.22926     12  < 1e-04 ***
nodematch.Assembly  2.48031    0.74204      6 0.000852 ***
triangle.Assembly   3.87343    0.67331      1  < 1e-04 ***
gwdegree            2.40227    1.51019      5 0.111906
graphlet.3.Count    0.04962    0.02964      7 0.094298 .
graphlet.4.Count   -0.03917    0.05467      1 0.473841
graphlet.6.Count   -0.15361    0.04993      0 0.002137 **
graphlet.7.Count   -0.47295    0.17782      0 0.007910 **
graphlet.8.Count   -2.49869    0.72543      0 0.000590 ***
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 1910.3  on 1378  degrees of freedom
 Residual Deviance:  593.9  on 1369  degrees of freedom

AIC: 611.9    BIC: 658.9    (Smaller is better.)
```

Model 2: The first ERGM model that is estimated for the protein structure network based on terms, including the *graphletCount* term.

Our model finds a significant, positive effect for within-assembly homophily (as represented by the significance of *nodematch.Assembly* term), a positive effect for triadic closure within complexes (as represented by the significance of *triangle.Assembly* term), and a propensity for the graph to be biased against formation of graphlets G_6, G_7, and G_8, assuming all other terms are held constant. We find no significant results for graphlets G_3 and G_4.

As explained in Section 1.5.3, models containing less parameters are preferred over more complex models, and the trade-off between the model complexity and goodness-of-fit can be assessed using the AIC and BIC scores.

We remove the non-significant terms of Model 2, and test whether we can obtain a simpler model with a better fit. AIC suffers slightly if we remove G_3 from the model (AIC: 612.97), while BIC improves (654.8). Both improve if we keep G_3 and remove G_4 (AIC: 610.73, BIC: 652.56). We find the best fit by removing *both* G_3 and G_4 (AIC: 610.7, BIC: 647.3). Accordingly, we fit our final model as shown in Model 3.

```
R> summary(spi.ergm.all)

==========================
Summary of model fit
==========================

Formula:   spi ~ edges + nodematch("Assembly") + triangle("Assembly") +
 gwdegree(0.5, fixed = T) + graphletCount(c(6, 7, 8))

Iterations:  20

Monte Carlo MLE Results:
                   Estimate Std. Error MCMC %  p-value
edges              -4.80106    0.73658      8  < 1e-04 ***
nodematch.Assembly  2.11636    0.66232      5 0.001428 **
triangle.Assembly   3.27864    0.53805      0  < 1e-04 ***
gwdegree            1.12902    1.21795      1 0.354095
graphlet.6.Count   -0.12037    0.04122      2 0.003560 **
graphlet.7.Count   -0.46225    0.16905      0 0.006330 **
graphlet.8.Count   -2.31074    0.68949      0 0.000826 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

     Null Deviance: 1910.3  on 1378  degrees of freedom
 Residual Deviance:  596.7  on 1371  degrees of freedom

AIC: 610.7    BIC: 647.3    (Smaller is better.)
```

Model 3: The second, simplified ERGM model that is estimated for the protein structure network based on terms, including the *graphletCount* term.

Once again we find positive, significant effects for homophily within complexes (*nodematch.Assembly*) and triadic closure within complexes (*triangle.Assembly*). Controlling for this, we find negative, significant effects for graphlet terms G_6, G_7, and G_8. Our final model appears to have converged without any notable issues [90].

We assess our final model for its goodness of fit. As Figure 5.6 shows, our model closely approximates the observed protein structure network of matriptase-aprotinin complex; our simulated networks show no clear deviations from the observed statistics on degree distribution, shortest-path length (geodesic) distance distribution, edge-wise shared partner distribu-

tion, or the triad census.

It is interesting to compare the results of our joint, multivariate analysis with the marginal tests conducted by [139]. Milo et al. find that the network overrepresents graphlets G_6, G_7, and G_8 and underrepresents G_3 and G_4. After controlling for other factors (particularly clustering within each complex), we find no evidence of additional underrepresentation or overrepresentation of G_3 or G_4; further, we actually find that the network appears biased against formation of graphlets G_6, G_7, and G_8, once other terms are accounted for. The discrepancy here is due to the use of marginal tests by [139]. To determine whether a graphlet occurs more or less often relative to chance, they compare the number of observed graphlets to the number observed in a set of random graphs conditioned on the degree distribution (a form of conditional uniform graph test). For this protein structure network, such random graphs bear little resemblance to the data in question (Figure 5.7), and in particular do not include effects related to the fact that the structure is a composite of two distinct complexes. While this does not make the results of such tests wrong per se, it does render them unable to distinguish between structural biases arising from simple features produced by the data collection process, and those arising from more subtle and informative biochemical mechanisms. The marginal approach is also unable to unravel the *joint* influence of multiple biases simultaneously; because graphlet structures are dependent upon one another, over or underrepresentation of multiple graphlets (relative to a uniform baseline) may actually be the result of biases to a smaller number of features. Such complexities are difficult to unravel using marginal tests, and are more flexibly handled via the ERGM framework.

Our analysis underscores the fact that one can obtain misleading conclusions when trying to use marginal tests to assess graphlet counts, particularly when the baseline distribution being employed does not incorporate extremely basic features of the studied system. While inference for complex, highly dependent systems is difficult under the best of conditions, the generative nature of the ERGM framework allows us to assess the adequacy of our models by comparison to features of the original data; given that we have identified a model that is both sensible and that successfully regenerates the important properties of the observed network, we have a stronger basis for subsequent investigation than would be obtained from simple rejection of a

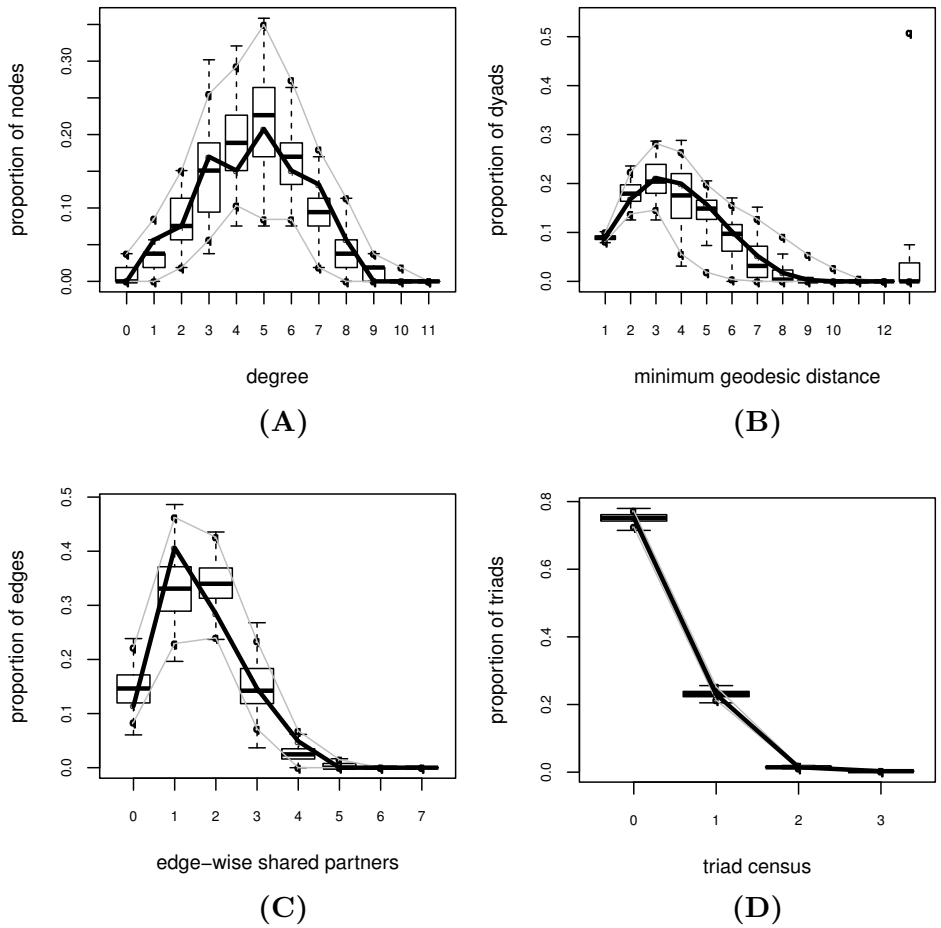


Figure 5.6: The goodness-of-fit test results of the ERGM estimated for the protein structure network of matriptase-aprotinin complex. The panels illustrate the results for the tests of: A – degree distribution, B – shortest path length distribution, C – edge-wise shared partner distribution, and D – triad census. The solid black line in each plot represents the protein structure network's observed statistics. The box plots illustrate the statistics for our simulated networks, as produced by the MLE.

