

ASYMPTOTICALLY UNBIASED AND CONSISTENT  
ESTIMATION OF MOTIF COUNTS  
IN BIOLOGICAL NETWORKS  
FROM NOISY SUBNETWORK DATA

TRAN NGOC HIEU

(Bachelor of Science, Moscow State University, Russia)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS & APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2013



# Acknowledgements

I would like to express my deepest gratitude to my supervisor Prof. Choi Kwok Pui who has been patiently guiding me during my PhD candidature. His invaluable advice and fruitful ideas have been the most crucial to the completion of this thesis and my future research career. I would not have been able to finish my PhD without his endless support, encouragement and inspiration.

I would also like to thank my co-supervisor Prof. Louis Chen for giving me the opportunity to pursue the PhD degree and supporting me through these years.

I am truly grateful to Prof. Zhang Louxin for his guidance in the project of motif count estimation, which contributes the most important results of this thesis. During the project, I have really learned a lot from Prof. Zhang, especially the analysis skills and the writing skills. I also wish to thank all members in the Network Biology group for their helpful discussion and warm friendship.

I would like to thank the Agency for Science, Technology and Research (A\*STAR) and the National University of Singapore (NUS) for the Singapore International Graduate Award (SINGA), which has provided me with the chance and financial support to fulfill my dream of pursuing the PhD degree. I also wish to express my gratitude to the Department of Statistics and Applied Probability, especially the management staffs for their helpful assistance during my PhD study.

I have been studying abroad for almost ten years, and that would not have been possible without my family's endless support. I am greatly indebted to my parents for their love and always being there to encourage me. Finally, my special thank goes to my love, Jenny, for her faith in me, understanding and love, always being on my side during every difficult time.

Thank you!



# Summary

Increasing availability of genomic and proteomic data has propelled Network Biology to the frontier of biomedical research. Using graph models with nodes and links to study the interactions between cellular components, Network Biology aims to understand topological structures of biological networks, the flow of information inside those networks, and how they control biological processes in living organisms. One of the main research topics in Network Biology focuses on *motifs*, which are usually defined as small connected subgraphs that appear in biological networks much more often than in their random counterparts. Several over-represented motifs such as feed-forward loop, bi-fan, bi-parallel, etc., have been highlighted in the literature as functional units or building blocks of many complex networks in the real world.

A natural question is to gauge whether a motif occurs abundantly or rarely in a biological network. However, counting motifs faces a challenging problem: current high-throughput biotechnology is only able to interrogate a portion of an entire biological network. For instance, recently updated high-throughput yeast two-hybrid assays are only able to detect up to 20% of the protein-protein interactions in living organisms. Moreover, there are a substantial number of spurious interactions that have been wrongly detected. Due to these low coverage and inaccuracy limitations, currently available biological networks actually only represent noisy subnetworks of the real ones. These facts underscore the importance of a reliable method to estimate the number of motif occurrences in biological networks from their noisy observed subnetworks.

In this thesis we develop a powerful method to address the problem of estimating motif counts. Following the *extrapolation* idea, we first apply a scaling-based method to estimate the number of occurrences of a motif in a network from its subnetworks. The proposed estimation, however, is biased if there is noise, that is, spurious and missing links in the subnetworks. Hence, we further refine the method by taking into

account the link error rates, namely, false positive and false negative rates, and develop the bias-corrected estimators. Our theoretical analysis show that the proposed estimators are asymptotically unbiased and consistent for several types of motifs and a wide class of commonly used random network models, including Erdos-Renyi, preferential attachment, duplication, and geometric models. More importantly, the asymptotically unbiased property holds without any assumption on the underlying network and the motif of interest.

Next, we perform extensive simulation validation of the proposed estimators on networks generated from random graph models as well as networks constructed from real datasets. We fully explore how the accuracy of the estimators depends on the underlying network, the subnetworks, and the motif type. Altogether, the theoretical and simulation results confirm that our proposed method is universal and can be easily applied to any complex network, including, but not limited to, biological networks, social networks, the World-Wide-Web, etc.

We then apply the estimators to the protein-protein interaction and gene regulatory networks of four species, namely, Human, Yeast, Worm, and Arabidopsis. Our estimation reveals several important features of these networks while only using their noisy observed subnetwork data. The main findings include the significant enrichment of functional motifs, the linear correlation between motif counts, the association between motif counts and cell functions, etc. The properties of the protein-protein interaction and gene regulatory networks uncovered in our study are consistent with our biological intuition about the complexity of living organisms.

The main findings of this work were first presented at the 17th Annual International Conference on Research in Computational Molecular Biology (RECOMB) 2013, Beijing, China. The revised version with substantial improvements was later accepted for publication in the journal Nature Communication.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Introduction to Network Biology . . . . .	3
1.1.1	What is Network Biology? . . . . .	3
1.1.2	Types of biological networks, data sources and analysis tools . .	7
1.1.3	Topologies of biological networks and their implications . . . . .	12
1.1.4	Random network models . . . . .	16
1.2	Inferring topological properties of biological networks from subnetworks	20
1.2.1	Limitation of biological networks data . . . . .	20
1.2.2	From observed subnetworks to the entire networks: motif count estimation . . . . .	22
1.3	Thesis organization . . . . .	26
<b>2</b>	<b>Theoretical Analysis for Motif Count Estimation</b>	<b>27</b>
2.1	Asymptotically unbiased and consistent estimators . . . . .	28
2.1.1	Estimator for the number of links in an undirected network . . .	30
2.1.2	Estimator for an arbitrary motif $\mathcal{M}$ . . . . .	37
2.2	Noisy subnetwork data and biased-corrected estimators . . . . .	43
2.2.1	Example of calculating the bias-corrected estimator $\tilde{N}_{\mathcal{M}}$ for the feed-forward loop motif . . . . .	48

2.3	Summary . . . . .	52
<b>3</b>	<b>Simulation Validation and Application to Protein-Protein Interaction &amp; Gene Regulatory Networks</b>	<b>54</b>
3.1	Simulation validation . . . . .	56
3.1.1	Simulation from random graph models . . . . .	56
3.1.2	Simulation from real network data . . . . .	63
3.2	Computational time efficiency of the sampling-estimating approach . .	69
3.3	Estimating motif counts in PPI networks . . . . .	73
3.3.1	Comparison of our estimator $\tilde{N}_1$ and CCSB estimator $\tilde{N}^{\text{CCSB}}$ . .	74
3.3.2	Estimating the number of links in PPI networks . . . . .	79
3.3.3	Estimating the number of triangles in PPI networks . . . . .	80
3.3.4	Gene Ontology (GO) analysis of triangles in the observed PPI subnetwork of Yeast . . . . .	81
3.4	Estimating motif counts in gene regulatory networks . . . . .	84
3.4.1	Significant enrichment of motifs . . . . .	85
3.4.2	Linear correlation of motif counts . . . . .	87
3.5	Summary . . . . .	90
<b>4</b>	<b>Discussion</b>	<b>92</b>
4.1	Networks with different types of nodes . . . . .	92
4.1.1	Baits and Preys in PPI networks . . . . .	92
4.1.2	Transcription factors and target genes in gene regulatory networks	94
4.2	Effects of sampling schemes on the estimation . . . . .	95
4.3	Linear correlation of motif counts . . . . .	99
4.4	Conclusion . . . . .	102





# List of Tables

2.1	Detailed expressions of function $f_{\mathcal{M}}()$ for 9 undirected motifs. . . . .	38
2.2	Detailed expressions of function $f_{\mathcal{M}}()$ for 11 directed motifs. . . . .	39
2.3	Detailed expressions of the bias-corrected estimator $\tilde{N}_{\mathcal{M}}$ for 9 undirected motifs. . . . .	49
2.4	Detailed expressions of the bias-corrected estimator $\tilde{N}_{\mathcal{M}}$ for 11 directed motifs. . . . .	50
3.1	Number of nodes and links in the observed PPI subnetworks of <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>H. sapiens</i> , and <i>A. thaliana</i> . . . . .	63
3.2	Observed PPI subnetworks of <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>H. sapiens</i> , & <i>A. thaliana</i> , and their quality parameters. . . . .	73
3.3	The interactome size and the number of triangles in the PPI networks of <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>H. sapiens</i> , and <i>A. thaliana</i> , estimated based on recently published datasets from the Center for Cancer Systems Biology (CCSB). . . . .	80
3.4	The estimated network size and the estimated counts of triad and quadriad motifs (in thousands). . . . .	85

4.1	Re-estimation of the interactome size and the number of triangles in the PPI networks from the intersection of the set of bait proteins and the set of prey proteins. . . . .	94
-----	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

# List of Figures

1.1	Protein-protein interaction network of <i>Saccharomyces cerevisiae</i> . . . . .	8
1.2	Gene regulatory network of <i>Escherichia coli</i> . . . . .	10
1.3	An illustration of the degree distributions of networks generated from four random graph models: Erdos-Renyi (ER), preferential attachment, duplication, and geometric models. As the node degrees in networks generated from the ER model follow a Poisson distribution, we use a histogram to plot the degree distribution for the ER model. The distribution is symmetric, unimodal, and illustrates that nodes tend to have similar degrees. We also use a histogram to plot the degree distribution for the geometric model as there is no any significant skewness. On the other hand, the node degrees in networks generated from the preferential attachment model are scale-free, that is, there are a lot of nodes with low degrees and a small, but significant number of nodes with high degrees. In particular, the node degrees follow a power-law distribution, that is, $P(k) \sim k^{-\lambda}$ , which is best illustrated by the linear pattern between $\log P(k)$ and $\log k$ when the degree distribution is plotted in the log-log scale. The degree distribution for the duplication model is also scale-free and is plotted in the log-log scale. . . . .	17
1.4	Schematic view of the motif count estimation problem. . . . .	23

2.1	All possible (9) undirected motifs that have up to 4 nodes. . . . .	44
2.2	11 selective directed motifs, some of which such as feed-forward loop, bi-fan, bi-parallel have been highlighted in literature as building blocks or functional units in many real-world complex networks [59]. . . . .	45
3.1	MSE of the estimators $\widehat{N}_9$ and $\widetilde{N}_9$ for the number of occurrences of FFL motif in networks generated from the ER model. . . . .	59
3.2	MSE of the estimators $\widehat{N}_9$ and $\widetilde{N}_9$ for the number of occurrences of FFL motif in networks generated from the preferential attachment (upper) and the duplication (lower) models. . . . .	60
3.3	The convergence rate of $\text{Var}\left(\frac{\widehat{N}_1}{N_1}\right)$ in equation (2.11) and the dominated term $\frac{N_2}{N_1^2}$ (denoted by $\pi_1$ in the legend) for the preferential attachment model. . . . .	64
3.4	The convergence rate of $\frac{N_2}{N_1^2}$ (denoted by $\pi_1$ in the legend) is bounded by $\frac{\log(n)}{n}$ as shown in Proposition 1 for the preferential attachment model. . . . .	64
3.5	Observed PPI subnetwork of <i>H. sapiens</i> from Y2H experiment. . . . .	65
3.6	Degree distribution of four observed PPI subnetwork of <i>S. cerevisiae</i> , <i>C. elegans</i> , <i>H. sapiens</i> , and <i>A. thaliana</i> (log-log scale). The linear pattern between the log of the number of nodes and the log of the node degree implies the scale-free structure of these subnetworks. . . . .	65
3.7	Performance of the estimator $\widehat{N}_M$ with respect to the node sampling probability $p$ in the PPI network of <i>S. cerevisiae</i> for different undirected motifs. . . . .	68
3.8	Performance of the estimator $\widetilde{N}_M$ with respect to the node sampling probability in the PPI network of <i>S. cerevisiae</i> . . . . .	68
3.9	Performance of the estimator $\widetilde{N}_M$ of the number of links with respect to false positive and false negative rates in the PPI network of <i>S. cerevisiae</i> . . . . .	70

3.10	Performance of the estimator $\tilde{N}_M$ of the number of triangles with respect to false positive and false negative rates in the PPI network of <i>S. cerevisiae</i> .	70
3.11	Computational time efficiency and MSE of the estimator $\hat{N}_3$ for estimating the number of triangles in an example network of $n = 5,000$ nodes and link density $\rho = 0.1$ .	71
3.12	Limitation of gold-standard datasets.	75
3.13	The enrichment in shared GO annotations of triangles in the observed PPI subnetwork of <i>S. cerevisiae</i> .	82
3.14	Motif count of feedback foop in forty-one observed networks (red “x”) and in their randomly rewired replicates ( $\mu \pm 3\sigma$ from 50 replicates for each observed network).	88
3.15	Motif count of feed-forward loop in forty-one observed networks (red “x”) and in their randomly rewired replicates ( $\mu \pm 3\sigma$ from 50 replicates for each observed network).	88
3.16	Correlation of motif counts in forty-one Human cell-specific transcription factor (TF) regulatory networks.	89
4.1	Plots of the MSE of the estimator $\hat{N}_3$ for triangle count with respect to four different sampling schemes and average sampling proportion $p$ .	99
4.2	Plots of the mean of the ratio $\frac{\hat{N}_3}{N_3}$ for triangle count with respect to four network models and increasing average sampling proportion $p$ .	100
4.3	Plots of the mean of the ratio $\frac{\hat{N}_3}{N_3}$ for triangle count with respect to the power-law exponent $\gamma$ and average sampling proportion $p$ .	100
4.4	Linear correlation of the motif counts in random networks which are generated from the forty-one Human cell-specific TF regulatory networks using the link rewiring process.	101

4.5	Linear correlation of the residuals of the motif counts' regression with respect to the number of links in forty-one Human cell-specific TF regulatory networks. . . . .	101
-----	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----

# Chapter 1

## Introduction

### 1.1 Introduction to Network Biology

#### 1.1.1 What is Network Biology?

Following the discovery of the double helix structure of the DNA molecule in 1953 by James Watson and Francis Crick [1], the completion of the Human Genome Project (HGP) in 2003 has been the greatest achievement ever in the history of biology and medicine [2]. This has enormous impacts on scientific research activities as well as biomedicine related industries [3]. The HGP was then followed by an explosion of new research areas which open up promising opportunities and challenges for the scientific community in the post-genomic era. The ultimate goal is to enhance our knowledge of Human Health and Diseases, and subsequently to provide humankind with better living conditions, health-care services, and other benefits.

As one of the most active fields in biomedical research, Molecular Biology has attracted a great deal of attention from scientists across different disciplines such as biologists, chemists, mathematicians, computer scientists, etc. Intensive efforts have been put into Molecular Biology to study cellular molecules (i.e., genes, proteins, en-



zymes, metabolites, etc.), substantially improving human knowledge of the structures and biological functions of the smallest elements of life.

However, information of individual cellular molecules alone is not enough to infer a cell's functions, and similarly, information of individual cells alone cannot tell us the whole picture of biological processes in a living organism. While keeping focus on each individual, one may ignore the interrelation between them. Cellular molecules must be studied in the context of integrated systems of interacting components, and as parts of the systems, they do not function in isolation, but in cooperation. That is the underlying principle of Network Biology: biological functions, as well as dysfunctions (i.e., genetic disorders or diseases), in a cell or in a living organism are co-regulated by multiple types of complex networks of interacting cellular components.

Network Biology is a rapidly emerging field in post-genomic biomedical research. It was first introduced at the beginning of the 21st century [4, 5], and recently has become one of the most attractive fields because it has demonstrated potential impacts in biology and medicine, especially on studies related to Human Health and Diseases [6]. Network Biology even serves as the fundamental background for the development of two latest research topics, namely Network Medicine [7] and Network Disease [8]. Roughly speaking, Network Biology is a multi-discipline research field in which different theories, frameworks, models and techniques from diverse fields of science, including, but not limited to biology, chemistry, physics, mathematics, statistics, computer science, are integrated to study different types of biological networks, to explore their topological structures and properties, and most importantly, to understand how these networks control cellular functions and biological processes in living organisms.

Two basic elements in biological networks are nodes and links. Nodes are cellular components (i.e., genes, proteins, enzymes, metabolites, etc.) and links represent the interactions between the components (Fig. 1.1 and 1.2). Links can be undirected (e.g.,

in protein-protein interaction networks) or directed (e.g., in gene regulatory networks, metabolic networks, signaling pathways). A biological network thus represents a complex system of interacting cellular molecules, and the flow of biological information inside such systems regulates all activities of the cell.

The most surprising result of complete genome sequencing projects, perhaps, is that the number of genes in the whole genome is not significantly different among species. For example, the human genome contains approximately 22,000 protein coding genes [2], which is much lower than expected, especially when compared to simple model organisms such as yeast (6,500 genes) [9], worm (20,000) [10], fruit fly (17,000) [11]. The estimated number of genes of human is even smaller than that of Arabidopsis which is estimated as 27,000 [12]. Moreover, there are just over two hundred genes that are unique to human. Thus it is obvious that the number of genes alone cannot explain the nature of biological complexity of living organisms as previously expected. However, biological networks, which possess much more complicated architectural features rather than the simple number of nodes, may provide us with better explanations to the question of species diversity and evolution.