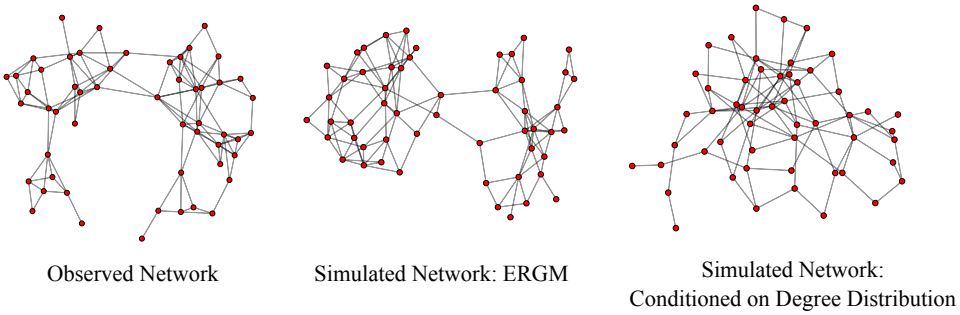


Figure 5.7: Observed protein structure network (left), typical protein structure network simulated by our final model (middle), and typical random network produced by holding the input network's degree distribution constant (right).

null hypothesis.

By using an ERGM approach and incorporating our graphlet terms, we are able to produce more sophisticated models of protein networks that include not only network motifs but also other important biological and/or chemical properties of the system in question. Scholars in a variety of biological sub-disciplines have begun to use ERGMs to model many different types of networks, including protein-protein interaction networks [20, 31], neural networks [83, 173, 174], and metabolic networks [166]. Introducing the tools from the *ergm.graphlets* package to the network community should enhance the field's ability to model graphlet counts in the context of network motifs or any other application where one is interested in counts of small, undirected, induced subgraphs.

5.5 Model Degeneracy, Instability, and Sensitivity

Model degeneracy, instability and sensitivity are currently the main challenges in network modelling within the ERGM framework [74, 169]. For some combination of model terms, the Markov Chain - Monte Carlo (MCMC) procedure may fail to converge to appropriate model coefficients (θ) within a reasonable number of iterations. This is generally because the network distribution associated with the specified model family are ill-behaved. Like most dependence terms, the terms in the *ergm.graphlets* package sometimes suffer from these instability issues depending on the input network and

the other terms in the ERGM. Typically, model degeneracy problems are currently handled either by using user-selected terms whose effects partially cancel (e.g., using sparse graphlets and complete graphlets together) or using curved exponential family models [25, 91, 169] that systematically combine large numbers of terms in a manner that balances their total effect.

The former technique requires some intuition about the topology of the input network and a number of trials with different combinations of terms under this intuition. It can be hard to identify the best terms for generating an ERGM and there is no general solution that works well in all settings. Our experience suggests that graphlet terms for which the change score is non-zero for most of the steps in the MCMC procedure are good candidates with which to start the modelling process. For example, it is not reasonable to model a sparse network using clique-like graphlets, as the change score will be 0 for most of the MCMC steps. In this respect, the graphlet terms that are expected to be overrepresented in the network can also be good candidate terms to start ERGM modelling. Using terms of the same graphlet size together usually improves the convergence of the MCMC process, since smaller graphlets might already be contained in a number of larger graphlets and this causes dependency issues among the model terms. The list of graphlet orbit dependencies in Table 2.1 can be useful for choosing independent model terms. We also observed that MCMC procedure converges faster when graphlets containing closed-loop structures (e.g., triangles, cycles) are excluded from the ERGM: this is mainly because of the instability of these terms, as explained in [169].

Past work with (partial) subgraph terms has suggested that curved exponential family models can also be used for improving degeneracy issues. In curved exponential families, parameters associated with model statistics are constrained to lie on a non-linear surface of reduced dimension, forcing them to remain in a fixed relationship with one another; this can be helpful when dealing with intrinsically correlated graph statistics, as very precise weighting may be needed to avoid the degenerate regime. Examples of curved terms include the *gwdegree*, *gwdsp*, and *gwesp* terms of the *ergm* package, as well as the closely related *alternating k-star* and *alternating path* statistics of [177]. Because graphlet statistics do not "nest" in the same way as non-induced subgraph statistics, they may benefit from novel formal development. On the other hand, some ideas used in existing curved families –

e.g., geometrically weighted degree distributions – could potentially be applied to graphlet orbit degrees in a relatively straightforward manner. This would seem to be a promising direction for future research.

5.6 Author's Contributions

Ömer Nebil Yaveroğlu collaborated with Sean M. Fitzhugh, Maciej Kurant, Athina Markopoulou, Carter T. Butts, and Nataša Pržulj for the work presented in this chapter.

In this collaboration, Ömer Nebil Yaveroğlu implemented the *ergm.graphlets* package after the idea is initiated by Dr. Carter Butts. Ömer designed the algorithms for the efficient, but exact computation of the change scores for graphlet-based terms, and also thoroughly tested the implementation for possible implementation errors. Ömer also wrote the first version of the paper, "ergm.graphlets: A package for ERG modelling based on graphlet statistics", which is in submission.

6 Conclusion

In this chapter, we provide a brief summary of our results and contribution in this dissertation. We conclude the dissertation by presenting some future directions that our graphlet correlations based methodology can be applied on.

6.1 Summary of the Dissertation

During the past decade, graphlet properties have been widely applied for the analysis of networks; in particular, for contrasting structural similarities among networks as well as for identifying topologically similar nodes in networks. Though graphlet based methods are shown to be successful, there is still room for improving these techniques because: (1) current methods do not accurately filter out the effects of redundancies and dependencies among the graphlet degrees of nodes, and (2) the computational complexity of the graphlet identification procedure makes these techniques impractical for analysing very large networks (e.g., social networks containing thousands of nodes and millions of edges such as the Facebook network). These limitations reduce the accuracy and applicability of the graphlet based techniques.

Keeping these limitations in mind, we propose graphlet based solutions to two fundamental graph theoretic problems: (1) topological network comparison problem, and (2) network modelling problem. Topological network comparison problem aims to quantify structural similarities between two networks, without any intention of producing a node mapping that highlights these similarities. The similarity scores identified by the solutions to this problem have been used in tracking the topological changes in a network, identifying topologically similar network pairs to enable the transfer of knowledge between them, and evaluating the fit of alternative network models on an input network. On the other hand, the network modelling problem aims to identify rules that govern the formation and evolution of a

network in a topological context. By identifying well-fitting network models, it is possible to understand the structural organisation in a network, evaluate the effects of some edge-formation rules on the topology of a network, and mine the correspondence between the node and edge characteristics with the observed patterns of links. These two problems are not completely independent in the sense that, in order to evaluate the fit of a network model to a network, we need to compare the topologies of the networks generated from the model with the input network. Therefore, accurate and efficient graphlet based network comparison techniques are needed for solving both problems.

First, in order to define such a graphlet based method without suffering from the above listed limitations, we identify all the redundancies and dependencies in the graphlet degree vectors (GDVs) of nodes. The redundancies in the GDVs arise from the fact that combinations of smaller graphlets form the larger graphlets in a number of different configuration possibilities. Therefore, graphlet degrees of some orbits can be derived from weighted linear combinations of the graphlet degrees of other orbits. Considering all possible combinations of smaller graphlets, we identify 26 orbit redundancy equations; for which 17 are independent and 9 of them can be derived from the combinations of the 17 others. This means that we can eliminate the graphlet degrees of 17 orbits from GDVs, one from each independent equation, without losing any topological information encoded in the graphlet degree vectors. This elimination obviously does not reduce the computational cost of identifying graphlets, as the information content of the non-redundant orbits is identical to the complete set of orbits. However, the elimination helps us to define more accurate distance measures without using any redundant graphlet degree information. Even with the elimination of redundancies, there still exist dependencies among the remaining graphlet orbits. These dependencies are caused by the appearance of smaller graphlets in larger ones. Since the counts of the smaller graphlets limit the counts of larger dependent graphlets, the graphlet degrees of dependent orbits are expected to be correlated.

We discover that investigating the dependencies among orbits is a very powerful way of analysing the structure of a network. We quantify the dependencies among the graphlet degrees of all orbit pairs using the Spearman's Correlation Coefficient. The existence of positive correlations for

the dependent orbit pairs is expected, but what summarizes the network's complex topology is the correlations among the independent orbits. Exploiting the correlations among orbits, we propose a new network topology statistic, named Graphlet Correlation Matrix, that is an 11×11 symmetric matrix in which each cell (i, j) corresponds to the Spearman's Correlation Coefficient among the graphlet degrees of orbits i and j of all nodes in the network. Furthermore, we exploit this new network statistic to quantify the topological similarities of two networks, by defining the Graphlet Correlation Distance (GCD). Graphlet Correlation Distance is the Euclidean distance between the upper triangular values of the Graphlet Correlation Matrices of two networks. We validate this new network distance measure by testing its model identification performance on synthetic networks that are generated from seven different random models; i.e., ER, ER-DD, SF-BA, SF-GD, GEO, GEO-GD, and STICKY. Based on this set of models, we performed the first systematic comparison of the model identification performances among the state-of-the-art network distance measures. In these tests, GCD outperforms all other methods, even when it is defined based on the statistics of 2- to 4-node graphlets. Moreover, we validated that GCD is highly noise-tolerant both for networks containing false interactions and also for networks with missing interactions (i.e., incomplete networks). The computational cost of GCD is also less than all other graphlet based network distance measures, as it performs better than those methods even when using only the statistics of 2- to 4-node graphlets (without the need for identifying the 5-node graphlet statistics in a network).

Second, we apply our new graphlet correlation based methods for the analysis of the world trade networks. Instability of the world economy and the recent financial crises urges the researchers to understand the functional mechanisms in these complex systems better. International world trade is one of the major factors that shape the world economy. Graph theoretic analysis of the complex world trade system can shed light onto possible sources of malfunctioning in this system. In this respect, we first analyse the graphlet correlation matrices of the world trade networks. The correlations observed in these matrices show that the world trade network has a three-layer organisation: the layers of core (i.e., densely connected), broker (i.e., mediators among disconnected nodes) and periphery (i.e., weakly connected to the rest of the network). The core and broker layers are softly separated,

while the separation between these two layers and the periphery layer is more strict; i.e., countries do not appear both in the periphery and the core/broker layers at the same time. We continue by analysing the dynamic changes in the world trade network topology over time using the GCD. In particular, since the crude oil price is one of the most important wealth indicators of the world financial system, we identify the correlations between the topological changes in world trade network and the changes in crude oil price. According to this analysis, the changes in the crude oil price change the topology of the world trade network in 1 to 2 years, but not vice versa. To understand the nature of the change in the topology, we analyse the graphlet count changes during crisis periods. We observe that during all global recessions, weakly connected graphlets (e.g., G_5, G_{15}, G_{16}, G_{20}) first deteriorate when entering the crises, and then recover after the crises. The counts of the densely connected or broker type graphlets do not change during the crises. Next, we analyse the correspondence between a country's network position and its wealth by computing the canonical correlation coefficients of graphlet degrees and economic wealth indicators. This analysis shows that among the three layers, the brokerage position is the strongest indicator of a country's wealth, and that peripheral position is strongly associated with poverty. This observation gives us the idea of defining brokerage and peripheral scores based on the graphlet degrees of relevant orbits, in order to track the change of a country's position in the world trade network over the years. Tracking these two scores, we find that: (1) the brokerage scores of well developed countries perfectly reflect the changes in their economies, (2) the peripheral scores of the developing countries match well with the economic crises that they experienced, and (3) accession of developing countries to European Union make them more peripheral in the world trade network.

Third, we focus on modelling five different types of networks from different domains: (1) autonomous networks, (2) Facebook networks, (3) metabolic networks, (4) protein structure networks, and (5) world trade networks. In these analyses, we evaluate the correspondence between the model networks and the input networks using the GCD, and evaluate which of the seven network models fit to these network types. We identify that: (1) autonomous networks are best modelled by the ER-DD model although the fit of this model is also not strong, (2) Facebook, metabolic and protein structure net-

works are all well-fit by the SF-GD, GEO, and GEO-GD models showing the resemblance between these networks from the social and biological domains for the first time, and (3) none of the seven network models fit the world trade networks. Due to the last observation, we propose two new generative random network models that are dedicated to modelling the world trade networks. The first, Gravitational Random model, has its roots from the well-known Gravity model of trade, but defined first time as a random generative model. The second, the brokerage model, is a completely graphlet dependent network model that aims to maximize the number of graphlet G_{23} in a random network. While both of these models approximate the topology of world trade networks well, the brokerage model shows a better fit as identified by smaller GCDs. We extend the analysis on world trade networks further, by analysing the three graphlet degrees from graphlet G_{23} (due to its success in modelling the world trade networks), and show that a country's economic wealth indicators are predictive of its future brokerage position.

Finally, being motivated by the success of the graphlet based brokerage model, we develop a generic framework for network modelling using any of the graphlet properties of a network. We exploit the exponential-family random graph models (ERGMs) in this respect, and embed graphlet statistics based modelling terms in the *ergm* package, which enables network analysis within an ERGM framework. Our modelling terms not only integrate the statistics of the number of appearances of each graphlet and graphlet degree distributions of the nodes with the ERGM framework, but also enable defining ERGMs that evaluate the association of a graphlet pattern with node attributes. We illustrate the application of our new network modelling framework by successfully defining ERGMs for networks from two different domains: (1) a social network representing an inter-organizational communication network, and (2) a biological network representing the tertiary protein structure of a protein.

Since our methods are solely based on the graphlet properties of the networks, they have endless application domains. In their current state, our methodology is specific to the analysis and modelling of undirected and unweighted networks, since graphlets are only defined for simple graphs. However, the idea of graphlet correlations are easily extendible to handle any type of network, expanding their applicability to a wider range of network

types. Even in their current state, these methods are successfully applied for the analysis of world trade networks, which naturally appear in the form of weighted and directed networks, giving insights into their organisational principles and their changes during the times of crises. This is only the first example, illustrating the power of our techniques in untangling the complexity of even such sophisticated networks. Furthermore, we exploit the descriptive power of graphlets in network modelling, illustrating their success in summarizing the network characteristics and reproducing these characteristics randomly. We believe that graphlet based network modelling methods will be fancied by network analysts, as they ease the exploration of any type of relational structure in a statistical context.

6.2 Future Directions

In this section, we present four ideas on the applications of our new methodology and show some preliminary results on these ideas.