Another interesting phenomenon is the robustness of some model organisms against gene mutations. For example, Wagner (2000) has reported the great tolerance of yeast against gene removal in [13]. This resilience suggests that under genetic mutations, some genes can be somehow functionally replaced by the others, and thus indicating that there must be some functional connections between the genes. Indeed, one of the most crucial findings of Biological Networks Alignment [14], a key research topic in Network Biology, has reported that some specific groups of proteins, and more importantly, the physical interactions between them are conserved and stick together through thousands years of evolution across multiple species. Such unusual conservations suggest that those protein pathways and complexes must play some critical roles in the survival,

reproduction and evolution of an organism. Moreover, their functions are determined not only by individual proteins, but also by the physical interactions between them. Those are just a few examples from thousands of important findings that support the underlying principle of Network Biology and underscore the importance of this new perspective in biomedical research.

As a new research area, Network Biology opens up promising opportunities as well as challenging problems for the scientific community. Fortunately, Network Biology inherits a solid theoretical background from graph theory, the fundamental field of mathematics which uses graphs to model pair-wise relations between objects [15]. Moreover, networks, or graphs representations are the most ubiquitous models that have been used to study various complex systems in other fields of science such as physics, computer science, social science [16, 17]. Some prominent examples include the World-Wide-Web, human social networks, scientific citations networks, electrical and power systems, neuron networks, etc. Most importantly, initial studies have pointed out that several complex networks in the real world, including biological networks, unexpectedly share some fundamental architectural features such as scale-free degree distribution [18], small-world properties [19], hierarchical and clustering structures [20]. Thus, this surprising universality allows us to apply well-developed and ready-to-use techniques, tools, and soft-ware applications from other well-established domains to Network Biology. The strong support from theoretical and technical sites as well as the rapidly increasing availability of genomic and proteomic data accumulated from high-throughput experiments have propelled Network Biology to the frontier of biomedical research. Network Biology is expected to revolutionize our understanding and knowledge of biology, medicine, Human Health and Diseases in this post-genomic era.

### 1.1.2 Types of biological networks, data sources and analysis tools

There are three major types of biological networks that have been the target of most studies in Network Biology: protein-protein interaction (PPI) networks, gene regulatory networks (GRNs) and metabolic networks.

In protein-protein interaction networks (Fig. 1.1), each node represents a particular protein and each link represents an interaction between two proteins. Links are undirected as an interaction means that the two proteins bind to each other. There are currently two high-throughput experimental techniques that are widely used to produce large-scale PPI networks [21]. Yeast two-hybrid (Y2H) assays, which were first introduced by Fields and Song in [22], can detect direct physical, or binary, interactions between any two proteins. This technology was used by Uetz *et al.* and Ito *et al.* in [23, 24] to produce the first PPI maps of *Saccharomyces cerevisiae*, or yeast, a well-studied model organism that has the most comprehensive and reliable data currently available on PPIs. Later, Y2H assays were also applied to other model organisms such as *Caenorhabditis elegans* (i.e., worm) and *Drosophila melanogaster* (i.e., fruit fly).

In 2005, two independent groups Rual *et al.* and Stelzl *et al.* successfully mapped the first versions of the human PPI network [25, 26]. In particular, Rual *et al.* were able to detect  $\sim 2,800$  new interactions connecting  $\sim 7,000$  protein-encoding genes, especially  $\sim 300$  interactions among them are linked to over 100 disease-associated proteins. Recently, Y2H assays have been improved by the experts from the Center for Cancer Systems Biology, Dana-Farber Cancer Institute, and are associated with an empirical framework that allows us to estimate the overall accuracy and sensitivity of high-throughput PPI mapping [27, 28, 29, 30].

Unlike Y2H assays which are able to detect direct binary interactions, affinity pu-

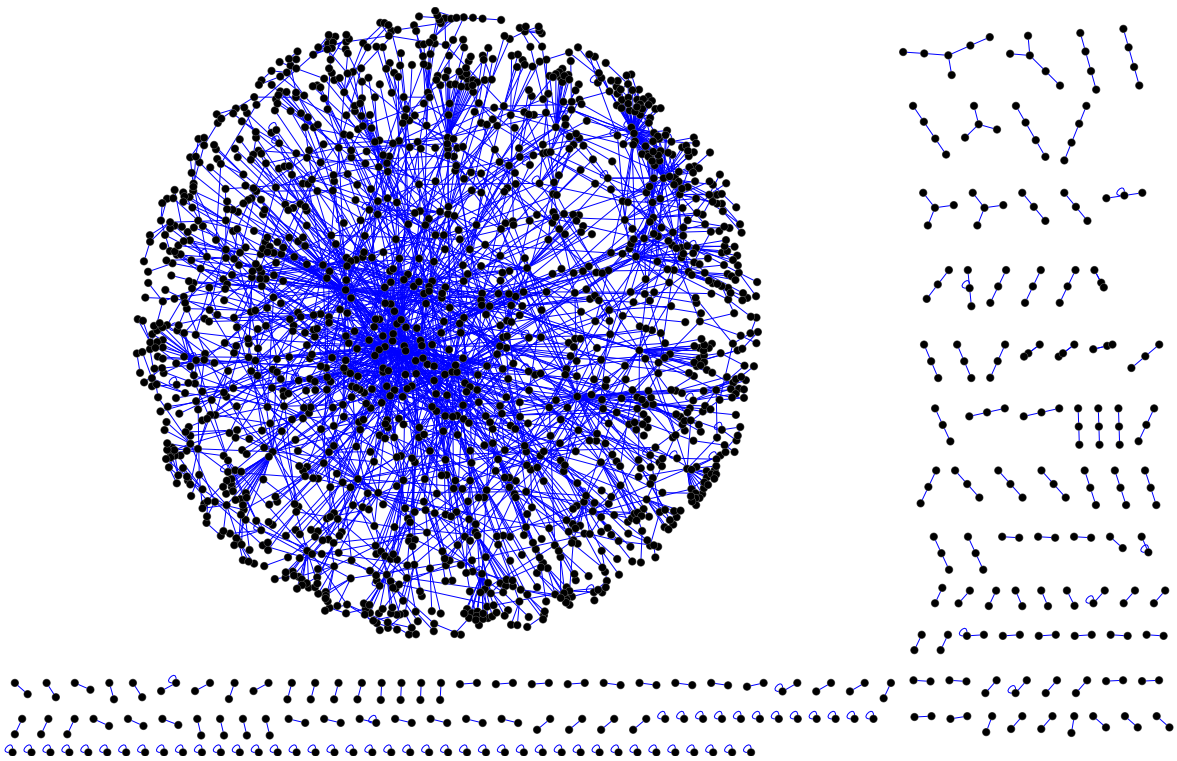


Figure 1.1: Protein-protein interaction network of *Saccharomyces cerevisiae*. There are 2018 nodes (proteins) and 2930 links (interactions). Data from Center for Cancer Systems Biology, Dana-Farber Cancer Institute [28]. Network visualization: Cytoscape [51].

rification followed by mass spectrometry (AP-MS) assays, which were first introduced by Rigaut *et al.* in [31], can detect protein complexes and indirect associations between proteins [32, 33]. Thus, a link detected from Y2H assays represents a direct physical interaction between two proteins, whereas a link detected from AP-MS assays implies that the two proteins belong to the same complex and there may be direct or indirect interactions between them. For the same organism, PPI networks generated by these two approaches may exhibit different structures and properties [21, 28]. In this thesis, we mainly focus on PPI networks that are generated from high-throughput Y2H assays.

Another major type of biological networks is gene regulatory networks (GRNs) (Fig. 1.2). There are two different kinds of nodes in a gene regulatory network: transcription factors and target genes. A transcription factor is a DNA-binding protein that can bind to specific DNA regions, which are called binding motifs, of a target gene or another transcription factor and subsequently regulates the expression of that gene. A target gene is regulated by transcription factors and cannot regulate any other gene. Thus, links in GRNs represent regulatory (protein-DNA binding) interactions and they are directed.

There are currently two experimental systems that can be used to reconstruct gene regulatory networks in a high-throughput fashion. In yeast one-hybrid (Y1H) assays [34], a specific regulatory DNA sequence of interest, called promoters, is used as bait to identify all putative transcription factors (preys) that bind to that sequence. On the other hand, Chromatin Immunoprecipitation (ChIP) experiments [35] are usually used to determine all potentially associated DNA binding sites for a DNA-binding protein of interest. Obviously, the two approaches are complementary and their combination is required for the reconstruction of gene regulatory networks. As for PPI networks, the most comprehensive and accurate GRN is that of *Saccharomyces cerevisiae* [36]. Recently, different research groups have attempted to map the entire GRN of human

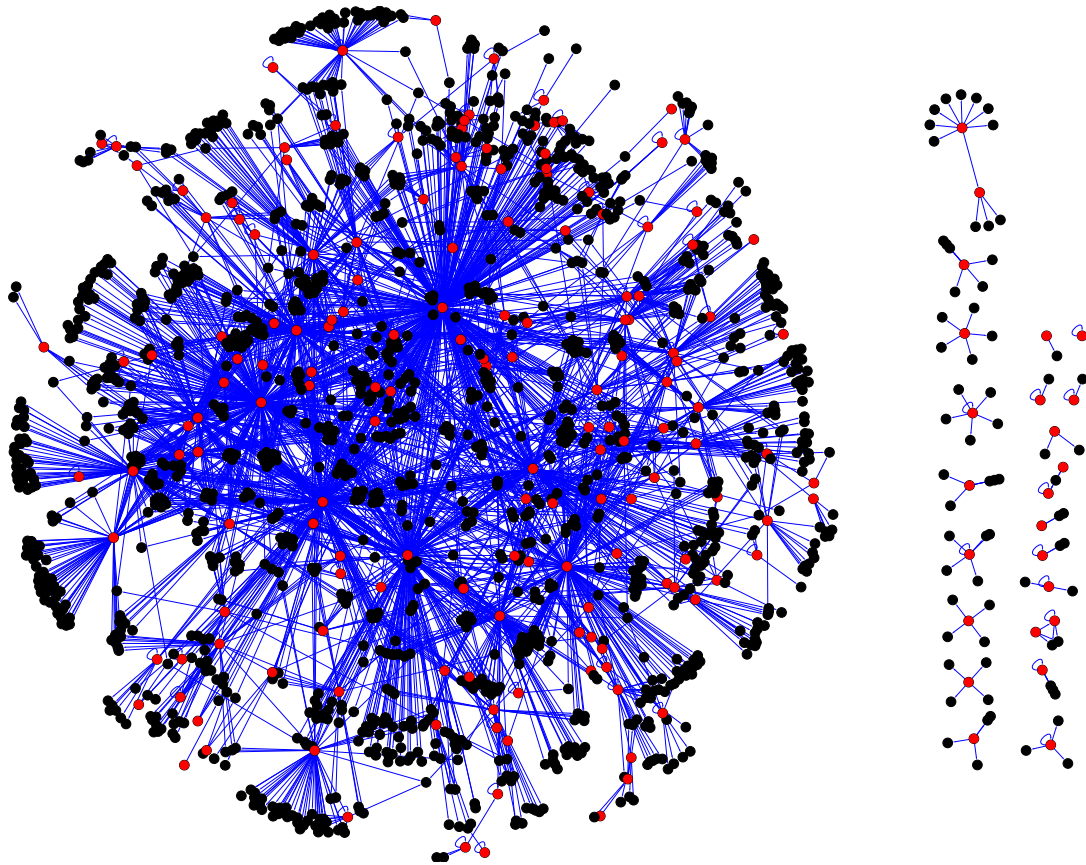


Figure 1.2: Gene regulatory network of *Escherichia coli*. There are 186 transcription factors (red nodes), 1,510 target genes (black nodes) and 3809 directed links. Data from RegulonDB (version 7.0) [45]. Network visualization: Cytoscape [51].

[37], and moreover, this can be done across multiple cell and tissue types [38].

The last major type of biological networks is metabolic networks, which actually appeared even before protein-protein interaction networks and gene regulatory networks [39]. In metabolic networks, nodes are biochemical metabolites and links represent the reactions, or the enzymes catalyzing the reactions that convert one metabolite to another. Links may be directed or undirected, depending on whether the reactions are reversible or not. In some context, nodes may represent enzymes and an link between two enzymes indicates that the product of one enzyme is the substrate of the other. Metabolic networks have been constructed mostly by the meticulous literature-curation of large numbers of publications for decades, and thus, are the most comprehensive among all biological networks [40]. With recent advanced computational technologies, metabolic network reconstruction also involves predicting orthologous reactions across multiple species.

Emerging at the beginning of the 21st century, biological networks data has increased rapidly, especially in the last few years thanks to novel advances in high-throughput experimental technologies. Nowadays a huge amount of data are widely available, not only from original publications, but also from several open-access databases as a result of enormous efforts of literature-curation experts. Protein-protein interaction networks data of multiple species including human are available in DIP [41], BioGRID [42], STRING [43], etc. Gene regulatory networks data can be downloaded from TRANSFAC [44], RegulonDB [45], AtRegNet [46], etc. KEGG [40], perhaps, is the most comprehensive database for metabolic networks and pathways. Some other useful resources include MIPS [47], BIND [48], BioCyc [49], Reactome [50], etc. The databases listed here just represent a few prominent examples among several hundreds of resources that have been developed and maintained by diverse groups of scientists from over the world. A brief summary of more than 300 resources related to biological networks and



pathways can be found in the meta-database Pathguide ([www.pathguide.org](http://www.pathguide.org)).

In order to deal with that huge amount of data on biological networks, where each network is a complex system of thousands of nodes and hundreds of thousands of links, several tools and applications have been developed to facilitate research activities in Network Biology. Among them, Cytoscape is the most outstanding bioinformatics software for network visualization, analysis and biomedical discovery [51]. This software incorporates different formats of biological networks data and is linked to several popular databases and resources. Cytoscape also allows the integration of other types of information such as gene expression profiles, Gene Ontology [52], functional annotations, etc., as node or link attribute data. The most beautiful feature of Cytoscape is that this is a freely available and open source Java platform that allows the research community to develop their own plug-ins for more specific and advanced analysis tasks. Cytoscape has been effectively supporting the research community for almost 10 years and will continue to play its crucial role in Network Biology with the next major version released soon in the near future. More importantly, following this flagship tool, an open source suite of software technologies dedicated to biological networks visualization, analysis and discovery is under development by the National Resource for Network Biology (NRNB, [www.nrn.org](http://www.nrn.org)) with support from the National Institutes of Health (NIH). Such bioinformatics packages provide the research community with powerful tools to gain more insights into those complicated systems of interacting cellular components.

### **1.1.3 Topologies of biological networks and their implications**

As biological networks are presented as graphs of nodes and links, a fundamental question to ask is “what are their topologies?”, and the immediate next question will be “how do those topological properties facilitate the flow of information inside the net-

works?”. This is basically the underlying framework of any analysis in Network Biology: the topological structure of a network of interest and the biological information of its nodes and links (e.g., gene expression profiles and functional annotations of nodes, types and scores of links, etc.) are combined to explore the functions of the entire network. Moreover, as mentioned earlier, complex networks from diverse fields, including biological networks, have been reported to share remarkable similarities in their structure. This surprising observation further emphasizes the importance of understanding the topologies of biological networks and their implications.

The most striking feature, perhaps, is the scale-free property that has been observed in most biological networks of multiple species. In particular, the degree distribution in PPI networks and the out-degree distribution in GRNs are believed to have the scale-free property. For any node  $u$  in an undirected network, its degree is defined as the number of links adjacent to it, or in other words, the number of its neighbors. In a directed network, the out-degree of a node  $u$  is the number of links pointing-out of that node. Scale-free property implies a coexistence of a large number of low-degree nodes and a small, but significant, number of high-degree nodes, which are often referred to as “hubs”. This scale-free structure has also been observed in many real-world networks such as social networks, the World-Wide-Web, and other technological networks. More importantly, it is suggested that the degree distribution in those networks follows a power law: the probability that a randomly chosen node has degree  $k$ , that is, it has  $k$  incident links, follows  $P(k) \sim k^{-\lambda}$ , where the exponent  $\lambda$  is network-specific and ranges between 2 and 3 [18].

The scale-free topology attracts a great deal of attention from the research community because such networks exhibit surprising tolerance against random perturbations. Random failures mainly affect nodes of low degree, and usually such nodes do not play important roles in a network. That also explains the robustness against gene muta-

tions that has been observed in some model organisms [13]. However, deletion of hubs, even just a few, may lead to the corruption of the entire network. This robustness and vulnerability is a signature feature of scale-free networks, including biological networks [53]. From a biological point of view, this double-edge feature suggests that hubs may represent essential proteins for the survival and reproduction of a cell [54]. The relationship between topological centrality and biological essentiality of proteins in PPI networks has been the target of several studies in Network Biology [54, 55, 56].

Another notable feature of biological networks is the small-world effect which is characterized by the two properties: small shortest path length and large clustering coefficient [19]. The shortest path length, or characteristic path length, between any two nodes  $u, v$  in a network is the length of the shortest path connecting  $u$  and  $v$ . Although there may be some alternative paths between  $u$  and  $v$ , it is believed that information always flows via the shortest path. Thus, the average over the shortest paths between all possible pairs of nodes of a network, which is usually called the mean path length, can be used to measure the efficiency of information flow in the network. The smaller the mean path length is, the more well-connected the network is.

Another measure of the interconnectivity in a network is the clustering coefficient. Intuitively, if node  $u$  is connected to  $v$ , and  $v$  is connected to  $w$ , then it is more likely  $u$  is also connected to  $w$ . The clustering coefficient of a node  $u$  is defined as  $C_u = \Delta_u / \binom{k_u}{2}$ , where  $k_u$  is the degree of  $u$  and  $\Delta_u$  is the number of links connecting its neighbors. In other words,  $C_u$  describes how likely any two neighbors of  $u$  will interact with each other. The average clustering coefficient of a network measures the overall tendency of its nodes to form highly interconnected local clusters which represent potential candidates for predicting functional modules. It has been observed that biological networks exhibit significantly shorter path length and higher clustering coefficient than those of a random network of equivalent size and degree distribution

[57], indicating that biological networks are small-world.

The combination of scale-free and small-world topologies, in particular the coexistence of hubs and highly interconnected clusters suggests that biological networks may exhibit a hierarchical architecture [20]. The most important signature of a hierarchical architecture is the dependence of the clustering coefficient on the degree of a node  $u$ , which follows  $C_u \sim k_u^{-1}$ . Low-degree nodes tend to form small, densely interconnected local clusters and hence have a high clustering coefficient. On the other hand, highly connected hubs tend to have a low clustering coefficient because they do not participate in any local clusters, but play their role as bridges to connect different clusters. Thus, small clusters are connected via hubs to form larger ones, which in turn are connected again via hubs to form even much larger clusters. Eventually, a hierarchical architecture emerges and incorporates both scale-free topology and local clustering structure [20, 6]. Furthermore, it has been found that transcription factors in gene regulatory networks are organized in a pyramid-shaped hierarchical structure in which a few master transcription factors on the top level regulate those at the middle levels, and altogether regulate those at the bottom levels, where most transcription factors are located [58]. The hierarchical architecture is believed to best describe the global structure of most biological networks.

Besides the global architecture, the local structure also plays a crucial role in biological networks. Network motifs, that is, small subgraphs that are significantly over-represented in biological networks than in randomized networks, are believed to represent functional units of biological processes. Some prominent network motifs such as single-input modules, feed-forward loops, bi-fans, bi-parallels have been detected in many real-world networks, including biological networks [59, 60]. Detecting motifs in a given network and exploring their properties are essential for the understanding of the network's functions [62, 61].

### 1.1.4 Random network models

The goal of understanding the topological structures and properties of biological networks cannot be achieved without appropriate random network models that play as null hypotheses based on which unusual features of biological networks can be detected. For instance, as mentioned above, in order to detect network motifs in a given biological network, one needs to verify if a subgraph is significantly over-represented in the observed network in comparison to randomized networks that have the same size (numbers of nodes and links) and the same degree distribution [59, 60]. On the other hand, a suitable random network model that well captures the topological structures and properties of a real biological network can be used to facilitate theoretical as well as simulation analyses to further explore more features of that biological network, make predictions and estimations, etc. These analyses cannot be done if we only look at the observed network.

A classical random graph model in graph theory is the Erdos-Renyi (ER) model [63]. This model has two parameters: the number of nodes  $n$  and the link density  $\rho$ . A random network  $\mathcal{G}$  is generated from the ER model as follows: first,  $n$  singletons are created, and then a link is placed independently and uniformly at random with probability  $\rho$  between any two nodes. The node degrees in a random network  $\mathcal{G}$  generated from the ER model follow a Poisson distribution in which all nodes tend to have similar degrees, approximately equal to the average degree of the network. This can be clearly seen from the symmetric and unimodal histogram in Fig. 1.3. Moreover, this network has a symmetric structure, that is, subnetworks which are randomly sampled from  $\mathcal{G}$  tend to have similar topological properties.

However, the ER model is too simple to describe topological structures and properties of real-world networks, for example, the well-known scale-free property. In [18] Barabasi and Albert proposed the first model that can capture this scale-free structure.

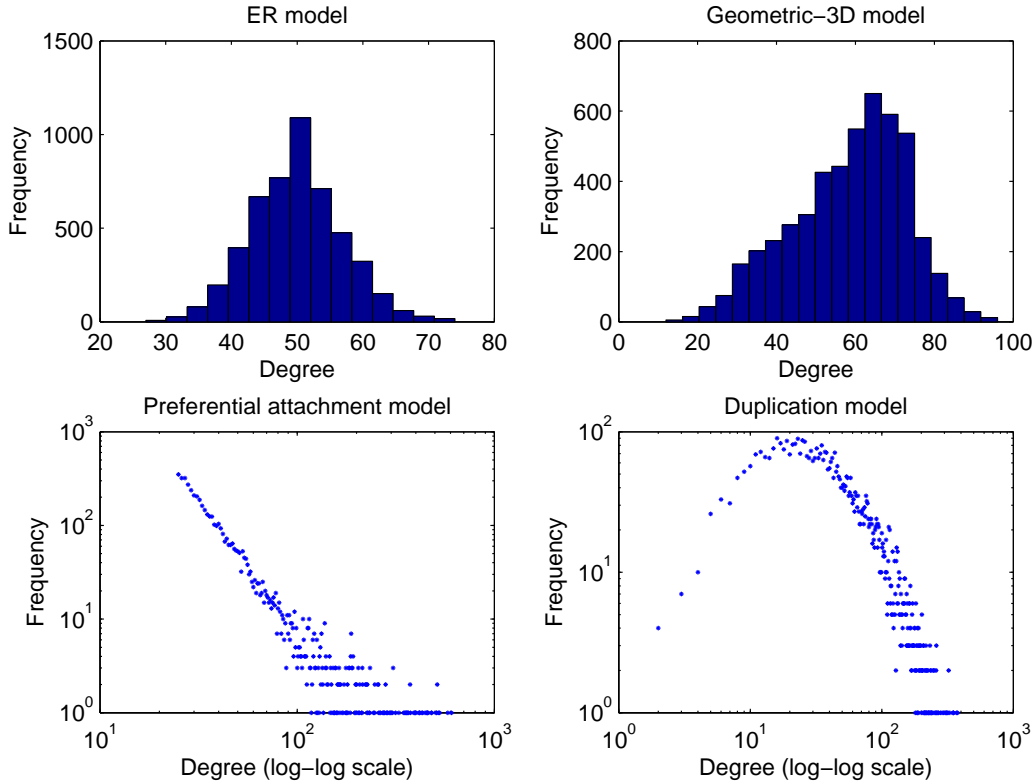


Figure 1.3: An illustration of the degree distributions of networks generated from four random graph models: Erdos-Renyi (ER), preferential attachment, duplication, and geometric models. As the node degrees in networks generated from the ER model follow a Poisson distribution, we use a histogram to plot the degree distribution for the ER model. The distribution is symmetric, unimodal, and illustrates that nodes tend to have similar degrees. We also use a histogram to plot the degree distribution for the geometric model as there is no any significant skewness. On the other hand, the node degrees in networks generated from the preferential attachment model are scale-free, that is, there are a lot of nodes with low degrees and a small, but significant number of nodes with high degrees. In particular, the node degrees follow a power-law distribution, that is,  $P(k) \sim k^{-\lambda}$ , which is best illustrated by the linear pattern between  $\log P(k)$  and  $\log k$  when the degree distribution is plotted in the log-log scale. The degree distribution for the duplication model is also scale-free and is plotted in the log-log scale.

Their model is based on two fundamental features of real-world networks which are not considered in the ER model: the growth process and the preferential attachment mechanism. Firstly, real-world networks grow and nodes are continuously added to existing networks. Secondly, the authors have found a common phenomenon in real-world networks which they referred to as the preferential attachment mechanism: when added to an existing network, a new node is more likely to connect itself to a highly connected node rather than a node of low degree. Indeed, a newly created web-site will prefer to link itself to already well-known ones such that it can attract more users from those web-sites. Similarly, a new research article is more likely to cite well-known ones, since such highly-cited papers usually include important results that can be applied in new research manuscripts. In this way, a highly connected node will have more chances to get new links from newly created nodes, and hence its degree is more and more increasing. This is the rich-get-richer phenomenon in real-world networks, a consequence of the preferential attachment mechanism.

In particular, a random network  $\mathcal{G}$  is generated from the preferential attachment model as follows:

- A small initial random network  $\mathcal{G}_0$  is generated from the ER model.
- At each iteration, a new node with  $l$  incident links is added to the current network. Neighbors of the newly added node are chosen with probabilities proportional to their current degrees.

Barabasi and Albert have shown that networks generated from the preferential attachment model are scale-free, that is, there are a lot of nodes with low degrees and a small, but significant number of nodes with high degrees. Moreover, they have shown that the node degrees follow a power-law distribution, that is,  $P(k) \sim k^{-\lambda}$ . This power-law distribution is best illustrated by the linear pattern between  $\log P(k)$  and  $\log k$  when

the degree distribution is plotted in the log-log scale (Fig. 1.3).

The preferential attachment model, however, cannot be applied directly in the context of biological networks. The evolution of biological networks requires more detailed explanations: how a new gene is created and how it is connected to existing genes in the current network. In [64] Chung *et al.* (2003) proposed duplication models to describe the gene duplication event, which is believed to represent one of the two driving forces of genome evolution [65]. In the full duplication model, a new gene is created by full duplication from an existing gene. As a result, the newly created gene inherits all functions of its original, including interactions with other genes. In the network context, an existing node  $u$  is chosen from the current network and is duplicated to create a new node  $u'$  which is subsequently connected to all neighbors of  $u$ . Interestingly, if the duplicated node  $u$  is chosen uniformly at random, that is, all existing nodes are equally likely to be duplicated, a highly connected hub is more likely to have one of its neighbors to be duplicated, and hence has a higher chance to get a new link. This is indeed the “rich-get-richer” phenomenon. The newly created node  $u'$  is more likely to be duplicated from a neighbor of a highly connected hub, and hence is more likely to be connected to that hub. This is the preferential attachment phenomenon.

The second driving force of genome evolution is the gene mutation event and it is captured by the partial duplication model [64]. In particular, after the new node  $u'$  is created by full duplication from the duplicated node  $u$ ,  $u'$  is allowed to “mutate”, that is, to lose some of its current links and to gain some new links, according to some controlling parameters of the model. Chung *et al.* (2003) have also demonstrated that networks generated from the partial duplication model are scale-free and their degree distributions follow a power law (Fig. 1.3). Perhaps this is currently the best model that is strongly supported by biological theories and can capture important features of biological networks such as scale-free, power-law degree distribution, preferential



attachment and “rich-get-richer” phenomena.

The geometric model was also proposed in [66] to study biological networks. A random network  $\mathcal{G}$  is generated from the geometric model as follows: first,  $n$  nodes are placed uniformly at random in a unit cube, and then any two nodes are connected if the distance between them is less than a given threshold  $\delta$ . Using graphlet frequency and graphlet degree distribution as distance measures of similarity between two networks, Przuli *et al.* have shown that the geometric model yielded better fit to biological networks than the other three random network models [66] (the term graphlet was used in that paper to denote a small connected subgraph with 3-5 nodes). As shown in Fig. 1.3, the degree distribution for the geometric model is left-skewed with more nodes of high degrees and less nodes of low degrees. However, the skewness is not as extreme as in the scale-free degree distribution.

## 1.2 Inferring topological properties of biological networks from subnetworks

### 1.2.1 Limitation of biological networks data

The most challenging problem in Network Biology is the low coverage and the inaccuracy of biological networks data due to the limitation of current experimental techniques. Moreover, even measuring the quality and error rates of experimental high-throughput datasets is also a difficult task.

Traditional assessment approaches which use gold-standard reference sets to benchmark interactions detected from high-throughput experiments have some considerable limitations [67, 68, 69, 70]. In particular, gold-standard reference sets, which are usually collected from literature curation, are themselves incomplete and biased. An interaction

which is detected from high-throughput experiments but was not reported previously in any gold-standard reference set may be considered as a false positive, but may also represent a novel interaction. Computational methods developed to assess biological relevance of detected interactions, e.g. expression profile reliability (EPR) index in [71], cannot tell the whole picture of the quality of a high-throughput dataset. For instance, two interacting proteins are not necessary to have their expression highly correlated.

Fortunately, an empirical framework was proposed recently to rigorously evaluate quality parameters in association with “second-generation” high-quality Y2H assays [27, 28]. The framework uses multiple cross-assay validation to estimate four quality parameters, that is screening completeness, precision, assay sensitivity, and sampling sensitivity, which altogether describe the overall performance of a high-throughput experiment. For instance, the precision for the human PPI dataset CCSB-HI1 was estimated at  $\sim 79.4\%$  in [27], which corresponds to a false discovery rate  $\sim 20.6\%$ , whereas this false discovery rate had been previously overestimated up to 87%-93% using traditional comparison approaches [70, 25]. The precision for a new high-quality PPI dataset of *Saccharomyces cerevisiae*, CCSB-YI1, was also estimated at  $\sim 94\%$  in [28]. Although new Y2H assays achieve very high precision, the sensitivity is quite low, where the best sensitivity is at  $\sim 17\%$  for *Saccharomyces cerevisiae*.

In general, even for the most well-studied model organism like *Saccharomyces cerevisiae*, what we actually observe merely reflects a minor part of the whole picture, i.e. a noisy subnetwork of a real complete network, which is much more complicated and still remains unknown to the research community. While intensive efforts are ongoing in laboratories to improve large-scale high-throughput experimental technologies, it is highly desirable to infer some initial ideas on the global and local features of a complete biological network, given its observed subnetwork. Such predictions are of critical importance to shed light on the organizational architecture and topological properties

of real biological networks, as well as to guide wet-lab experiments to focus on the right target.

### **1.2.2 From observed subnetworks to the entire networks: motif count estimation**

In this thesis we study the problem of inferring topological features of biological networks from their noisy observed subnetworks, which may contain spurious and missing links (Fig. 1.4).

The simplest case of this problem is to estimate the size of an interactome, that is, the number of interactions in a PPI network, has been the target of several studies. This task is especially important to evaluate the progress of current PPI mapping projects and to estimate how much work still needs to be done. Moreover, it is expected that the size of interactomes may partially explain the question of biological diversity of living organisms, which the number of genes has failed to answer. For example, one may expect that Human interactome should have more interactions than other simple organisms do.

There are two approaches to address the problem of estimating the size of interactomes. In [72] the author proposed the first approach to estimate the size of an interactome by modeling the overlap between two independent datasets of that interactome using hypergeometric distribution. Hart *et al.* further extended this method by taking into account the false positive rate, which was evaluated by comparing the two datasets of interest with another reference dataset [70]. However, this method requires that the two datasets must be generated from identical, or at least similar experimental conditions, and they must be independent from the reference set. Unfortunately, this is rarely the case for biological networks data. Most importantly, this approach is difficult to generalize to the case of larger motifs.