6.2.1 Phylogeny Reconstruction from Metabolic Network Similarities

Metabolic networks explain the chemical reactions that occur in a cell. Given the complete map of reactions from all species, metabolic networks differ among different species with respect to the genes and gene products that catalyse these reactions. If a gene is expressed in a species, then this gene product works as an enzyme for some reactions within the species' cell, so the elements (i.e., metabolites and enzymes) of these reactions are included into the species' metabolic network. In this respect, phylogenetically similar species are expected to have similar metabolic network topologies. So far, the phylogenetic similarities among species are studied based only on the sequence similarities and phenotypical similarities. Investigating whether the network topology contain some extra information to uncover about phylogenetic similarities is an open question. As we have shown in Chapter 2, Graphlet Correlation Distance is the best network distance measure for identifying the topological similarities among networks. In this respect, it would be a good solution for the metabolic network comparison problem.

We obtain and construct the metabolic networks of all species in the form of enzyme – enzyme interactions from KEGG database [97] as explained in Section 1.2. For each species in KEGG, we identify the phylogenetic kingdom, phylum, class, order, family, and genus from NCBI Taxonomy database [54]. If the phylogenetic classification information of a species is not included in the NCBI Taxonomy database, we exclude those species from our experiment. Our main hypothesis is: Metabolic networks of species with similar phylogenies should have similar topologies. We test this hypothesis by computing the graphlet correlation distances among all species and plotting the Receiver-Operator Characteristic (ROC) curves obtained from the comparison of GCD distances according to the 6 phylogenetic classes; i.e., genus, family, order, class, phylum, kingdom (phylogenetic classes are ordered from most specific to most generic). Figure 6.1 presents the resulting ROC curves and the corresponding AUC scores. We observe that smaller GCD distances are observed among phylogenetically more similar species, as evident with the high AUC score obtained for the Genus level.

These results encourage us to investigate the graphlet correlation distances among metabolic networks of different species. A first step in this investigation would be understanding the organizational differences in the metabolic networks of different phylogenetic groups, based on the orbit clustering patterns observed in their graphlet correlation matrices. The homogeneity of the classes defined at the Genus level could also be further analysed in order to identify the genus groups that have inconsistent topologies. Investigating the possible causes for these inconsistent genus groups is another open research question that can be investigated by the graphlet correlation distances.

6.2.2 Uncovering Topological Disease - Pathway Similarities

There is a recently increasing interest on studying diseases in terms of the pathways associated with them rather than individual gene associations [8, 49, 114, 120]. We hypothesize that the topological disease - pathway similarities may reveal novel relations that might lead to disease gene predictions and insights into drug targeting. In this respect, we investigate the topological similarities among disease and pathway genes in the protein-protein interaction (PPI) network.

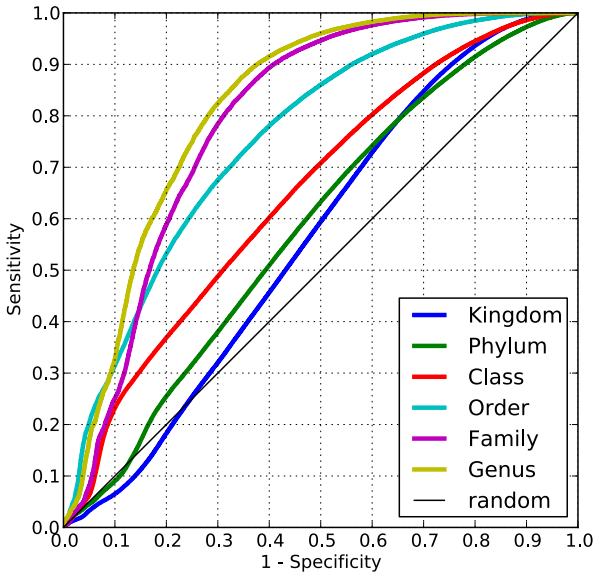


Figure 6.1: ROC curves representing the performance of metabolic network distances in identifying phylogenetic classes. The AUC scores for the six phylogenetic classes are: genus – 0.820, family – 0.793, order – 0.756, class – 0.651, phylum – 0.584, and kingdom – 0.563 (ordered from most specific to most generic).

We obtain the PPIs of human proteins from BioGRID Database (version 3.2.101 - June 2013), resulting with a network containing 110,528 interactions among 13,276 proteins. Disease–gene associations are obtained from DisGeNet database (version 2.0) [11], which integrates disease-gene associations that are available at UniProt [5], OMIM [70], Comparative Toxicogenomics Database (CTD) [130], Genetic Association Database (GAD) [12], Mouse Genome Database (MGD) [50], and Literature-derived Human Gene-Disease Network (LGHDN) [21] databases. Although the DisGeNet database contains both curated and predicted associations, we focus only on the curated ones; resulting with $28,287$ associations among $5,493$ diseases and $6,936$ genes. Finally, pathway-gene associations for human pathways are obtained from KEGG database (Release 66.1 - downloaded on 17.06.2013) [97].

We define the topological profile of a disease or pathway using graphlet

correlation matrices that are computed from the graphlet degree vectors (obtained from the human PPI network) of the genes associated with the disease or pathway. We compute the graphlet correlation matrices of 233 diseases and 217 pathways that are associated with ≥ 20 genes in DisGeNet and KEGG databases. We quantify the topological similarity among these diseases and pathways by computing Euclidean distances between the upper triangular values of their graphlet correlation matrices; i.e., their graphlet correlation distances. Evaluating the validity of the topological disease–disease and disease–pathway similarities is challenging. Some alternative methods for validating the similarity of a disease pair are: (1) the number of shared genes, (2) the number of shared drugs, (3) commorbidity – the frequency of two diseases being observed on a person at the same time, and (4) correlated expression profiles in genome-wide association studies [8, 49, 114, 120]. Disease–pathway relations can be similarly evaluated with methods 1, 2, and 4.

We construct two networks that encode the distances among the 233 diseases and 217 pathways on edge weights: (1) a bipartite network, that is constructed by computing the pairwise GCDs among disease–pathway pairs. This network contains $233 \times 217 = 50,561$ weighted edges among 450 nodes. (2) a complete network, that is constructed by computing pairwise GCDs among all pathways and diseases, including disease–disease and pathway–pathway comparisons. This network contains $\binom{450}{2} = 101,025$ weighted edges among 450 nodes. Constructing the two networks, we encode a huge amount of topological similarity information among pathways and diseases into a single network. These two weighted networks need to be mined in detail for uncovering novel disease–pathway, or even disease–disease relations; e.g., by applying weighted network clustering algorithms such as the affinity propagation clustering [58]. The uncovered relations would give insights into disease gene prediction, and drug repositioning problems; e.g., knowing that pathway P is topologically similar to disease D, one can obtain the drug-pathway associations from KEGG database [97], and reconsider using the drugs that are effective on pathway P for possible cures on disease D. Similarly, such drug predictions can be made from the uncovered disease-disease relations. Furthermore, after clustering the weighted networks, the highly shared genes in the obtained clusters can be tested for being associated with the diseases in the cluster. Evaluating the topological characteristics

of diseases and pathways from the perspective of their graphlet correlation matrices is another open research question, which might give insights into understanding the positioning of diseases in protein interaction networks.

Interestingly, graphlet correlation distances do not only consider the topological similarities, but also the number of shared genes between two diseases/pathways as a side effect. If the number of shared genes is high for a pair of diseases/pathways, then it is expected that the graphlet correlation matrices will be similar for these pairs. This is something desirable since the high number of shared genes indicates similar positioning in the protein interaction network. Nevertheless, the most novel disease – pathway associations are the ones for which the number of shared genes are low, but the topological similarity identified by the graphlet correlation distances is high. Identifying these disease – pathway pairs is another data mining problem that will uncover novel relations among diseases and pathways.