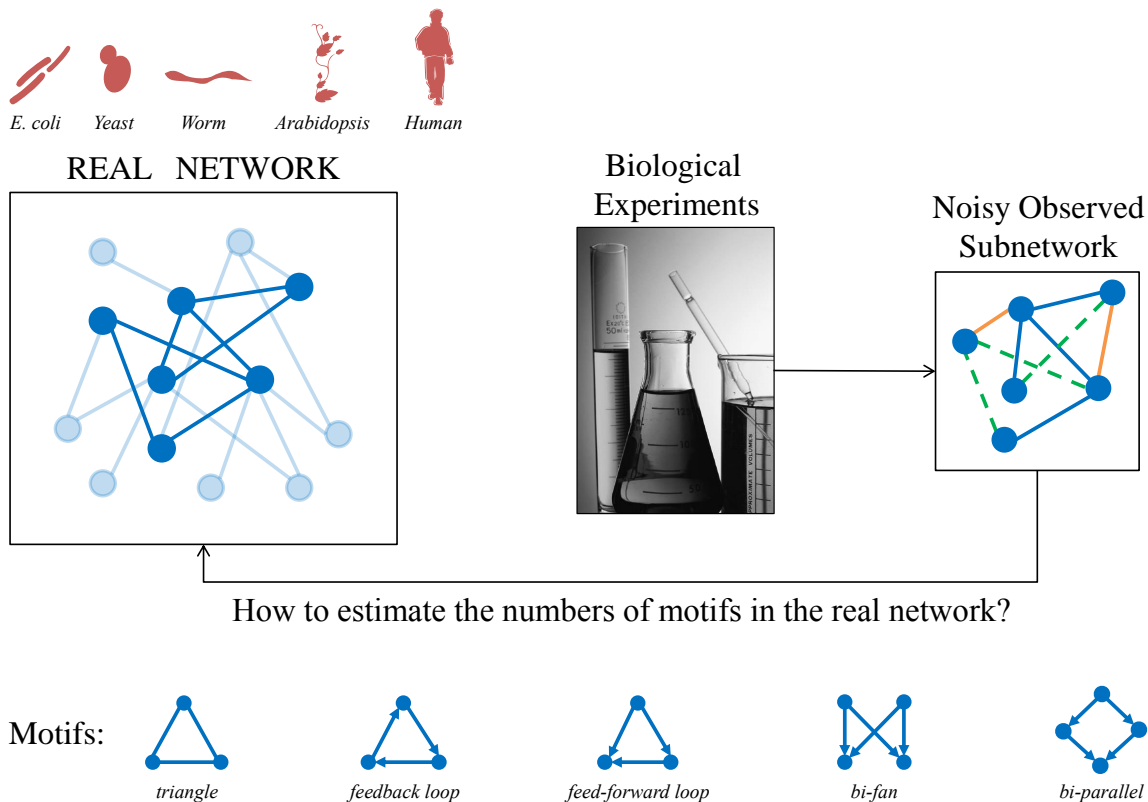


Figure 1.4: Schematic view of the motif count estimation problem. Biological networks of most species are not completely known due to limitations of current biotechnologies. Their subnetworks are usually inferred with errors, that is, spurious (orange) and missing (dashed and green) links, from high-throughput experiments such as Yeast Two-Hybrid, Affinity Purification followed by Mass Spectrometry, etc. Spurious links are the links that do not exist in the real network but are wrongly detected by the experiments. Missing links are the links that exist in the real network but cannot be detected by the experiments. In this study, we propose a method to estimate the number of motif occurrences in biological networks from their noisy observed subnetworks. Some motifs such as triangle, feedback loop, feed-forward loop, bi-fan, bi-parallel have been highlighted in literature as building blocks or functional units of many complex networks in the real world [59, 60]. Our method is further applied to estimate the motif count in protein-protein interaction networks of Yeast, Worm, Human, and *Arabidopsis*, as well as in gene regulatory networks of *E.coli* and 41 different cell and tissue types of Human.

Using a different approach, which can be named as “extrapolation”, the authors in [73] scaled up the number of interactions in observed PPI subnetworks to estimate the size of real interactomes, assuming that the link density of a real network can be approximated by the link density of its observed subnetworks. The unbiasedness and consistence, two important requirements of any estimator, however were not justified in this study. Moreover, the effect of experimental errors, that is spurious and missing links, on the estimation has not been considered carefully in [73].

Using the same “extrapolation” approach together with the empirical framework to assess quality parameters in association with Y2H assays mentioned in the previous section, the authors from Center for Cancer Systems Biology, Dana-Farber Cancer Institute, have accurately estimated the interactome size of *Homo sapiens*, *Saccharomyces cerevisiae* (Yeast), *Caenorhabditis elegans* (Worm), and *Arabidopsis thaliana* (Arabidopsis) in [27, 28, 29, 30]. Recently, Rottger *et al.* further applied this method to gene regulatory networks [74]. Thus, they needed to distinguish between two different types of nodes: transcription factors (TFs) and target genes (TGs). Subsequently, they estimated the number of three different types of interactions: TF-regulating-TF, TF-regulating-TG, and TF self-regulations. Unfortunately, the authors did not take into account error rates of the datasets.

In chapter 2 of this thesis, we generalize the “extrapolation” idea in [27, 28, 29, 30, 73, 74] to the case of larger motifs. As mentioned earlier, network motifs are believed to represent functional units of biological processes in living organisms [59, 60, 61]. They have been observed at unusually high frequency in many real-world networks, including biological networks. For example, Mangan and Alon (2004) have carefully studied the structure and function of the feed-forward loop motif, a three-gene pattern which is composed of two input transcription factors, one of which regulates the other, and both jointly regulating a target gene [61]. The authors found that different types of the

feed-forward loop motif can either accelerate or delay the response time of the target gene and the abundance of those motifs in transcription networks can be partially explained by their functionality. Some other examples include single-input modules, bi-fan, dense overlapping regulons, etc [60]. Given their important roles in biological processes, it is highly desirable to detect motifs in a biological network of interest. The key idea to address that problem is to compare the frequency of a motif in the biological network and that in random networks to find out if that motif occurs more often than expected [59, 60, 62]. However, directly counting motifs is difficult and inaccurate due to the incompleteness and the noise in biological networks data. We propose a simple, yet powerful, method to estimate the number of occurrences of different types of motifs in both directed and undirected networks from their observed subnetworks (Fig. 1.4). Next, we perform rigorous theoretical analysis on the properties of the proposed estimators and prove that our proposed estimators are asymptotically unbiased and consistent. Most importantly, the unbiased property holds for any arbitrary motif and regardless of the topological structures of the underlying network. Finally, we further refine the estimation method to take into account spurious and missing interactions, and develop bias-corrected estimators for noisy data.

In chapter 3, the estimators are extensively validated for networks generated from each of the following four widely used random graph models: Erdos-Renyi (ER) [63], preferential attachment [18], duplication [64], and geometric [66] models. We carefully study the accuracy of the proposed estimators with respect to random graph models, network parameters, and sampling parameters. We also perform simulation validation on real biological networks. Both of the theoretical and the simulation results show that our proposed method performs consistently well on all four random network models, suggesting that the method is universal and can be easily applied to any type of networks, including, but not limited to, biological networks, social networks, the

World-Wide-Web, etc.

Finally, we apply our method to estimate the number of different types of motifs in protein-protein interaction networks and gene regulatory networks of four species: Human, Yeast, Worm, and Arabidopsis. Our estimation reveals several interesting features of these networks while only using their noisy observed subnetwork data. For example, we found that the estimated triangle density in Human and Worm are 2.5 times larger than that in Yeast and Arabidopsis, whereas the later have higher link density than the formers, indicating a higher clustering and well-connected structure of the PPI network of Human and Worm. We also discover a strong positive linear correlation between the number of occurrences of different three-node and four-node motifs in forty-one Human cell-specific transcription factor regulatory networks. Our estimation also shows that the feed-forward loop and bi-fan are significantly enriched in these forty-one networks, and the motif counts are highly associated with the functional class of the cell.

### **1.3 Thesis organization**

This thesis is organized as follows. In chapter 2, we present our method to estimate the number of motif occurrences in biological networks from noisy observed subnetworks data. We provide rigorous analysis on the properties of the proposed estimators and prove that they are asymptotically unbiased and consistent. In chapter 3, we first perform extensive simulation validations to study the accuracy of the estimators. Then, we demonstrate how to apply them to real biological networks. In chapter 4, we concludes this thesis with a summary of our contributions and discussion on the limitations of our study, how to address those problems, and potential topics for future research.

## Chapter 2

# Theoretical Analysis for Motif Count Estimation

In this chapter, we propose the method to estimate the number of motif occurrences in biological networks from their noisy observed subnetworks. The problem is schematically illustrated in Figure 1.4. Due to limitations of current biotechnologies, biological networks of most species are not completely known. Their subnetworks are usually inferred with errors, that is, spurious and missing links, from high-throughput experiments such as Yeast Two-Hybrid, Affinity Purification followed by Mass Spectrometry, etc. Spurious links are the links that do not exist in the real network but are wrongly detected by the experiments. Missing links are the links that exist in the real network but cannot be detected by the experiments. In this study, we also refer to spurious links as false positives and missing links as false negatives. Exact motif enumeration in real biological networks is impossible due to the incompleteness (i.e., only subnetworks are inferred) and the inaccuracy (i.e., link errors) in the observed data.

In this proposed method, we first count the number of occurrences of the motif of interest in the observed subnetwork, and then extrapolate to the entire network. In the



second step, our estimation further takes into account the error rates to correct the bias caused by missing and spurious links. We show, theoretically and empirically, that such estimators are asymptotically unbiased and consistent, assuming that the subnetwork is obtained from the original network under a uniformly random node sampling process. More importantly, the results hold for different types of networks and motifs, allowing our method to be easily extended beyond biological networks to apply to any other type of complex networks such as social networks, the World-Wide-Web, engineering and electrical circuitries, etc.

All theoretical analysis on the proposed estimators are given in this chapter. In the next chapter, the estimators are extensively validated in networks generated from four random graph models: Erdos-Renyi [63], preferential attachment [18], duplication [64], and geometric [66] models. For each model, the performance of the estimators is examined with respect to different parameters of the generated networks and the sampling process. In addition to random networks, we also make use of observed PPI subnetworks from real datasets of *S. cerevisiae*, *C. elegans*, *H. sapiens*, and *A. thaliana*. In particular, we consider each observed subnetwork as an entire real network from which we sample even smaller subnetworks (i.e., sub-subnetworks) and then do the estimation & validation.

Details of our method are described in the following sections.

## 2.1 Asymptotically unbiased and consistent estimators

In this section, we develop the estimators for the number of motif occurrences and show that they are asymptotically unbiased and consistent, the two essential properties of any estimator.

First, we introduce some notations that will be used in our analysis. Any network can be represented by a graph which consists of a set of nodes and a set of links. Let  $\mathcal{G}(V, E)$  denote a real biological network, where  $V$  is the set of nodes and  $E = \{(u, v) | u, v \in V, u \neq v\}$  is the set of links. Here nodes represent proteins or genes, and links correspond to their interactions. Note that we do not consider self-interactions. Let  $n$  be the number of nodes in  $\mathcal{G}$ , that is,  $n = |V|$ , and for simplicity, we enumerate the nodes in  $\mathcal{G}$  as  $V = \{1, 2, 3, \dots, n\}$ .

A network can also be represented by an adjacency matrix. Let  $\mathbf{A} = [a_{ij}]_{1 \leq i, j \leq n}$  denote the adjacency matrix of  $\mathcal{G}$ , where  $a_{ij} = 1$  if there is a link from  $i$  to  $j$ , and  $a_{ij} = 0$  otherwise. Note that if  $\mathcal{G}$  is undirected, then  $\mathbf{A}$  is a symmetric matrix, that is,  $a_{ij} = a_{ji}$ . For example, PPI networks are undirected whereas TF regulatory networks are directed.

We use similar notations with the superscript “obs” for observed subnetworks. In particular, let  $\mathcal{G}^{\text{obs}}(V^{\text{obs}}, E^{\text{obs}})$  be an observed subnetwork of the real network  $\mathcal{G}$ . The number of nodes in the observed subnetwork is denoted as  $n^{\text{obs}}$ , that is,  $n^{\text{obs}} = |V^{\text{obs}}|$ . Our goal is to estimate the number of motif occurrences in the real network  $\mathcal{G}$  from the observed subnetwork  $\mathcal{G}^{\text{obs}}$ .

In order to do that estimation, we need to know the relationship between the real network  $\mathcal{G}$  and the observed subnetwork  $\mathcal{G}^{\text{obs}}$ , in particular, how  $\mathcal{G}^{\text{obs}}$  is obtained from  $\mathcal{G}$ . Following [73], we model the observed subnetwork  $\mathcal{G}^{\text{obs}}$  as the outcome of a uniformly random node sampling process in the following sense: each node from  $V$  is independently sampled with some probability  $p$ ,  $0 < p < 1$ ; and the subgraph induced from  $\mathcal{G}$  by the sampled nodes is the observed subnetwork  $\mathcal{G}^{\text{obs}}$ .

More specifically, this uniformly random node sampling scheme can be modelled using a Bernoulli process. Let independent random variables  $X_i \sim \text{Bernoulli}(p)$  denote whether node  $i$  is sampled ( $X_i = 1$ ) or not ( $X_i = 0$ ),  $1 \leq i \leq n$ . Then,  $V^{\text{obs}}$  is the

set of nodes  $i$  with  $X_i = 1$  (sampled nodes), and  $E^{\text{obs}}$  is induced from  $E$  by  $V^{\text{obs}}$ , that is,  $E^{\text{obs}}$  is a subset of  $E$  that consists of all links connecting the sampled nodes. It should be noted that by “induced” we mean there is neither spurious nor missing links in  $E^{\text{obs}}$ . In this section, we first introduce our method for the case when the observed subnetwork  $\mathcal{G}^{\text{obs}}$  is free from experimental errors (that is, there is neither missing nor spurious links), and then generalize it to handle noisy observed subnetworks in the next section.

For the clarity of presentation, we first describe the estimator for the simplest motif type, that is, the number of links in an undirected network. The analysis will then be generalized to handle any arbitrary motif in both undirected or directed networks.

### 2.1.1 Estimator for the number of links in an undirected network

Consider the case when the real network  $\mathcal{G}(V, E)$  is undirected. Let  $N_1$  and  $N_1^{\text{obs}}$  denote the number of links in  $\mathcal{G}$  and  $\mathcal{G}^{\text{obs}}$  respectively. The number of links (and any motifs) of any network can always be computed from its adjacency matrix. In the case of the observed subnetwork  $\mathcal{G}^{\text{obs}}$ , instead of using its adjacency matrix  $A^{\text{obs}}$ , we can also use the adjacency matrix  $\mathbf{A}$  of the real network  $\mathcal{G}$  and Bernoulli random variables  $X_i$ ,  $1 \leq i \leq n$ . In particular,  $N_1$  and  $N_1^{\text{obs}}$  can be written as following:

$$N_1 = \sum_{1 \leq i_1 < i_2 \leq n} a_{i_1 i_2}, \quad (2.1)$$

$$N_1^{\text{obs}} = \sum_{1 \leq i_1 < i_2 \leq n} a_{i_1 i_2} (X_{i_1} X_{i_2}), \quad (2.2)$$

where  $a_{i_1 i_2} (X_{i_1} X_{i_2})$  implies that one can observe link  $(i_1, i_2)$  in  $\mathcal{G}^{\text{obs}}$  if and only if link  $(i_1, i_2)$  exists in  $\mathcal{G}$  (that is,  $a_{i_1 i_2} = 1$ ), and both nodes  $i_1$  and  $i_2$  are sampled (that is,

$X_{i_1}X_{i_2} = 1$ ).

Remind that in this work we study the problem of estimating the number of occurrences of any arbitrary motif in the real network  $\mathcal{G}$  using the observed subnetwork  $\mathcal{G}^{obs}$ . In particular, we assume that the real network  $\mathcal{G}$  and its adjacency matrix  $\mathbf{A}$  are not fully known due to the low coverage and inaccuracy of biological networks data. The only available information is the observed subnetwork  $\mathcal{G}^{obs}$  and the total number of nodes in the real network  $\mathcal{G}$ ,  $n$ , which corresponds to the number of genes (or proteins) in the species under consideration. Thus,  $N_1$  is unknown and to be estimated using  $n$  and  $N_1^{obs}$ , which can be counted in  $\mathcal{G}^{obs}$ .

Under the uniformly random node sampling process, one could expect that the real network  $\mathcal{G}$  and the observed subnetwork  $\mathcal{G}^{obs}$  should have many similar properties. In particular, the link density in the real network  $\mathcal{G}$  is intuitively very close to that in the observed subnetwork  $\mathcal{G}^{obs}$ . This gives rise to the following estimator

$$\hat{N}_1 = \frac{\binom{n}{2}}{\binom{n^{obs}}{2}} N_1^{obs}. \quad (2.3)$$

This is the “extrapolation” approach for estimating the size of interactomes that we have reviewed in chapter 1. This idea has been widely used to estimate the number of interactions in PPI networks of different species in [27, 28, 29, 30, 73]. However, the unbiased and consistent properties of this estimator have not been considered thoroughly in previous works. In the following theorem, we first prove that the estimator  $\hat{N}_1$  is asymptotically unbiased. More importantly, this property holds for any topological structure of the underlying network  $\mathcal{G}$ , thus making this estimator widely applicable to any real-world undirected networks.

**Theorem 1.** *Let  $\mathcal{G}$  be an arbitrary undirected network of  $n$  nodes, and  $\mathcal{G}^{obs}$  be a sub-network obtained from  $\mathcal{G}$  via a uniformly random node sampling process that selects*

a node with probability  $p$ ,  $0 < p < 1$ . Let  $N_1$  denote the number of links in  $\mathcal{G}$ ,  $N_1^{obs}$  denote the number of links in  $\mathcal{G}^{obs}$ , and  $N_1$  is estimated by the estimator  $\widehat{N}_1$  defined in Equation (2.3). We have:

$$E\left(\frac{\widehat{N}_1}{N_1}\right) = 1 - q^n - npq^{n-1}, \quad (2.4)$$

where  $q = 1 - p$ . Therefore,  $\widehat{N}_1$  is an asymptotically unbiased estimator for  $N_1$  in the sense that  $E\left(\widehat{N}_1/N_1\right) \rightarrow 1$  as  $n$  goes to infinity. Moreover, the convergence is exponentially fast in  $n$ .