6.2.3 Improvements on the Graphlet Degree Vector Similarities of Nodes

Apart from using the graphlet properties of networks for the quantification of topological network similarities, these properties can also be used for identifying the topologically similar nodes in a network. Milenkovic et al. [138] proposed the graphlet degree vector (GDV) similarity measure that compares the Graphlet Degree Vectors of nodes to quantify topological node similarities. Given the graphlet degree vectors of two nodes, C_u and C_v, GDV similarity is the weighted and normalized absolute difference of all orbits in their GDVs. The weighting is performed based on the number of dependencies of each orbit, o_i, where the orbit dependencies are defined as in Table 2.1. In particular, the GDV similarity of nodes u and v are computed as:

$$w_i = 1 - \frac{\log(o_i)}{\log(73)},$$

$$D_i(u,v) = w_i \times \frac{|\log(C_u[i]+1) - \log(C_v[i]+1)|}{\log(\max\{C_u[i], C_v[i]\}+2)},$$

$$D(u,v) = \frac{\sum_{i=0}^{72} D_i}{\sum_{i=0}^{72} w_i},$$

$$S(u,v) = 1 - D(u,v), \tag{6.1}$$

where $C_u[i]$ represents the graphlet degree of node u for orbit i, and o_i is the number of orbits that orbit i is dependent on (including the orbit itself). GDV similarity, $S(u, v)$, is a real number in $[0, 1]$, where 0 represents that the nodes are topologically different and 1 represents that the GDV's are identical. This node similarity measure is shown to be successful in identifying the melanogenesis-related cancer genes [138], and also guided four different algorithms for producing high quality network alignments [109, 110, 133, 137].

Although GDV similarity is a very successful measure for identifying the node similarities, there is still room for improving the way the node similarities are computed from the graphlet properties of nodes. First of all, in Section 2.2, we show that the graphlet degrees of 17 orbits are redundant in the GDVs of nodes. These orbits can be removed from the node similarity computation, in order to have a better node similarity measure. Second, the performance of GDV similarities that are obtained by including the orbits of 5-node graphlets are not systematically compared with the results obtained by excluding the 5-node graphlet orbits. Inclusion of 5-node graphlet orbits increase the computational complexity of computing node similarities, as explained in Section 2.5. If the performance of the GDV similarity is comparable (or even better in the case of Graphlet Correlation Distances), then this would lead to another important improvement on the computation of node similarities. Finally, the weighting function of GDV similarities can be redefined based on the graphlet correlation matrix of the network. For example, in Figure 2.6-C, we observe that orbits 2 and 5 are perfectly correlated, and their correlations with the other orbits are extremely similar. This means that the information contained in the graphlet degrees of these orbits contribute almost identically to the GDV distance among nodes. This type of redundant information can be eliminated from GDV similarity by re-weighting all orbits based on the similarities of their correlation profiles.

We also consider adapting the graphlet correlation distances to compare the topological similarities among nodes. In this approach, we define the topological profile of a node as the graphlet correlation matrix that is defined by the GDV of the node and its neighbours. Then the topological distance between two nodes is defined as the graphlet correlation distance of their topological profiles. However, this approach has a limitation: For the Spearman's Correlation Coefficients to be meaningful, the node should

have a minimum number of neighbours (e.g., a minimum of 20 neighbours) so that the change in the graphlet degrees can be observed. This restricts the number of nodes for which the graphlet correlation matrix can be defined, and so the graphlet correlation distances to the other nodes. Still, the graphlet correlation matrices of the nodes can provide a simplified description of the topological organization around a node. Implementing these ideas, and validating the performance of them is among the future directions of our methodology.

6.2.4 Integration of Graphlet Correlation Distances with ergm package

Statnet [72] is a collection of packages that allows exponential-family random graph modelling (ERGM). It is a flexible framework that enables defining network models based on any choice of network properties (more details provided in Section 1.5.2). Apart from the built-in modelling terms, we also introduced some new graphlet-based model terms into this package, as explained in Chapter 5. Within the wide range of network models that can be defined based on any network property, evaluating the fit of a network model to a network is a challenging task, since it requires network comparison (see Section 1.4). In *Statnet* package, the fit of a network model to a network is tested based on four different types of network properties: (1) shortest path length distribution, (2) edge-wise shared partner distribution (i.e., the distribution of number of node pairs that are connected with an edge and share $\{0, 1, 2, \ldots, |V| - 2\}$ neighbours), (3) degree distribution, and (4) triad census distribution (i.e., the proportion of 3-node sets that have 0, 1, 2, 3 edges among them) [92]. Once an ERGM model is estimated for an input network, a number of networks from this model are simulated, the above listed network properties of simulations are computed, and these network properties are summarized by the quartile statistics on a plot. The network properties of the input network are plotted against the simulation statistics for evaluating their agreement (as illustrated in Figure 5.4 and 5.6).

Although this model evaluation approach successfully identifies various network characteristics that are different among the model networks and the input network, it has two shortcomings: (1) the comparison does not

produce any quantified statistics that would help to choose among alternative ERGMs, and (2) the tested network properties are not detailed enough to capture any subgraph pattern properties other than the triangles. In this respect, we believe that comparing the graphlet properties of the networks will be an important contribution to *Statnet* package for model evaluation.

The graphlet based model-fitting comparisons can be done based on distributions as in the built-in model-fitting tests of *Statnet*. The two types of graphlet-based distributions that can be used in this way are: (1) The distribution of the 30 graphlets, and (2) the 73 distributions of graphlet degree of each orbit of 2- to 5-node graphlets (i.e., orbits 0-72). Although this would provide a more detailed evaluation of models, this technique still does not quantify the similarity of a model to an input network. Therefore, this evaluation would be a good technique for evaluating the fit of a single ERGM, but not for comparing alternative ERGMs for the best fit.

The second group of techniques that consists of the RGF distance, the GDD-Agreement and the Graphlet Correlation Distance, fills this shortcoming. Based on the averages and standard deviations of the above listed network distances between the input network and the model networks, one can choose which network model would be the best possible explanation for the observed structure of input network. Among these alternative distances, graphlet correlation distance is of particular importance, as it is shown to outperform the others in terms of model identification and it has lower computational complexity than the other methods (Chapter 2). We believe that these additional model-fitting tests will improve the capabilities of the *Statnet* package, though there is still the need for integrating the implementation for these model-fitting tests with the *Statnet* package.

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] I. Albert and R. Albert. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346–3352, 2004.

[3] U. Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

[4] J. E. Anderson. A theoretical foundation for the gravity equation. *The American Economic Review*, 69(1):106–116, 1979.

[5] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. ODonovan, N. Redaschi, and L. L. Yeh. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 32(Suppl. 1):D115–D119, 2004.

[6] C. Arellano. Default risk and income fluctuations in emerging economies. *American Economic Review*, 98(3):690–712, 2008.

[7] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone. Comment on "Network Motifs: Simple Building Blocks of Complex Networks" and "Superfamilies of Evolved and Designed Networks". *Science*, 305(5687):1107, 2004.

[8] A. L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[9] A. L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.

[10] S. Battiston, J. F. Rodrigues, and H. Zeytinoglu. The network of inter-regional direct investment stocks across Europe. *Advances in Complex Systems*, 10(1):29–51, 2007.

[11] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. A. Mayer, F. Sanz, and L. I. Furlong. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS ONE*, 6(6):e20284, 06 2011.

[12] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang. The genetic association database. *Nature Genetics*, 36(5):431–432, 2004.

[13] M. Berka, M. B. Devereux, and C. Engel. Real exchange rate adjustment in and out of the Eurozone. *American Economic Review*, 102(3):179–85, 2012.

[14] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[15] L. Biggiero and M. Basevi. Testing the gravity model through network analysis. *Paper presentato al Convegno*, 2009.

[16] B. Bollobás. *Random graphs*, volume 73. Cambridge University Press, 2001.

[17] M. Boss, H. Elsinger, M. Summer, and S. Thurner. Network topology of the interbank market. *Quantitative Finance*, 4(6):677–684, 2004.

[18] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[19] K. R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatics*, 21(9):2076–2082, 2005.