*Proof.* By Equations (2.2) and (2.3),

$$\begin{aligned} E(\widehat{N}_1) &= E\left(\frac{\binom{n}{2}}{\binom{n^{obs}}{2}} \sum_{1 \leq i_1 < i_2 \leq n} a_{i_1 i_2}(X_{i_1} X_{i_2})\right) \\ &= E\left(\frac{n(n-1)}{n^{obs}(n^{obs}-1)} \sum_{1 \leq i_1 < i_2 \leq n} a_{i_1 i_2}(X_{i_1} X_{i_2})\right) \\ &= n(n-1) \sum_{1 \leq i_1 < i_2 \leq n} a_{i_1 i_2} E\left(\frac{X_{i_1} X_{i_2}}{n^{obs}(n^{obs}-1)}\right). \end{aligned} \quad (2.5)$$

Since random variables  $X_i$  are independent and identically distributed, for any  $1 \leq i_1 < i_2 \leq n$ , we have

$$E\left(\frac{X_{i_1} X_{i_2}}{n^{obs}(n^{obs}-1)}\right) = E\left(\frac{X_1 X_2}{n^{obs}(n^{obs}-1)}\right). \quad (2.6)$$

Subsequently, we have

$$\begin{aligned}
E(\widehat{N}_1) &= n(n-1) \sum_{1 \leq i_1 < i_2 \leq n} a_{i_1 i_2} E\left(\frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)}\right) \\
&= n(n-1) E\left(\frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)}\right) N_1 \\
E\left(\frac{\widehat{N}_1}{N_1}\right) &= n(n-1) E\left(\frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)}\right). \tag{2.7}
\end{aligned}$$

The number of nodes in the observed subnetwork  $\mathcal{G}^{\text{obs}}$ ,  $n^{\text{obs}}$ , can be written as:

$$n^{\text{obs}} = X_1 + X_2 + \dots + X_n. \tag{2.8}$$

By conditioning on the event that  $X_1 = X_2 = 1$ , we rewrite  $n^{\text{obs}}$  as

$$(n^{\text{obs}} | X_1 = X_2 = 1) = Z + 2, \tag{2.9}$$

where  $Z \sim \text{Binomial}(n-2, p)$ , since random variables  $X_i$  are independent.

Subsequently, we have

$$E\left(\frac{\widehat{N}_1}{N_1}\right) = n(n-1)p^2 E\left(\frac{1}{(Z+2)(Z+1)}\right). \tag{2.10}$$

It can be shown that

$$\begin{aligned}
E\left(\frac{1}{(Z+2)(Z+1)}\right) &= E \int_0^1 (1-u)u^Z du \\
&= \int_0^1 (1-u)Eu^Z du \\
&= \int_0^1 (1-u)[q+pu]^{n-2} du \\
&= \frac{1 - q^n - npq^{n-1}}{n(n-1)p^2},
\end{aligned}$$

where the first and the last equations are obtained by integration by parts, the second equation is obtained by interchange between integral and expectation, and the third equation is obtained from the moment generating function of the random variable  $Z \sim \text{Binomial}(n - 2, p)$ .

Finally, we have

$$E \left( \frac{\widehat{N}_1}{N_1} \right) = 1 - q^n - npq^{n-1}.$$

□

Thus, we have shown that the estimator  $\widehat{N}_1$  is an asymptotically unbiased estimator for  $N_1$  in the sense that  $E \left( \widehat{N}_1/N_1 \right)$  tends to one exponentially fast in  $n$ . We also study the consistent property of the estimator  $\widehat{N}_1$ . In the following theorem, we first obtain the compact close-form expression for  $Var \left( \widehat{N}_1/N_1 \right)$ . We notice that the variation of the estimator  $\widehat{N}_1$  depends on the topological structure of the underlying network  $\mathcal{G}$ . Subsequently, we show that the variance goes to zero when  $\mathcal{G}$  is generated from four widely used random graph models in studies of biological networks, including Erdos-Renyi, preferential attachment, duplication, and geometric models. This indicates that the estimator  $\widehat{N}_1$  is consistent for a wide class of random networks.

**Theorem 2.** *Let  $\mathcal{G}$  be an arbitrary undirected network of  $n$  nodes, and  $\mathcal{G}^{obs}$  be a sub-network obtained from  $\mathcal{G}$  via a uniformly random node sampling process that selects a node with probability  $p$ ,  $0 < p < 1$ . Let  $N_1$  denote the number of links in  $\mathcal{G}$ , and  $N_1$  is estimated by the estimator  $\widehat{N}_1$  defined in Equation (2.3). Let  $N_2$  denote the number of three-node paths in  $\mathcal{G}$  (that is, the number of pairs of links that share exactly one common node, see motif with ID  $u_2$  in Figure 2.1). We have:*

$$Var \left( \frac{\widehat{N}_1}{N_1} \right) = \frac{2q}{p} \frac{N_2}{N_1^2} [1 + O(n^{-1})] + \frac{(1+p)q}{p^2 N_1} [1 + O(n^{-1})] + O(n^{-1}), \quad (2.11)$$

where  $q = 1 - p$ .

The convergence of the variance in Equation (2.11) is dominated by the term  $\frac{N_2}{N_1^2}$ , which subsequently depends on  $N_1$  and  $N_2$ . In the following proposition, we obtain the convergence rate of  $\frac{N_2}{N_1^2}$ , and hence, show that  $\text{Var}\left(\widehat{N}_1/N_1\right)$  tends to zero as  $n$  goes to infinity.

**Proposition 1.** *When  $\mathcal{G}$  is generated by the Erdos-Renyi, preferential attachment, duplication, or geometric models, the corresponding convergence rate of  $\frac{N_2}{N_1^2}$  is as following:*

- *ER model:  $O(n^{-1})$*
- *Preferential attachment model:  $O\left(\frac{\log(n)}{n}\right)$*
- *Partial duplication model: let  $\beta$  be the approximated exponent of the power-law degree distribution of  $G$ , we have*
  - $\beta = 2$ :  $O\left(\frac{1}{(\log(n))^2}\right)$
  - $2 < \beta < 3$ :  $O\left(\frac{1}{n^{\beta-2}}\right)$
  - $\beta = 3$ :  $O\left(\frac{\log(n)}{n}\right)$
  - $\beta > 3$ :  $O(n^{-1})$ .
- *Geometric model:  $O(n^{-1})$*

Thus,  $\text{Var}\left(\widehat{N}_1/N_1\right) \rightarrow 0$  as  $n \rightarrow \infty$ .

The detailed proofs of Theorem 2 and Proposition 1 are too long to present here, and they can be found in the Appendix. It is important to note that the convergence rate for networks generated from the Erdos-Renyi and geometric models are faster than that for networks generated from the preferential attachment and duplication models. This is due to the fact that the former models generate networks with symmetric structure,



whereas the latter ones generate networks with the scale-free structure which subsequently increases the variation among sampled subnetworks, and hence, the variation of the estimator  $\widehat{N}_1$ .

We also note that the duplication model has a wide range of convergence rate, depending on the exponent of the power-law degree distribution  $\beta$ . When  $\beta = 3$ , the duplication model has the same convergence rate  $O(\frac{\log(n)}{n})$  as the preferential attachment model since both models have the same exponent. When  $\beta > 3$ , the duplication model has the same convergence rate  $O(n^{-1})$  as the ER model and the geometric model. In general, the convergence is faster when the exponent is higher, because there will be more nodes with low degree and less nodes with high degree, which subsequently makes the variation become smaller.

We wish to point out that the above notions “asymptotically unbiased” and “consistent” are not in the usual statistical sense where the population is fixed and the number of observations increases to infinity. In the context of biological networks, the number of subnetworks observed from high-throughput experiments (that is, “the number of observations”) is limited to just a few for each species. On the other hand, the genome size or proteome size of species (that is,  $n$ , the number of nodes in  $\mathcal{G}$ ) is sufficiently large, ranging from  $\sim 6,000$  for yeast [9] up to  $\sim 22,000$  for human [2] or  $\sim 27,000$  for Arabidopsis [12]. Thus it is reasonable to study the asymptotic properties of the estimators when the number of nodes,  $n$ , in the underlying network  $\mathcal{G}$  is sufficiently large. Indeed, our simulation validation in the next chapter will show that the proposed estimators perform accurately for  $n$  within the range of the genome size of the model organisms under consideration.

In summary, we have shown in Theorem 1, Theorem 2, and Proposition 1 that the simple estimator  $\widehat{N}_1$  derived in Equation 2.3 is asymptotically unbiased and consistent for estimating the number of links in the real network  $\mathcal{G}$  from the observed subnetwork

$\mathcal{G}^{obs}$ . Most importantly, Theorem 1 and Theorem 2 hold for any arbitrary network  $\mathcal{G}$ . In Proposition 1, we have presented the convergence rate of for a wide class of random graph models that are often used in studies of real-world complex networks. Thus, the estimator  $\widehat{N}_1$  is widely applicable to any network of interest. In the next section, we further generalize this idea to estimate the number of occurrences of any arbitrary motif.

### 2.1.2 Estimator for an arbitrary motif $\mathcal{M}$










The results obtained in the previous section can be generalized to the case of larger motifs in both directed and undirected networks. Let  $\mathcal{M}$  denote an arbitrary motif. Let  $N_{\mathcal{M}}$  and  $N_{\mathcal{M}}^{obs}$  denote the number of occurrences of  $\mathcal{M}$  in  $\mathcal{G}$  and  $\mathcal{G}^{obs}$  respectively. In general, the number of occurrences of any motif can be written in terms of the adjacency matrix  $\mathbf{A}$  and the Bernoulli random variables  $X_i$ ,  $1 \leq i \leq n$ , in a similar way to  $N_1$  and  $N_1^{obs}$  in Eqn. (2.1) and (2.2). In particular, the number of occurrences of  $\mathcal{M}$  in the whole network  $\mathcal{G}$ ,  $N_{\mathcal{M}}$ , can be written as a function of  $\mathbf{A}$ . Similarly,  $N_{\mathcal{M}}^{obs}$  can be written as a function of the adjacency matrix of  $\mathcal{G}^{obs}$ , or equivalently a function of  $\mathbf{A}$  and Bernoulli random variables  $X_i$ ,  $1 \leq i \leq n$ . For a subset  $J \subseteq \{1, 2, \dots, n\}$ , let  $\mathbf{A}[J]$  denote the submatrix consisting of entries in the rows and columns indexed by  $J$ . Then, for any motif  $\mathcal{M}$  of  $m$  nodes, we have:

$$N_{\mathcal{M}} = \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\mathbf{A}[i_1, i_2, \dots, i_m]), \quad (2.12)$$

$$N_{\mathcal{M}}^{obs} = \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\mathbf{A}[i_1, i_2, \dots, i_m]) X_{i_1} X_{i_2} \dots X_{i_m}, \quad (2.13)$$

where function  $f_{\mathcal{M}}()$  is suitably chosen to count the number of occurrences of motif  $\mathcal{M}$  among any  $m$  nodes  $i_1, i_2, \dots, i_m$ . Equation 2.13 can be interpreted as follows: we can observe motif  $\mathcal{M}$  among nodes  $i_1, i_2, \dots, i_m$  in the subnetwork  $\mathcal{G}^{obs}$  if and only if












Table 2.1: Detailed expressions of function  $f_{\mathcal{M}}()$  for 9 undirected motifs. For any motif  $\mathcal{M}$  of  $m$  nodes, function  $f_{\mathcal{M}}()$  is suitably chosen to count the number of occurrences of motif  $\mathcal{M}$  among any  $m$  nodes  $i_1 < i_2 < \dots < i_m$ .

ID	Motif	$f_{\mathcal{M}}()$
1		$f_1(\mathbf{A}[i_1, i_2]) = a_{i_1 i_2}$
2		$f_2(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_1 i_3} + a_{i_2 i_1} a_{i_2 i_3} + a_{i_3 i_1} a_{i_3 i_2}$
3		$f_3(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_1}$
4		$f_4(\mathbf{A}[i_1, i_2, i_3, i_4]) = a_{i_1 i_2} a_{i_1 i_3} a_{i_1 i_4} + a_{i_2 i_1} a_{i_2 i_3} a_{i_2 i_4} + a_{i_3 i_1} a_{i_3 i_2} a_{i_3 i_4} + a_{i_4 i_1} a_{i_4 i_2} a_{i_4 i_3}$
5 <sup>(*)</sup>		$f_5(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum a_{ij} (a_{ik} a_{jl} + a_{il} a_{jk})$
6 <sup>(**)</sup>		$f_6(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum a_{ij} a_{jk} a_{ki} (a_{il} + a_{jl} + a_{kl})$
7		$f_7(\mathbf{A}[i_1, i_2, i_3, i_4]) = a_{i_1 i_3} a_{i_1 i_4} a_{i_2 i_3} a_{i_2 i_4} + a_{i_1 i_2} a_{i_1 i_4} a_{i_3 i_2} a_{i_3 i_4} + a_{i_1 i_2} a_{i_1 i_3} a_{i_4 i_2} a_{i_4 i_3}$
8 <sup>(*)</sup>		$f_8(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum a_{ik} a_{il} a_{jk} a_{jl} a_{kl}$
9		$f_9(\mathbf{A}[i_1, i_2, i_3, i_4]) = a_{i_1 i_2} a_{i_1 i_3} a_{i_1 i_4} a_{i_2 i_3} a_{i_2 i_4} a_{i_3 i_4}$

(\*) The sum is taken over all possible combinations ( $i < j$ ) chosen from  $\{i_1, i_2, i_3, i_4\}$ , and ( $k < l$ ) being the two remaining nodes (totally, there are 6 such combinations).

(\*\*) The sum is taken over all possible combinations ( $i < j < k$ ) chosen from  $\{i_1, i_2, i_3, i_4\}$ , and  $l$  being the remaining node (totally, there are 4 such combinations).

Table 2.2: Detailed expressions of function  $f_{\mathcal{M}}()$  for 11 directed motifs. For any motif  $\mathcal{M}$  of  $m$  nodes, function  $f_{\mathcal{M}}()$  is suitably chosen to count the number of occurrences of motif  $\mathcal{M}$  among any  $m$  nodes  $i_1 < i_2 < \dots < i_m$ .

ID	Motif	$f_{\mathcal{M}}()$
10		$f_{10}(\mathbf{A}[i_1, i_2]) = a_{i_1 i_2} + a_{i_2 i_1}$
11		$f_{11}(\mathbf{A}[i_1, i_2, i_3]) = (a_{i_2 i_1} a_{i_1 i_3} + a_{i_3 i_1} a_{i_1 i_2}) + (a_{i_1 i_2} a_{i_2 i_3} + a_{i_3 i_2} a_{i_2 i_1}) + (a_{i_1 i_3} a_{i_3 i_2} + a_{i_2 i_3} a_{i_3 i_1})$
12		$f_{12}(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_1 i_3} + a_{i_2 i_1} a_{i_2 i_3} + a_{i_3 i_1} a_{i_3 i_2}$
13		$f_{13}(\mathbf{A}[i_1, i_2, i_3]) = a_{i_2 i_1} a_{i_3 i_1} + a_{i_1 i_2} a_{i_3 i_2} + a_{i_1 i_3} a_{i_2 i_3}$
14		$f_{14}(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_2 i_3} a_{i_3 i_1} + a_{i_3 i_2} a_{i_2 i_1} a_{i_1 i_3}$
15		$f_{15}(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_1 i_3} (a_{i_2 i_3} + a_{i_3 i_2}) + a_{i_2 i_1} a_{i_2 i_3} (a_{i_1 i_3} + a_{i_3 i_1}) + a_{i_3 i_1} a_{i_3 i_2} (a_{i_1 i_2} + a_{i_2 i_1})$
16 <sup>(*)</sup>		$f_{16}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum a_{ik} a_{il} (a_{jk} + a_{jl}) + a_{jk} a_{jl} (a_{ik} + a_{il})$
17 <sup>(*)</sup>		$f_{17}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum (a_{ik} a_{jl} + a_{il} a_{jk}) (a_{kl} + a_{lk})$
18 <sup>(*)</sup>		$f_{18}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum (a_{ik} a_{jl} + a_{il} a_{jk}) (a_{ij} + a_{ji})$
19 <sup>(*)</sup>		$f_{19}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum a_{ik} a_{il} a_{jk} a_{jl}$
20 <sup>(*)</sup>		$f_{20}(\mathbf{A}[i_1, i_2, i_3, i_4]) = \sum a_{ik} a_{il} a_{kj} a_{lj} + a_{jk} a_{jl} a_{ki} a_{li}$

(\*) The sum is taken over all possible combinations ( $i < j$ ) chosen from  $\{i_1, i_2, i_3, i_4\}$ , and ( $k < l$ ) being the two remaining nodes (totally, there are 6 such combinations).

motif  $\mathcal{M}$  exists in the whole network  $\mathcal{G}$  (indicated by function  $f_{\mathcal{M}}()$ ) and all  $m$  nodes  $i_1, i_2, \dots, i_m$  are sampled (indicated by the product  $X_{i_1}X_{i_2}\dots X_{i_m}$ ).