[20] S. Bulashevska, A. Bulashevska, and R. Eils. Bayesian statistical modelling of human protein interaction network incorporating protein disorder information. *BMC Bioinformatics*, 11(1):46, 2010.

[21] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, and H. P. Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9(1):207, 2008.

[22] R. Burt. *Structural Holes*. Harvard University Press, Cambridge, MA, 1992.

[23] C. T. Butts. network: a package for managing relational data in R. *Journal of Statistical Software*, 24(2):1–36, 2008.

[24] C. T. Butts. Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, 11(1):13–41, 2008.

[25] C. T. Butts. Bernoulli graph bounds for general random graphs. *Sociological Methodology*, 41(1):299–345, 2011.

[26] W. D. Bygrave. The structure of the investment networks of venture capital firms. *Journal of Business Venturing*, 3(2):137–157, 1988.

[27] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154, 2007.

[28] C. Chen. Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, 35(3):401–420, 1999.

[29] J. M. Cherry, C. Adler, C. Ball, S. A. Chervitz, S. S. Dwight, E. T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, et al. SGD: Saccharomyces genome database. *Nucleic Acids Research*, 26(1):73–79, 1998.

[30] T. Y. Choi, K. J. Dooley, and M. Rungtusanatham. Supply networks and complex adaptive systems: control versus emergence. *Journal of Operations Management*, 19(3):351–366, 2001.

[31] N. R. Clark, R. Dannenfelser, C. M. Tan, M. E. Komosinski, and A. Ma'ayan. Sets2Networks: network inference from repeated observations of sets. *BMC Systems Biology*, 6(1):89, 2012.

[32] R. Clark and J. Beckfield. A new trichotomous measure of world-system position using the international trade network. *International Journal of Comparative Sociology*, 50(1):5–38, 2009.

[33] S. Collins, P. Kemmeren, X. Zhao, J. Greenblatt, F. Spencer, F. Holstege, J. Weissman, and N. Krogan. Toward a comprehensive atlas of the physical interactome of Saccharomyces Cerevisiae. *Molecular and Cellular Proteomics*, 6(3):439–450, 2008.

[34] U. N. Comtrade. United nations commodity trade statistics database. *URL: http://comtrade.un.org*, 2010. Data downloaded on 15.11.2011.

[35] S. A. Cook. The complexity of theorem-proving procedures. In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, pages 151–158. ACM, 1971.

[36] M. Costanzo, A. Baryshnikova, J. Bellay, Y. Kim, E. D. Spear, C. S. Sevier, H. Ding, J. L. Koh, K. Toufighi, S. Mostafavi, et al. The genetic landscape of a cell. *Science*, 327(5964):425–431, 2010.

[37] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, London, 1994.

[38] D. Croft, G. OKelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Suppl. 1):D691–D697, 2011.

[39] A. Cukierman and M. Tommasi. When does it take a Nixon to go to China? *American Economic Review*, 88:180–197, 1998.

[40] F. Curtis. Peak globalization: Climate change, oil depletion and global trade. *Ecological Economics*, 69(2):427–434, 2009.

[41] M. R. Darby. The price of oil and world inflation and recession. *The American Economic Review*, 72(4):738–751, 1982.

[42] A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly. Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical–

gene–disease networks. *Nucleic Acids Research*, 37(Suppl. 1):D786–D792, 2009.

[43] L. De Benedictis and L. Tajoli. The world trade network. *The World Economy*, 34(8):1417–1454, 2011.

[44] F. Della Rossa, F. Dercole, and C. Piccardi. Profiling core-periphery network structure by random walkers. *Scientific Reports*, 3:1467, 2013.

[45] I. Destler. US trade policy-making in the eighties. In *Politics and Economics in the Eighties*, pages 251–284. University of Chicago Press, 1991.

[46] S. J. Dixon, M. Costanzo, A. Baryshnikova, B. Andrews, and C. Boone. Systematic mapping of genetic interaction networks. *Annual Review of Genetics*, 43(1):601–625, 2009.

[47] T. E. Drabek, H. L. Tamminga, T. S. Kilijanek, and C. R. Adams. *Managing Multiorganizational Emergency Responses: Emergent Search and Rescue Networks in Natural Disaster and Remote Area Settings*. University of Colorado Institute of Behavioral Science, Boulder, CO, 1981.

[48] M. Duenas and G. Fagiolo. Modeling the international-trade network: a gravity approach. *Journal of Economic Interaction and Coordination*, 8:155–178, 2013.

[49] H. Eleftherohorinou, V. Wright, C. Hoggart, A.-L. Hartikainen, M.-R. Jarvelin, D. Balding, L. Coin, and M. Levin. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PloS ONE*, 4(11):e8068, 2009.

[50] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, and the Mouse Genome Database Group. The mouse genome database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Research*, 40(D1):D881–D886, 2012.

[51] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[52] G. Fagiolo, J. Reyes, and S. Schiavo. World-trade web: Topological properties, dynamics, and evolution. *Physical Review E*, 79:036115, 2009.

[53] A. Falicov and F. E. Cohen. A surface of minimum area metric for the structural comparison of proteins. *Journal of Molecular Biology*, 258(5):871–892, 1996.

[54] S. Federhen. The NCBI taxonomy database. *Nucleic Acids Research*, 40(D1):D136–D143, 2012.

[55] O. Frank and D. Strauss. Markov graphs. *Journal of American Statistical Association*, 81(395):832–842, 1986.

[56] C. Freund. Current account adjustment in industrial countries. *Journal of International Money and Finance*, 24(8):1278–1298, 2005.

[57] C. Freund. The trade response to global downturns. *Research Working papers*, 1(1):1–30, 2009.

[58] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[59] R. Friedrich, P. Fuentes-Prior, E. Ong, G. Coombs, M. Hunter, R. Oehler, D. Pierson, R. Gonzalez, R. Huber, W. Bode, and E. L. Madison. Catalytic domain structures of MT-SP1/Matriptase, a matrix-degrading transmembrane Serine Proteinase. *The Journal of Biological Chemistry*, 277(2):2160–2168, 2002.

[60] I. M. Fund. World economic outlook database. *Washington (DC): International Monetary Fund*, 2006. Online: accessed October-2012.

[61] S. Gama-Castro, V. Jiménez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Peñaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muñiz-Rascado, I. Martínez-Flores, H. Salgado, et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic Acids Research*, 36(Suppl. 1):D120–D124, 2008.

[62] D. Garlaschelli and M. I. Loffredo. Structure and evolution of the world trade network. *Physica A: Statistical Mechanics and its Applications*, 355(1):138–144, 2005.

[63] M. R. Gary and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. WH Freeman and Company, New York, 1979.

[64] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A.-M. Michon, C.-M. Cruciat, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

[65] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[66] S. M. Goodreau, M. S. Handcock, D. R. Hunter, C. T. Butts, and M. Morris. A Statnet tutorial. *Journal of Statistical Software*, 24(9):1–26, 2008.

[67] R. V. Gould and R. M. Fernandez. Structure of mediation: A formal approach to brokerage in exchange networks. *Sociological Methodology*, 19:89–126, 1989.

[68] L. H. Greene and V. A. Higman. Uncovering network systems within protein structures. *Journal of Molecular Biology*, 334(4):781–791, 2003.

[69] J. Hair and R. Anderson. *Multivariate data analysis*. Prentice Hall Higher Education, 2010.

[70] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52–55, 2002.

[71] J.-D. J. Han, D. Dupuy, N. Bertin, M. E. Cusick, and M. Vidal. Effect of sampling on topology predictions of protein-protein interaction networks. *Nature Biotechnology*, 23(7):839–844, 2005.

[72] M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1548, 2008.

[73] M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris. ergm.userterms: User-specified terms for the Statnet suite of packages. *Journal of Statistical Software*, 52(2):1–25, 2013.

[74] M. S. Handcock, G. Robins, T. Snijders, J. Moody, and J. Besag. Assessing degeneracy in statistical models of social networks. *Journal of the American Statistical Association*, 76:33–50, 2003.

[75] G. T. Hart, A. K. Ramani, E. M. Marcotte, et al. How complete are current yeast and human protein-interaction networks. *Genome Biol*, 7(11):120, 2006.

[76] W. Hayes, K. Sun, and N. Pržulj. Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, 29(4):483–491, 2013.

[77] J. He and M. W. Deem. Structure and response in the world trade network. *Physical Review Letters*, 105(19):198701, 2010.

[78] D. Headey and S. Fan. Anatomy of a crisis: the causes and consequences of surging food prices. *Agricultural Economics*, 39(S1):375–391, 2008.

[79] C. Hertz-Fowler, C. S. Peacock, V. Wood, M. Aslett, A. Kerhornou, P. Mooney, A. Tivey, M. Berriman, N. Hall, K. Rutherford, et al. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Research*, 32(Suppl. 1):D339–D343, 2004.

[80] A. Heston, R. Summers, and B. Aten. PENN world table. https://pwt.sas.upenn.edu/, 2002. Online: accessed November-2011.

[81] C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis. A dynamic network approach for the study of human phenotypes. *PLoS Computational Biology*, 5(4):e1000353, 2009.

[82] D. Higham, M. Rasajski, and N. Pržulj. Fitting a geometric graph to a protein-protein interaction network. *Bioinformatics*, 24(8):1093–1099, 2008.

[83] M. Hinne, T. Heskes, C. F. Beckmann, and M. A. van Gerven. Bayesian inference of structural brain networks. *NeuroImage*, 66(0):543 – 552, 2013.

[84] D. A. Hojman and A. Szeidl. Core and periphery in networks. *Journal of Economic Theory*, 139(1):295–309, 2008.

[85] P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs: Rejoinder. *Journal of the American Statistical Association*, 76(373):62–65, 1981.

[86] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233(1):123–138, 1993.

[87] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[88] F. Hormozdiari, P. Berenbrink, N. Pržulj, and S. C. Sahinalp. Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution. *PLoS Computational Biology*, 3(7):e118, 2007.

[89] G. Hu and P. Agarwal. Human disease-drug network based on genomic expression profiles. *PLoS One*, 4(8):e6536, 2009.

[90] D. R. Hunter, S. M. Goodreau, and M. S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258, 2008.

[91] D. R. Hunter and M. S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3):565–583, 2006.

[92] D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, 24(3):1–29, 2008.

[93] G. Iori, G. De Masi, O. V. Precup, G. Gabbi, and G. Caldarelli. A network analysis of the italian overnight money market. *Journal of Economic Dynamics and Control*, 32(1):259–278, 2008.

[94] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.

[95] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences*, 97(3):1143–1147, 2000.

[96] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[97] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[98] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, 2002.

[99] P. D. Karp, C. A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahrén, S. Tsoka, N. Darzentas, V. Kunin, and N. López-Bigas. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research*, 33(19):6083–6089, 2005.

[100] N. Kashtan and U. Alon. Spontaneous evolution of modularity and network motifs. *Proceedings of National Academy of Sciences*, 102(39):13773–13778, 2005.

[101] T. Kastelle, J. Steen, and P. Liesch. Measurig globalisation: an evolutionary economic approach to tracking the evolution of international trade. In *DRUID Summer Conference on Knowledge, Inno-*

vation and Competitiveness: Dynamics of Firms, Networks, Regions and Institutions-Copenhagen, Denmark, June, pages 18–20, 2006.

[102] R. Kelley and T. Ideker. Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology*, 23(5):561–566, 2005.

[103] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muñiz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, et al. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Research*, 39(Suppl. 1):D583–D590, 2011.

[104] C. P. Kindleberger. Government policies and changing shares in world trade. *The American Economic Review*, 70(2):293–298, 1980.

[105] E. D. Kolaczyk. *Statistical Analysis of Network Data*. Springer-Verlag, Boston, 1st edition, 2009.

[106] D. Krackhardt. Cognitive social structures. *Social Networks*, 9(2):109–134, 1987.

[107] C. J. Krieger, P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee, and P. D. Karp. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research*, 32(Suppl. 1):D438–D442, 2004.

[108] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. P. Tikuisis, et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006.

[109] O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, 7(50):1341–1354, 2010.

[110] O. Kuchaiev and N. Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, 2011.

[111] O. Kuchaiev, M. Rašajski, D. J. Higham, and N. Pržulj. Geometric denoising of protein-protein interaction networks. *PLoS Computational Biology*, 5(8):e1000454, 2009.

[112] M. Kurant, M. Gjoka, C. T. Butts, and A. Markopoulou. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, volume 11, pages 281–292. ACM, 2011.

[113] M. Lappe and L. Holm. Unraveling protein interaction networks with near-optimal efficiency. *Nature Biotechnology*, 22(1):98–103, 2004.

[114] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, 4(11):e1000217, 2008.

[115] C. E. Leiserson, R. L. Rivest, C. Stein, and T. H. Cormen. *Introduction to Algorithms*. The MIT Press, 2nd edition, 2001.

[116] J. Leskovec. Stanford large network dataset collection, 2011. URL: http://snap. stanford. edu/data/index.html.

[117] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1(1):5, 2007.

[118] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[119] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[120] Y. Li and P. Agarwal. A pathway-based view of human diseases and disease relationships. *PloS One*, 4(2):e4346, 2009.

[121] A. Lin, R. T. Wang, S. Ahn, C. C. Park, and D. J. Smith. A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome Research*, 20(8):1122–1132, 2010.

[122] B. E. Lind, M. Tirado, C. T. Butts, and M. Petrescu-Prahova. Brokerage role in disaster response: Organisational mediation in the wake of hurricane katrina. *International Journal of Emergency Management*, 5(1/2):75–99, 2008.

[123] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Information Processing and Management*, 41(6):1462–1480, 2005.

[124] F. Luo, J. Z. Wang, and E. Promislow. Exploring local community structures in large networks. *Web Intelligence and Agent Systems*, 6(4):387–400, 2008.

[125] N. Malod-Dognin, R. Andonov, and N. Yanev. Maximum cliques in protein structure comparison. In *Experimental Algorithms*, pages 106–117. Springer, 2010.

[126] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[127] C. S. Marcum, C. A. Bevc, and C. T. Butts. Mechanisms of control in emergent interorganizational networks. *The Policy Studies Journal*, 40(3):516–546, 2012.

[128] G. Maria Milesi-Ferretti and A. Razin. Sharp reductions in current account deficits an empirical analysis. *European Economic Review*, 42(3):897–908, 1998.

[129] P. V. Marsden. Brokerage behavior in restricted exchange networks. volume 7, pages 341–410. Sage: Beverly Hills, CA, 1982.

[130] C. J. Mattingly, M. C. Rosenstein, A. P. Davis, G. T. Colby, J. N. Forrest, and J. L. Boyer. The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicological Sciences*, 92(2):587–595, 2006.

[131] V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, et al. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003.

[132] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *Advances in Neural Information Processing Systems 25*, pages 548–556, 2012.

[133] V. Memišević and N. Pržulj. C-GRAAL: Common-neighbors-based global graph alignment of biological networks. *Integrative Biology*, 4(7):734–743, 2012.

[134] H.-W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Research*, 30(1):31–34, 2002.

[135] T. Milenković, I. Filippis, M. Lappe, and N. Pržulj. Optimized null model for protein structure networks. *PLoS One*, 4(6):e5967, 2009.

[136] T. Milenković, V. Memišević, A. K. Ganesan, and N. Pržulj. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *Journal of The Royal Society Interface*, 7(44):423–437, 2010.

[137] T. Milenković, W. L. Ng, W. Hayes, and N. Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer informatics*, 9:121–137, 2010.