Table 2.1 give the expressions of function  $f_{\mathcal{M}}()$  for all possible undirected motifs that have up to 4 nodes. There are totally 9 of them. Table 2.2 give the expressions of function  $f_{\mathcal{M}}()$  for 11 selected directed motifs. In this section, we will develop the estimators for these 20 motifs, some of which such as feed-forward loop (FFL), bi-fan, bi-parallel have been highlighted in literature as building blocks or functional units in many real-world complex networks [59].

Similar to the case of the number of links in an undirected network with  $N_1$ ,  $N_1^{obs}$ , and  $\widehat{N}_1$ , we want to develop an estimator  $\widehat{N}_{\mathcal{M}}$  for  $N_{\mathcal{M}}$ , using  $n$  and  $N_{\mathcal{M}}^{obs}$ . We generalize the ‘‘extrapolation’’ idea of link density in the previous section to the case of larger motifs: for any motif  $\mathcal{M}$ , its density in the real network  $\mathcal{G}$  can be approximated by that in the observed subnetwork  $\mathcal{G}^{obs}$ . This motivates us to consider the following estimator:

$$\widehat{N}_{\mathcal{M}} = \frac{\binom{n}{m}}{\binom{n^{obs}}{m}} N_{\mathcal{M}}^{obs}. \quad (2.14)$$

We obtain the following theorem, which is an generalized extension of Theorem 1. In particular, we show that the estimator  $\widehat{N}_{\mathcal{M}}$  is asymptotically unbiased for any arbitrary motif  $\mathcal{M}$ . Most importantly, this theorem does not require any assumption regarding to the underlying network  $\mathcal{G}$ . Thus, the proposed estimators can be applied to any real-world complex network from diverse fields, including, but not limited to, biological networks, social networks, the World-Wide-Web, engineering and electrical circuitries, etc.

**Theorem 3.** *Let  $\mathcal{G}$  be an arbitrary network of  $n$  nodes, and  $\mathcal{G}^{obs}$  be a subnetwork obtained from  $\mathcal{G}$  via a uniformly random node sampling process that selects a node with probability  $p$ ,  $0 < p < 1$ . For any motif  $\mathcal{M}$  of  $m$  nodes, the estimator  $\widehat{N}_{\mathcal{M}}$  defined in*

Eqn. (2.14) satisfies

$$E \left( \frac{\widehat{N}_{\mathcal{M}}}{N_{\mathcal{M}}} \right) = 1 - q^n - npq^{n-1} - \dots - \binom{n}{j} p^j q^{n-j} - \dots - \binom{n}{m-1} p^{m-1} q^{n-(m-1)},$$

where  $q = 1 - p$ . Therefore,  $\widehat{N}_{\mathcal{M}}$  is an asymptotically unbiased estimator for  $N_{\mathcal{M}}$  in the sense that  $E \left( \widehat{N}_{\mathcal{M}}/N_{\mathcal{M}} \right) \rightarrow 1$  as  $n$  goes to infinity. Moreover, the convergence is exponentially fast in  $n$ .

*Proof.* The proof is similar to the proof of Theorem 1.

Since random variables  $X_i$  are *i.i.d.*, for any  $1 \leq i_1 < i_2 < \dots < i_m \leq n$ , we have

$$E \left( \frac{X_{i_1} X_{i_2} \dots X_{i_m}}{n^{\text{obs}}(n^{\text{obs}} - 1) \dots (n^{\text{obs}} - m + 1)} \right) = E \left( \frac{X_1 X_2 \dots X_m}{n^{\text{obs}}(n^{\text{obs}} - 1) \dots (n^{\text{obs}} - m + 1)} \right). \quad (2.15)$$

Thus, by Equations (2.12), (2.13), and (2.14), we have

$$\begin{aligned} E \left( \frac{\widehat{N}_{\mathcal{M}}}{N_{\mathcal{M}}} \right) &= E \left( \frac{1}{N_{\mathcal{M}}} \frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\mathbf{A}[i_1, i_2, \dots, i_m]) X_{i_1} X_{i_2} \dots X_{i_m} \right) \\ &= \binom{n}{m} \frac{\sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\mathbf{A}[i_1, i_2, \dots, i_m]) E \left( \frac{X_{i_1} X_{i_2} \dots X_{i_m}}{\binom{n^{\text{obs}}}{m}} \right)}{N_{\mathcal{M}}} \\ &= n(n-1) \dots (n-m+1) E \left( \frac{X_1 X_2 \dots X_m}{n^{\text{obs}}(n^{\text{obs}} - 1) \dots (n^{\text{obs}} - m + 1)} \right) \end{aligned} \quad (2.16)$$

By conditioning on the event that  $X_1 = X_2 = \dots = X_m = 1$ , we rewrite  $n^{\text{obs}}$  as

$$n^{\text{obs}} = Z + m, \quad (2.17)$$

where  $Z \sim \text{Binomial}(n - m, p)$ . Subsequently, we have

$$E \left( \frac{\widehat{N}_{\mathcal{M}}}{N_{\mathcal{M}}} \right) = n(n-1) \dots (n-m+1) p^m E \left( \frac{1}{(Z+m)(Z+m-1) \dots (Z+1)} \right). \quad (2.18)$$

It can be shown by integration by parts that

$$\frac{1}{(Z+m)(Z+m-1)\dots(Z+1)} = \int_0^1 \frac{(1-u)^{m-1}}{(m-1)!} u^Z du. \quad (2.19)$$

Thus, we further have

$$\begin{aligned} E\left(\frac{\widehat{N}_{\mathcal{M}}}{N_{\mathcal{M}}}\right) &= p^m n(n-1)\dots(n-m+1) E\left(\int_0^1 \frac{(1-u)^{m-1}}{(m-1)!} u^Z du\right) \\ &= p^m n(n-1)\dots(n-m+1) \left(\int_0^1 \frac{(1-u)^{m-1}}{(m-1)!} E(u^Z) du\right) \\ &= p^m n(n-1)\dots(n-m+1) \left(\int_0^1 \frac{(1-u)^{m-1}}{(m-1)!} (q+pu)^{n-m} du\right) \\ &= 1 - q^n - npq^{n-1} - \dots - \binom{n}{j} p^j q^{n-j} - \dots - \binom{n}{m-1} p^{m-1} q^{n-(m-1)}, \end{aligned}$$

where in the second equality we use interchange between expectation and integral, in the third equality we use the moment generating function of  $Z \sim \text{Binomial}(n-m, p)$ , and in the last equality we use integration by parts.  $\square$

To show the consistent property of the estimator  $\widehat{N}_{\mathcal{M}}$ , one needs to study the variation of the ratio  $\frac{\widehat{N}_{\mathcal{M}}}{N_{\mathcal{M}}}$ , like what has been done for  $\frac{\widehat{N}_1}{N_1}$  in Theorem 2. However, it is a difficult task to work out the close-form of the variance of  $\widehat{N}_{\mathcal{M}}$  for an arbitrary motif  $\mathcal{M}$ . Thus, we decided to show the consistent property of the estimator  $\widehat{N}_{\mathcal{M}}$  via simulation results in chapter 3.

*Summary:* in this section 2.1, we have developed the estimators of the number of occurrences of any arbitrary motif  $\mathcal{M}$  in any directed or undirected network. First, the analysis was done for the simplest case, i.e. the number of links in an undirected network, and then was generalized to any arbitrary motif  $\mathcal{M}$ . Using the ‘‘extrapolation’’ approach, we scaled up the motif count in the observed subnetwork  $\mathcal{G}^{\text{obs}}$  to estimate that in the real network  $\mathcal{G}$ . Although the idea is very intuitive, little effort has been

done to analytically explore the properties of these estimators. A previous attempt by Stumpf *et al.* in [73] only focused the estimation of the number of links, and even the analysis there is not complete yet. To the best of our knowledge, our study is the first work to thoroughly explore the asymptotically unbiased and consistent of the estimators  $\widehat{N}_{\mathcal{M}}$ . Most importantly, we have proved that the results obtained here hold for any motif  $\mathcal{M}$  and any underlying network  $\mathcal{G}$ .

However, it should be noted that we have assumed so far that the observed subnetwork  $\mathcal{G}^{\text{obs}}$  is free from experimental errors, that is, there is neither spurious nor missing links in  $\mathcal{G}^{\text{obs}}$ . In the next section 2.2, we shall further refine the method to handle noise in the observed subnetwork  $\mathcal{G}^{\text{obs}}$ .

## 2.2 Noisy subnetwork data and biased-corrected estimators

As mentioned earlier in section 1.2, the low coverage and inaccuracy of data prevail in Network Biology as one of the most challenging problems. In this section, we refine the estimator  $\widehat{N}_{\mathcal{M}}$  developed in the previous section 2.1 to take into account the error rates of real biological networks data.

There are two types of errors in biological networks data: spurious interactions (that is, false positives) and missing interactions (that is, false negatives) (see Fig. 1.4). Spurious links are the links that do not exist in the real network but are wrongly detected by the experiments. Missing links are the links that exist in the real network but cannot be detected by the experiments. We define the false positive rate  $r_+$  to be the probability that a non-existing link is incorrectly detected, and the false negative rate  $r_-$  to be the probability that a true link is not detected. Using independent random variables  $F_{i_1 i_2}^+ \sim \text{Bernoulli}(r_+)$  and  $F_{i_1 i_2}^- \sim \text{Bernoulli}(r_-)$ ,  $1 \leq i_1 < i_2 \leq n$ , to model



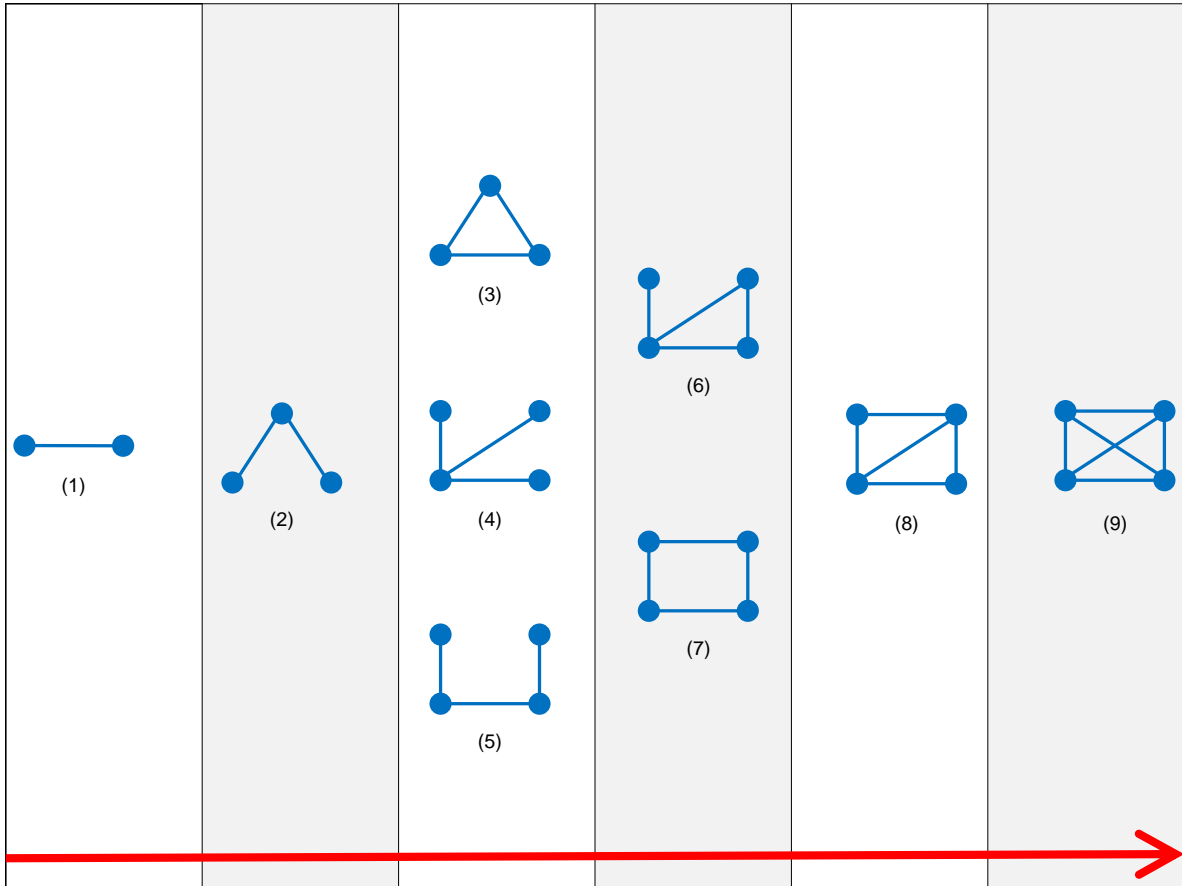


Figure 2.1: All possible (9) undirected motifs that have up to 4 nodes. The number associated with each motif indicates its reference ID in the main text. The long red arrow indicates the sub-motif relationship: motifs in each column are sub-motifs of the ones on their right side.

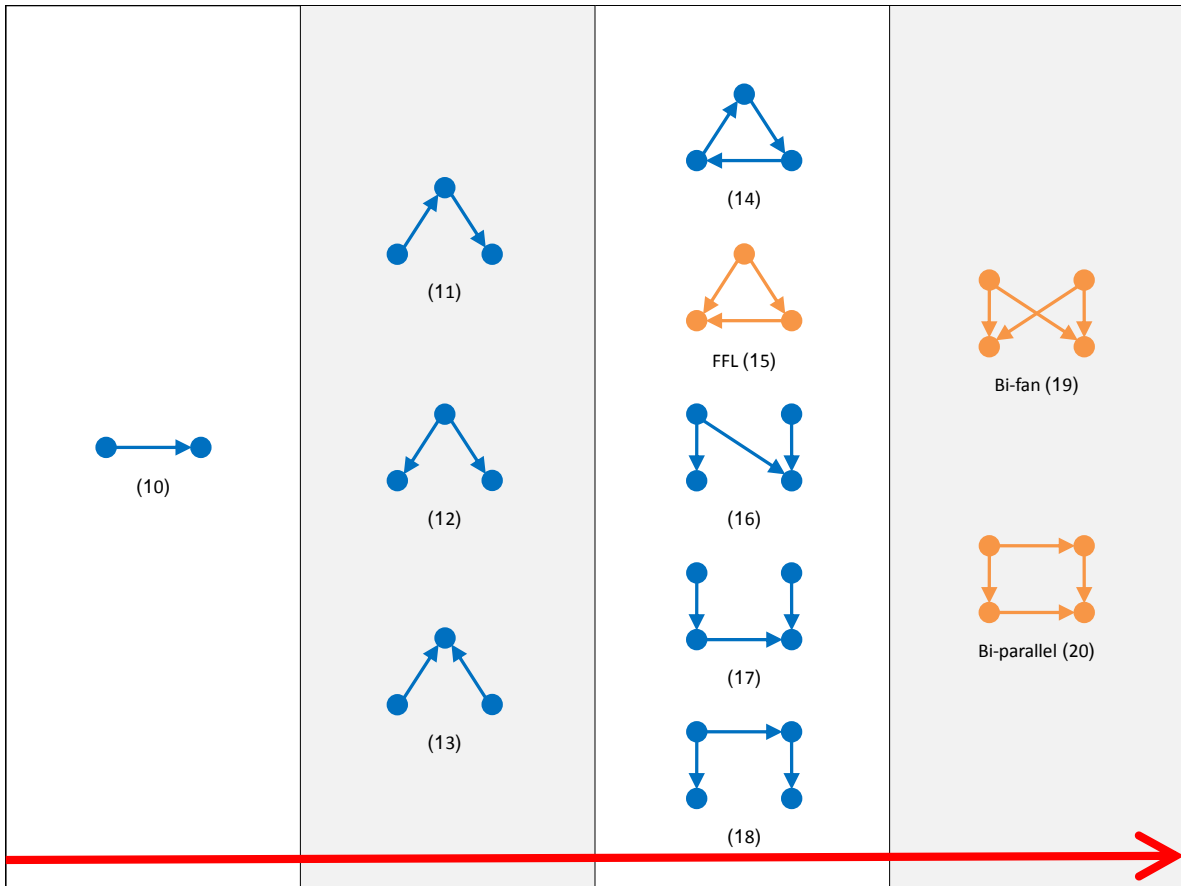


Figure 2.2: 11 selective directed motifs, some of which such as feed-forward loop, bi-fan, bi-parallel have been highlighted in literature as building blocks or functional units in many real-world complex networks [59]. The number associated with each motif indicates its reference ID in the main text. The long red arrow indicates the sub-motif relationship: motifs in each column are sub-motifs of the ones on their right side.

spurious and missing interactions in the observed subnetwork  $\mathcal{G}^{\text{obs}}$ , we can represent the effect of experimental errors on an ordered pair of nodes  $(i_1, i_2)$  as follows:

$$\tilde{a}_{i_1 i_2} = a_{i_1 i_2}(1 - F_{i_1 i_2}^-) + (1 - a_{i_1 i_2})F_{i_1 i_2}^+. \quad (2.20)$$

In other words, for any two nodes  $i_1, i_2 \in V^{\text{obs}}$ , a link  $(i_1, i_2)$  is observed in the subnetwork  $\mathcal{G}^{\text{obs}}$  (that is,  $\tilde{a}_{i_1 i_2} = 1$ ) if one of the following two mutually exclusive situations happens. In the first situation, there is an link  $(i_1, i_2)$  in the real network  $\mathcal{G}$  (that is,  $a_{i_1 i_2} = 1$ ), and there is no false negative (that is,  $F_{i_1 i_2}^- = 0$ ). This situation gives rise to the term  $a_{i_1 i_2}(1 - F_{i_1 i_2}^-)$ . In the second situation, link  $(i_1, i_2)$  does not exist in the real network  $\mathcal{G}$  (that is,  $a_{i_1 i_2} = 0$ ), but false positive occurs (that is,  $F_{i_1 i_2}^+ = 1$ ). This situation gives rise to the term  $(1 - a_{i_1 i_2})F_{i_1 i_2}^+$ .

To take these error rates into account, we need to replace each entry  $a_{i_1 i_2}$  in the adjacency matrix  $\mathbf{A}$  with  $\tilde{a}_{i_1 i_2}$  to obtain a new matrix,  $\tilde{\mathbf{A}}$ , and then replace  $\mathbf{A}$  by  $\tilde{\mathbf{A}}$  in the expression of  $N_{\mathcal{M}}^{\text{obs}}$  in Eqn. (2.13). As a result, the proposed estimator  $\hat{N}_{\mathcal{M}}$  defined in Eqn. (2.14) is no longer asymptotically unbiased. In particular,  $\hat{N}_{\mathcal{M}}$  has the new expression as follows:

$$\hat{N}_{\mathcal{M}} = \frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\tilde{\mathbf{A}}[i_1, i_2, \dots, i_m]) X_{i_1} X_{i_2} \dots X_{i_m}. \quad (2.21)$$

Since the random variables  $X_i$  are *i.i.d.*,  $1 \leq i \leq n$ , and we assume that they are

also independent of  $F_{i_1 i_2}^+$  and  $F_{i_1 i_2}^-$ ,  $1 \leq i_1 < i_2 \leq n$ , we have

$$\begin{aligned}
E(\widehat{N}_{\mathcal{M}}) &= E\left(\frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\widetilde{\mathbf{A}}[i_1, i_2, \dots, i_m]) X_{i_1} X_{i_2} \dots X_{i_m}\right) \\
&= E\left(\frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} \sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\widetilde{\mathbf{A}}[i_1, i_2, \dots, i_m]) X_1 X_2 \dots X_m\right) \\
&= E\left(\frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} X_1 X_2 \dots X_m\right) E\left(\sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\widetilde{\mathbf{A}}[i_1, i_2, \dots, i_m])\right) \quad (2.22)
\end{aligned}$$

As shown in the proof of Theorem 3, we have

$$\begin{aligned}
E\left(\frac{\binom{n}{m}}{\binom{n^{\text{obs}}}{m}} X_1 X_2 \dots X_m\right) &= 1 - q^n - npq^{n-1} - \dots - \binom{n}{j} p^j q^{n-j} - \dots - \binom{n}{m-1} p^{m-1} q^{n-(m-1)} \\
&= 1 - \sum_{j=0}^{m-1} \binom{n}{j} p^j q^{n-j} \\
&\Rightarrow 1 \text{ as } n \rightarrow \infty.
\end{aligned}$$

On the other hand, we can work out the second expectation in Eqn. (2.22) using the fact that  $F_{i_1 i_2}^+ \sim \text{Bernoulli}(r_+)$ ,  $F_{i_1 i_2}^- \sim \text{Bernoulli}(r_-)$ ,  $1 \leq i_1 < i_2 \leq n$ , and they are independent. In particular, the second expectation has the following form:

$$E\left(\sum_{1 \leq i_1 < i_2 < \dots < i_m \leq n} f_{\mathcal{M}}(\widetilde{\mathbf{A}}[i_1, i_2, \dots, i_m])\right) = (1 - r_+ - r_-)^s N_{\mathcal{M}} + W_{\mathcal{M}}, \quad (2.23)$$

where  $s$  is the number of links in motif  $\mathcal{M}$ ,  $W_{\mathcal{M}}$  is a function of  $n$ , the error rates  $r_-$  and  $r_+$ , and the number of occurrences of all sub-motifs  $\mathcal{M}'$  of  $\mathcal{M}$ , that is,  $N_{\mathcal{M}'}$ .

Hence, to correct the bias caused by the error rates, we first need to estimate  $N_{\mathcal{M}'}$  by  $\widetilde{N}_{\mathcal{M}'}$  for all sub-motifs  $\mathcal{M}'$  of  $\mathcal{M}$ . Subsequently, we obtain  $\widetilde{W}_{\mathcal{M}}$ , and adjust  $\widehat{N}_{\mathcal{M}}$  to

$$\widetilde{N}_{\mathcal{M}} = \frac{1}{r^s} (\widehat{N}_{\mathcal{M}} - \widetilde{W}_{\mathcal{M}}), \quad (2.24)$$

where

$$r = 1 - r_+ - r_-. \quad (2.25)$$

Thus, we have used a re-centering factor  $\widetilde{W}_{\mathcal{M}}$  and a scaling factor  $\frac{1}{r^s}$  to correct the bias caused by spurious and missing links. The bias-corrected estimator  $\widetilde{N}_{\mathcal{M}}$  given in Eqn. (2.24) now can be used to estimate the number of occurrences of any arbitrary motif  $\mathcal{M}$  in the entire network  $\mathcal{G}$  from its noisy observed subnetwork  $\mathcal{G}^{\text{obs}}$ . Note that our method recursively estimates the number of occurrences of a motif based on its sub-motifs. Fig. 2.1 and Fig. 2.2 respectively show the sub-motif relationship for 9 undirected motifs and 11 directed motifs considered in this study. Expressions of the bias-corrected estimator  $\widetilde{N}_{\mathcal{M}}$  are given in Table 2.3 (undirected motifs) and Table 2.4 (directed motifs). Detailed calculation of the bias-corrected estimator  $\widetilde{N}_{\mathcal{M}}$  for those motifs are omitted from this thesis since they are too long. Nevertheless, in the following section, we give an example of how to obtain the bias-corrected estimator  $\widetilde{N}_{\mathcal{M}}$  for the feed-forward loop motif.

### 2.2.1 Example of calculating the bias-corrected estimator $\widetilde{N}_{\mathcal{M}}$ for the feed-forward loop motif

For any three nodes  $1 \leq i_1 < i_2 < i_3 \leq n$ , the following function counts the number of occurrences of feed-forward loop (FFL) motif among them (Table 2.2):

$$f_{15}(\mathbf{A}[i_1, i_2, i_3]) = a_{i_1 i_2} a_{i_1 i_3} (a_{i_2 i_3} + a_{i_3 i_2}) + a_{i_2 i_1} a_{i_2 i_3} (a_{i_1 i_3} + a_{i_3 i_1}) + a_{i_3 i_1} a_{i_3 i_2} (a_{i_1 i_2} + a_{i_2 i_1}). \quad (2.26)$$

The numbers of occurrences of FFL motif in the real network  $\mathcal{G}$  and its observed

Table 2.3: Detailed expressions of the bias-corrected estimator  $\tilde{N}_{\mathcal{M}}$  for 9 undirected motifs.





















ID	Motif	Bias-corrected estimator $\tilde{N}_{\mathcal{M}}$
1		$\tilde{N}_1 = \frac{1}{r} \left[ \hat{N}_1 - \binom{n}{2} r_+ \right]$
2		$\tilde{N}_2 = \frac{1}{r^2} \left[ \hat{N}_2 - 2(n-2)r_+ r \tilde{N}_1 - 3 \binom{n}{3} r_+^2 \right]$
3		$\tilde{N}_3 = \frac{1}{r^3} \left[ \hat{N}_3 - r_+ r^2 \tilde{N}_2 - (n-2)r_+^2 r \tilde{N}_1 - \binom{n}{3} r_+^3 \right]$
4		$\tilde{N}_4 = \frac{1}{r^3} \left[ \hat{N}_4 - (n-3)r_+ r^2 \tilde{N}_2 - 2 \binom{n-2}{2} r_+^2 r \tilde{N}_1 - 4 \binom{n}{4} r_+^3 \right]$
5		$\tilde{N}_5 = \frac{1}{r^3} \left\{ \hat{N}_5 - r_+ r^2 \left[ 4 \binom{\tilde{N}_1}{2} + 2(n-5) \tilde{N}_2 \right] - 6 \binom{n-2}{2} r_+^2 r \tilde{N}_1 - 12 \binom{n}{4} r_+^3 \right\}$
6		$\tilde{N}_6 = \frac{1}{r^4} \left\{ \hat{N}_6 - r_+ r^3 \left[ 3(n-3) \tilde{N}_3 + 3 \tilde{N}_4 + 2 \tilde{N}_5 \right] - r_+^2 r^2 \left[ 4 \binom{\tilde{N}_1}{2} + (5n-19) \tilde{N}_2 \right] - 8 \binom{n-2}{2} r_+^3 r \tilde{N}_1 - 12 \binom{n}{4} r_+^4 \right\}$
7		$\tilde{N}_7 = \frac{1}{r^4} \left\{ \hat{N}_7 - r_+ r^3 \tilde{N}_5 - r_+^2 r^2 \left[ 2 \binom{\tilde{N}_1}{2} + (n-5) \tilde{N}_2 \right] - 2 \binom{n-2}{2} r_+^3 r \tilde{N}_1 - 3 \binom{n}{4} r_+^4 \right\}$
8		$\tilde{N}_8 = \frac{1}{r^5} \left\{ \hat{N}_8 - 2r_+ r^4 (\tilde{N}_6 + \tilde{N}_7) - 3r_+^2 r^3 \left[ (n-3) \tilde{N}_3 + \tilde{N}_4 + \tilde{N}_5 \right] - 4r_+^3 r^2 \left[ \binom{\tilde{N}_1}{2} + (n-4) \tilde{N}_2 \right] - 5 \binom{n-2}{2} r_+^4 r \tilde{N}_1 - 6 \binom{n}{4} r_+^5 \right\}$
9		$\tilde{N}_9 = \frac{1}{r^6} \left\{ \hat{N}_9 - r_+ r^5 \tilde{N}_8 - r_+^2 r^4 (\tilde{N}_6 + \tilde{N}_7) - r_+^3 r^3 \left[ (n-3) \tilde{N}_3 + \tilde{N}_4 + \tilde{N}_5 \right] - r_+^4 r^2 \left[ \binom{\tilde{N}_1}{2} + (n-4) \tilde{N}_2 \right] - \binom{n-2}{2} r_+^5 r \tilde{N}_1 - \binom{n}{4} r_+^6 \right\}$

Table 2.4: Detailed expressions of the bias-corrected estimator  $\tilde{N}_{\mathcal{M}}$  for 11 directed motifs.

ID	Motif	Bias-corrected estimator $\tilde{N}_{\mathcal{M}}$
10		$\tilde{N}_{10} = \frac{1}{r} \left[ \hat{N}_{10} - 2\binom{n}{2}r_+ \right]$
11		$\tilde{N}_{11} = \frac{1}{r^2} \left[ \hat{N}_{11} - 2(n-2)r_+r\tilde{N}_{10} - 6\binom{n}{3}r_+^2 \right]$
12		$\tilde{N}_{12} = \frac{1}{r^2} \left[ \hat{N}_{12} - (n-2)r_+r\tilde{N}_{10} - 3\binom{n}{3}r_+^2 \right]$
13		$\tilde{N}_{13} = \frac{1}{r^2} \left[ \hat{N}_{13} - (n-2)r_+r\tilde{N}_{10} - 3\binom{n}{3}r_+^2 \right]$
14		$\tilde{N}_{14} = \frac{1}{r^3} \left[ \hat{N}_{14} - r_+r^2\tilde{N}_{11} - (n-2)r_+^2r\tilde{N}_{10} - 2\binom{n}{3}r_+^3 \right]$
15		$\tilde{N}_{15} = \frac{1}{r^3} \left[ \hat{N}_{15} - r_+r^2(\tilde{N}_{11} + 2\tilde{N}_{12} + 2\tilde{N}_{13}) - 3(n-2)r_+^2r\tilde{N}_{10} - 6\binom{n}{3}r_+^3 \right]$
16		$\tilde{N}_{16} = \frac{1}{r^3} \left\{ \hat{N}_{16} - 2r_+r^2 \left[ 2\binom{\tilde{N}_{10}}{2} + (n-3)(\tilde{N}_{12} + \tilde{N}_{13}) \right] - 6\binom{n-2}{2}r_+^2r\tilde{N}_{10} - 24\binom{n}{4}r_+^3 \right\}$
17		$\tilde{N}_{17} = \frac{1}{r^3} \left\{ \hat{N}_{17} - r_+r^2 \left[ 2\binom{\tilde{N}_{10}}{2} + (n-3)(\tilde{N}_{11} + 2\tilde{N}_{13}) \right] - 6\binom{n-2}{2}r_+^2r\tilde{N}_{10} - 24\binom{n}{4}r_+^3 \right\}$
18		$\tilde{N}_{18} = \frac{1}{r^3} \left\{ \hat{N}_{18} - r_+r^2 \left[ 2\binom{\tilde{N}_{10}}{2} + (n-3)(\tilde{N}_{11} + 2\tilde{N}_{12}) \right] - 6\binom{n-2}{2}r_+^2r\tilde{N}_{10} - 24\binom{n}{4}r_+^3 \right\}$
19		$\tilde{N}_{19} = \frac{1}{r^4} \left\{ \hat{N}_{19} - r_+r^3\tilde{N}_{16} - r_+^2r^2 \left[ 2\binom{\tilde{N}_{10}}{2} + (n-3)(\tilde{N}_{12} + \tilde{N}_{13}) \right] - 2\binom{n-2}{2}r_+^3r\tilde{N}_{10} - 6\binom{n}{4}r_+^4 \right\}$
20		$\tilde{N}_{20} = \frac{1}{r^4} \left\{ \hat{N}_{20} - r_+r^3(\tilde{N}_{17} + \tilde{N}_{18}) - r_+^2r^2 \left[ 2\binom{\tilde{N}_{10}}{2} + (n-3)(\tilde{N}_{11} + \tilde{N}_{12} + \tilde{N}_{13}) \right] - 4\binom{n-2}{2}r_+^3r\tilde{N}_{10} - 12\binom{n}{4}r_+^4 \right\}$

subnetwork  $\mathcal{G}^{\text{obs}}$ , given that there is no error in  $\mathcal{G}^{\text{obs}}$ , are written as follows:

$$N_{15} = \sum_{1 \leq i_1 < i_2 < i_3 \leq n} f_{15}(\mathbf{A}[i_1, i_2, i_3]), \quad (2.27)$$

$$N_{15}^{\text{obs}} = \sum_{1 \leq i_1 < i_2 < i_3 \leq n} f_{15}(\mathbf{A}[i_1, i_2, i_3]) X_{i_1} X_{i_2} X_{i_3}. \quad (2.28)$$

Assuming that the density of FFL motif in  $\mathcal{G}$  can be approximated by that in  $\mathcal{G}^{\text{obs}}$ , the following estimator can be used to estimate the number of occurrences of FFL motif in  $\mathcal{G}$ , given that there is no error in  $\mathcal{G}^{\text{obs}}$ :

$$\widehat{N}_{15} = \frac{\binom{n}{3}}{\binom{n^{\text{obs}}}{3}} N_{15}^{\text{obs}}. \quad (2.29)$$

Taking error rates into account, we replace each entry  $a_{i_1 i_2}$  in the adjacency matrix  $\mathbf{A}$  with  $\tilde{a}_{i_1 i_2}$ , defined in Eqn. (2.20), to obtain a new matrix  $\tilde{\mathbf{A}}$ , and then replace  $\mathbf{A}$  by  $\tilde{\mathbf{A}}$  in the expressions of  $N_{15}^{\text{obs}}$  and  $\widehat{N}_{15}$ . As a result, the estimator  $\widehat{N}_{15}$  is no longer unbiased. In particular, its expectation has the following form:

$$\begin{aligned} E(\widehat{N}_{15}) &= E \left( \frac{\binom{n}{3}}{\binom{n^{\text{obs}}}{3}} \sum_{1 \leq i_1 < i_2 < i_3 \leq n} f_{15}(\tilde{\mathbf{A}}[i_1, i_2, i_3]) X_{i_1} X_{i_2} X_{i_3} \right) \\ &= E \left( \frac{\binom{n}{3}}{\binom{n^{\text{obs}}}{3}} \sum_{1 \leq i_1 < i_2 < i_3 \leq n} f_{15}(\tilde{\mathbf{A}}[i_1, i_2, i_3]) X_1 X_2 X_3 \right) \\ &= E \left( \frac{\binom{n}{3}}{\binom{n^{\text{obs}}}{3}} X_1 X_2 X_3 \right) E \left( \sum_{1 \leq i_1 < i_2 < i_3 \leq n} f_{15}(\tilde{\mathbf{A}}[i_1, i_2, i_3]) \right) \\ &= \left[ 1 - q^n - npq^{n-1} - \binom{n}{2} p^2 q^{n-2} \right] [(1 - r_+ - r_-)^3 N_{15} + W_{15}], \quad (2.30) \end{aligned}$$

where in the second and the third equalities we use the assumption that  $X_i$  are *i.i.d.*,  $1 \leq i \leq n$ , and independent of  $F_{i_1 i_2}^+$  and  $F_{i_1 i_2}^-$ ,  $1 \leq i_1 < i_2 \leq n$ .