[138] T. Milenković and N. Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:257–273, 2008.

[139] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[140] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[141] D. Mitchell. A note on rising food prices. *World Bank Policy Research Working Paper No. 4682*, 2008.

[142] R. A. Mundell. A reconsideration of the twentieth century. *American Economic Review*, 90(3):327–340, 2000.

[143] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–540, 1995.

[144] U. Nations. Conference on trade and development (UNCTADstat) database. URL: http://unctadstat.unctad.org. Accessed: 03/11/2012.

[145] M. Newman. *Networks: an introduction*. Oxford University Press, 2009.

[146] M. E. Newman. The structure and function of networks. *Computer Physics Communications*, 147(1):40–45, 2002.

[147] S. L. Ooi, D. D. Shoemaker, and J. D. Boeke. DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nature Genetics*, 35(3):277–286, 2003.

[148] J. Osborne, J. Flatow, M. Holko, S. Lin, W. Kibbe, L. Zhu, M. Danila, G. Feng, and R. Chisholm. Annotating the human genome with disease ontology. *BMC Genomics*, 10(Suppl. 1):S6, 2009.

[149] R. Overbeek, N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov, K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, et al. The ERGOTM genome analysis and discovery system. *Nucleic Acids Research*, 31(1):164–171, 2003.

[150] P. Pattison and S. Wasserman. Logit models and logistic regressions for social networks: Ii. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169–193, 1999.

[151] M. Penrose. *Random geometric graphs*, volume 5. Oxford University Press, 2003.

[152] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371, 2003.

[153] V. Piana. The "pattern" approach to world trade structures and their dynamics. In *Princeton Institude for International and Regional Studies – Observing Trade: Revealing International Trade Networks*, 2006.

[154] E. S. Prasad and R. G. Rajan. Modernizing China's growth paradigm. *American Economic Review*, 96(2):331–336, 2006.

[155] T. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, et al. Human protein reference database – 2009 update. *Nucleic Acids Research*, 37(Suppl. 1):D767–D772, 2009.

[156] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[157] N. Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[158] N. Pržulj and D. J. Higham. Modelling protein–protein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, 3(10):711–716, 2006.

[159] N. Przulj, O. Kuchaiev, A. Stevanović, and W. Hayes. Geometric evolutionary dynamics of protein interaction networks. In *Pacific Symposium on Biocomputing*, pages 178–189, 2010.

[160] J. Quiggin. Interpreting globalization: Neoliberal and internationalist views of changing patterns of the global trade and financial system. *United Nations Research Institude for Social Development (UNRISD)*, Overarching Concerns: 7:1–45, 2005.

[161] R Development Core Team. R: A language and environment for statistical computing. *Vienna Austria R Foundation for Statistical Computing*, 1(10):ISBN 3–900051–07–0, 2008.

[162] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*, 17(10):1030–1032, 1999.

[163] T. Rito, Z. Wang, C. M. Deane, and G. Reinert. How threshold behaviour affects the use of subgraphs for network comparison. *Bioinformatics*, 26(18):i611–i617, 2010.

[164] G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173–191, 2007.

[165] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

[166] Z. M. Saul and V. Filkov. Exploring biological network structure using exponential random graph models. *Bioinformatics*, 23(19):2604–2611, 2007.

[167] F. Schacherer, C. Choi, U. Götze, M. Krull, S. Pistor, and E. Wingender. The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics*, 17(11):1053–1057, 2001.

[168] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[169] M. Schweinberger. Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496):1361–1370, 2011.

[170] M. A. Serrano and M. Boguná. Topology of the world trade web. *Physical Review E*, 68(1):015101, 2003.

[171] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, 31(1):64–68, 2002.

[172] N. Simonis, J.-F. Rual, A.-R. Carvunis, M. Tasan, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, J. M. Sahalie, K. Venkatesan, F. Gebreab, et al. Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nature Methods*, 6(1):47–54, 2008.

[173] S. L. Simpson, S. Hayasaka, and P. J. Laurienti. Exponential random graph modeling for complex brain networks. *PloS One*, 6(5):e20039, 2011.

[174] S. L. Simpson, M. N. Moussa, and P. J. Laurienti. An exponential random graph modeling approach to creating group-based representative whole-brain connectivity networks. *NeuroImage*, 60(2):1117–1126, 2012.

[175] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

[176] D. A. Smith and D. R. White. Structure and dynamics of the global economy: Network analysis of international trade 1965–1980. *Social Forces*, 70(4):857–893, 1992.

[177] T. A. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153, 2006.

[178] R. W. Solava, R. P. Michaels, and T. Milenković. Graphlet-based edge clustering reveals pathogen-interacting proteins. *Bioinformatics*, 28(18):i480–i486, 2012.

[179] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[180] E. S. Spiro, R. M. Acton, and C. T. Butts. Extended structures of mediation: Re-examining brokerage in dynamic networks. *Social Networks*, 35(1):130 – 143, 2013.

[181] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, et al. The BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 39(Suppl. 1):D698–D704, 2011.

[182] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Suppl. 1):D535–D539, 2006.

[183] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005.

[184] M. P. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4221–4224, 2005.

[185] S. Subbiah, D. Laurents, and M. Levitt. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Current Biology*, 3(3):141–148, 1993.

[186] A. Surana, S. Kumara, M. Greaves, and U. N. Raghavan. Supply-chain networks: a complex adaptive systems perspective. *International Journal of Production Research*, 43(20):4235–4265, 2005.

[187] S. J. Swamidass, C.-A. Azencott, K. Daily, and P. Baldi. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics*, 26(10):1348–1356, 2010.

[188] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(Suppl. 1):D561–D568, 2011.

[189] R. Tanaka. Scale-rich metabolic networks. *Physical Review Letters*, 94(16):168101, 2005.

[190] T. Thorne and M. P. Stumpf. Graph spectral analysis of protein interaction network evolution. *Journal of The Royal Society Interface*, 9(75):2653–2666, 2012.

[191] A. H. Y. Tong, M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368, 2001.

[192] A. H. Y. Tong, G. Lesage, G. D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G. F. Berriz, R. L. Brost, M. Chang, et al. Global mapping of the yeast genetic interaction network. *Science*, 303(5659):808–813, 2004.

[193] A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[194] S. Tweedie, M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, R. Seal, et al. FlyBase: enhancing Drosophila gene ontology annotations. *Nucleic Acids Research*, 37(Suppl. 1):D555–D559, 2009.

[195] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, et al. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, 2000.

[196] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *ComPlexUs*, 1:38–44, 2001.

[197] C. Von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.

[198] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[199] S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrika*, 61(3):401–425, 1996.

[200] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[201] R. C. Wilson and P. Zhu. A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, 41(9):2833–2841, 2008.

[202] I. Wohlers, L. Petzold, F. Domingues, and G. Klau. PAUL: Protein structural alignment using integer linear programming and Lagrangian relaxation. *BMC Bioinformatics*, 10(Suppl. 13):P2, 2009.

[203] I. Xenarios, L. Salwinski, J. Duan, P. Higney, S. Kim, and D. Eisenberg. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30:303–305, 2002.

[204] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, and U. Alon. Network motifs in intergrated cellular networks of transcription-regulation and protein-protein interaction. *Proceedings of National Academy of Sciences*, 101(16):5934–5939, 2004.

[205] Y.-K. Yu, E. M. Gertz, R. Agarwala, A. A. Schäffer, and S. F. Altschul. Retrieval accuracy, statistical significance and compositional similarity in protein sequence database searches. *Nucleic Acids Research*, 34(20):5966–5973, 2006.

[206] B. Zhang, R. Liu, D. Massey, and L. Zhang. Collecting the internet AS-level topology. *ACM SIGCOMM Computer Communication Review*, 35(1):53–61, 2005.