The function  $W_{15}$  has the following form (detailed calculation is omitted):

$$W_{15} = r_+ r^2 (N_{11} + 2N_{12} + 2N_{13}) + 3(n-2)r_+^2 r N_{10} + 6 \binom{n}{3} r_+^3, \quad (2.31)$$

in which  $r = 1 - r_+ - r_-$ ,  $N_{10}, N_{11}, N_{12}, N_{13}$  are the numbers of occurrences of the corresponding sub-motifs of FFL motif (Fig. 2.2).

Subsequently, we replace  $N_{10}, N_{11}, N_{12}, N_{13}$  by their bias-corrected estimators and obtain

$$\tilde{N}_{15} = \frac{1}{r^3} \left[ \hat{N}_{15} - r_+ r^2 (\tilde{N}_{11} + 2\tilde{N}_{12} + 2\tilde{N}_{13}) - 3(n-2)r_+^2 r \tilde{N}_{10} - 6 \binom{n}{3} r_+^3 \right]. \quad (2.32)$$

Thus,  $\tilde{N}_{15}$  is the bias-corrected estimator for the number of occurrences of FFL motif (Table 2.4).

*Summary:* in this section 2.2, we have refined the estimator  $\hat{N}_{\mathcal{M}}$  developed in section 2.1 by taking into account the false positive and false negative rates. As a result, the bias-corrected estimator  $\tilde{N}_{\mathcal{M}}$  now can be used to handle noise in the observed subnetwork  $\mathcal{G}^{\text{obs}}$ . To the best of our knowledge, our bias correction approach is the first attempt in Network Biology, as well as in other fields of Network Sciences, to **directly** estimate the number of occurrences of motifs from **noisy** subnetwork data. As discussed in the Introduction, previous works only focused on direct estimation of the number of links, or indirect approaches that try to reconstruct the real network by inferring spurious and missing links.

## 2.3 Summary

In this chapter, we have proposed a method to estimate the number of occurrences of motifs in a real network from its noisy subnetwork data. Rigorous theoretical analysis

has been done to thoroughly explore the asymptotically unbiased and consistent properties of the proposed estimators. Interestingly, the results have been obtained without any assumption regarding to the structure of the real network and the type of the motif of interest. Thus, the proposed estimators are widely applicable to different fields of Network Sciences, including Network Biology, Social Networks, World-Wide-Web, etc. Most importantly, our method is the first attempt that can take into account spurious and missing links to directly estimate the number of occurrences of motifs from noisy subnetwork data. In the next chapter, we shall present the simulation results and the analysis on real network datasets that further confirm the accuracy of the estimators  $\hat{N}_{\mathcal{M}}$ ,  $\tilde{N}_{\mathcal{M}}$  and our theoretical results obtained in this chapter.

## Chapter 3

# Simulation Validation and Application to Protein-Protein Interaction & Gene Regulatory Networks

In this chapter, we first present the simulation validation for the theoretical results obtained in chapter 2. We study the accuracy of the two estimators  $\hat{N}_{\mathcal{M}}$  and  $\tilde{N}_{\mathcal{M}}$  with respect to different parameters of the whole network  $\mathcal{G}$ , as well as parameters of the sampling process used to obtain the subnetwork  $\mathcal{G}^{\text{obs}}$ . In particular, we consider four random graph models: Erdos-Renyi [63], preferential attachment [18], duplication [64], and geometric [66]. As mentioned in the chapter Introduction, these four models are the most widely used in Network Sciences because they are able to capture several topological structures of real-worlds networks. Parameters required to generate network  $\mathcal{G}$  from these random graph models include the number of nodes  $n$ , the link density  $\rho$ , and the power-law exponent  $\beta$  for scale-free networks. Sampling parameters include the

sampling probability  $p$  and the error rates  $r_+, r_-$ . At this point we only focus on the uniform random node sampling scheme, more detailed discussion on sampling schemes will be given in the next chapter. In addition to random graph models, which can only explain some, but not all characteristics of any complex network in the real-world, we also examine the performance of the proposed estimators in “real” networks. In particular, we consider observed subnetworks from real data sets as the entire complete network  $\mathcal{G}$  from which we draw smaller subnetworks, that is, sub-subnetworks, and then perform similar estimation and validation. Detailed simulation results given in the following sections confirm the accuracy of our proposed estimators and the theoretical results obtained in chapter 2, including Theorems 1, 2, 3 and convergence rates in Proposition 1.

In the second part of this chapter, we apply our method to estimate the number of occurrences of different motif types from real datasets of protein-protein interaction (PPI) and gene regulatory networks. Our results reveal several interesting features of the PPI networks in four species: *S. cerevisiae* (Yeast), *C. elegans* (Worm), *H. sapiens* (Human), and *A. thaliana* (Arabidopsis), as well as the transcription factor (TF) regulatory networks in forty-one different cell types of Human. In particular, we found that although PPI networks of Yeast and Arabidopsis have similar, or even higher link densities than Human, the triangle density in the PPI network of Human is 5 times that of Yeast, and 1.7 times that of Arabidopsis. This indicates a highly clustering and well-connected structure of the PPI network of Human. We also discovered a very strong positive linear correlation between the number of occurrences of important triad and quadriad motifs in forty-one cell-specific TF regulatory networks of Human. Moreover, the numbers of occurrences of motifs seem to be associated with the functional class of the cell types. Details are given in the following sections.

## 3.1 Simulation validation

### 3.1.1 Simulation from random graph models

In this section, we examine the accuracy of the proposed estimators in networks generated from four random graph models: Erdos-Renyi, preferential attachment, duplication, and geometric. Detailed properties and features of these models have been discussed in the chapter Introduction. For the clarity of presentation, here we briefly describe again how a network  $\mathcal{G}$  is generated from these random graph models.

1. Erdos-Renyi model:

- We start with  $n$  singletons.
- Between any two nodes a link is placed independently and uniformly at random with probability  $\rho$ .

2. Preferential attachment model:

- An initial random network  $\mathcal{G}_0$  is generated from the ER model with 10 nodes and edge density 0.5.
- At each iteration, a new node with  $l$  incident links is added to the current network. Neighbors of the newly added node are chosen with probabilities proportional to their current degrees.

3. Duplication model:

- An initial random network  $\mathcal{G}_0$  is generated from the ER model with 10 nodes and edge density 0.5.
- At each iteration, an existing node  $u$  is chosen uniformly at random. A new node  $u'$  is duplicated from  $u$ , that is,  $u'$  is connected to each neighbor of  $u$

with probability  $p_{dup}$ . The new node  $u'$  is also connected to the duplicated node  $u$ .

#### 4. Geometric model:

- First,  $n$  nodes are placed uniformly at random in a unit cube.
- Any two nodes are then connected if the distance between them is less than a given threshold  $\delta$ .

Although these four models have their own parameters, for a meaningful comparison, we will study the accuracy of the proposed estimators with respect to two common important properties: the number of nodes,  $n$ , and the link density,  $\rho$ . The parameters of each model are then adjusted so that the model can produce networks with desired  $n$  and  $\rho$ .

The simulation and validation process is carried out as follows. Using each model, we generate directed and undirected networks with different values of the number of nodes  $n$  and the link density  $\rho$ . Each generated network is then considered as the real network  $\mathcal{G}$  from which 100 subnetworks are sampled using the uniformly random node sampling scheme with different values of the sampling probability  $p$ . For each motif  $\mathcal{M}$ , we count its occurrences in each sampled subnetwork, that is  $N_{\mathcal{M}}^{\text{obs}}$ , and scale up to obtain the estimator  $\widehat{N}_{\mathcal{M}}$ , as described in chapter 2. To proceed further with the noisy case, spurious and missing interactions are planted to each of those 100 sampled subnetworks using different values of the error rates  $r_+$  and  $r_-$ . Then, the bias-corrected estimator  $\widetilde{N}_{\mathcal{M}}$  is calculated from those noisy sampled subnetworks using the formula derived in chapter 2.

Finally, we compare the estimators  $\widehat{N}_{\mathcal{M}}$  and  $\widetilde{N}_{\mathcal{M}}$  with the real number  $N_{\mathcal{M}}$ , which is calculated directly from  $\mathcal{G}$ . The accuracy of each estimator depends on the random graph model, the network parameters  $n$  and  $\rho$ , and the node sampling probability  $p$ .

The bias-corrected estimator also depends on the error rates  $r_+$  and  $r_-$ . We use the mean square error (MSE) of the ratios  $\frac{\widehat{N}_{\mathcal{M}}}{N_{\mathcal{M}}}$  and  $\frac{\widetilde{N}_{\mathcal{M}}}{N_{\mathcal{M}}}$  to measure the accuracy of the estimators. As we expect the ratios to be close to one, the MSE is calculated as follows:

$$MSE\left(\frac{\widehat{N}_{\mathcal{M}}}{N_{\mathcal{M}}}\right) = \frac{1}{100} \sum_{1 \leq i \leq 100} \left(\frac{\widehat{N}_{\mathcal{M}}^{(i)}}{N_{\mathcal{M}}} - 1\right)^2, \quad (3.1)$$

where  $\widehat{N}_{\mathcal{M}}^{(i)}$  is calculated from the sampled subnetwork  $i$ ,  $1 \leq i \leq 100$ . The MSE of  $\frac{\widetilde{N}_{\mathcal{M}}}{N_{\mathcal{M}}}$  is calculated similarly. Since we have proved that the estimators are asymptotically unbiased, the MSE will basically reflect their variation.

First, we study how the MSE of the estimators  $\widehat{N}_{\mathcal{M}}$  and  $\widetilde{N}_{\mathcal{M}}$  depends on the parameters of the underlying network  $\mathcal{G}$ . We fix the node sampling probability  $p = 0.1$ , that is, only 10% of the nodes from  $\mathcal{G}$  will be sampled. We choose the error rates  $r_- = 0.85$  and  $r_+ = 0.00001$ , which are similar to the error rates in real datasets (more details of real datasets will be given in the next section). For each of the four random graph models, we consider increasing number of nodes  $n = 500, 1000, \dots, 10000$  and increasing link density  $\rho = 0.01, 0.02, \dots, 0.1$ . For each combination of model,  $n$ , and  $\rho$ , a network  $\mathcal{G}$  is generated. The sampling and estimation process is then carried out as described above.

Panels (A) and (B) in Fig. 3.1 show the MSE of the estimators  $\widehat{N}_9$  and  $\widetilde{N}_9$  for the number of occurrences of FFL motif with respect to the network parameters  $n$  and  $\rho$ , when the underlying network  $\mathcal{G}$  is generated from the ER model. In both cases, for each value of the link density  $\rho$ , the MSE decreases and converges to zero as  $n$  increases. This confirms that our proposed estimators are asymptotically unbiased and consistent. The MSE also decreases as the link density  $\rho$  increases, indicating that the estimators have better accuracy when applied to denser networks. If we compare panel (A) versus panel (B), it can be seen that for the same values of the parameters  $n$  and  $p$ , the MSE

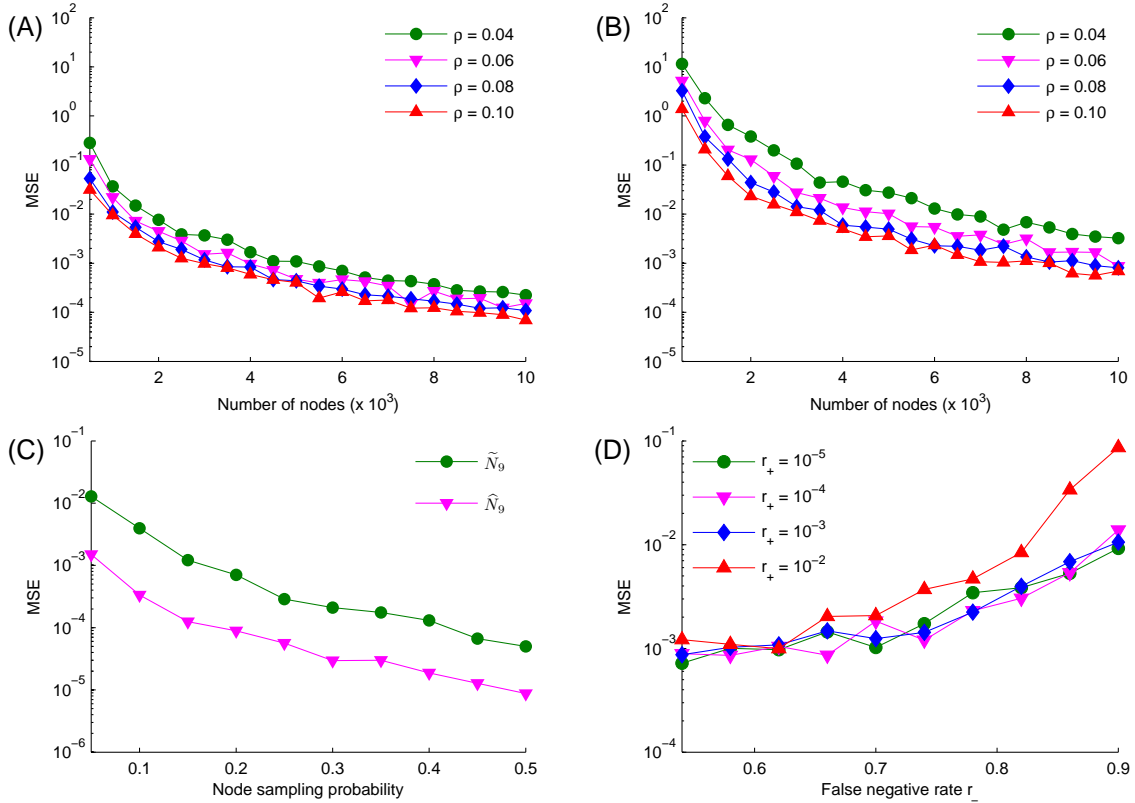


Figure 3.1: MSE of the estimators  $\hat{N}_9$  and  $\tilde{N}_9$  for the number of occurrences of FFL motif in networks generated from the ER model. **(A)** MSE of the estimator  $\hat{N}_9$ , **(B)** MSE of the estimator  $\tilde{N}_9$  with respect to the number of nodes  $n$  and the link density  $\rho$ . For each value of the link density  $\rho$ , we generate networks with increasing number of nodes,  $n = 500, 1,000, \dots, 10,000$ . From each network, we sample 100 subnetworks with the node sampling probability  $p = 0.1$ , and calculate the estimator  $\hat{N}_9$  from each sampled subnetwork. Missing and spurious links are then introduced to each sampled subnetwork with the error rates  $r_+ = 0.00001$ ,  $r_- = 0.85$ , and the estimator  $\tilde{N}_9$  is calculated from each noisy sampled subnetwork. **(C)** MSE of the estimators  $\hat{N}_9$  and  $\tilde{N}_9$  with respect to the node sampling probability  $p$ . We first generate a network with the number of nodes  $n = 5,000$  and the link density  $\rho = 0.1$ , and repeat the sampling-estimating process with increasing node sampling probability  $p = 0.05, 0.1, \dots, 0.5$ , while the error rates are fixed at  $r_+ = 0.00001$  and  $r_- = 0.85$ . **(D)** MSE of the estimator  $\tilde{N}_9$  with respect to the error rates  $r_+$  and  $r_-$ , when the other parameters are fixed at  $n = 5,000$ ,  $\rho = 0.1$ ,  $p = 0.1$ .



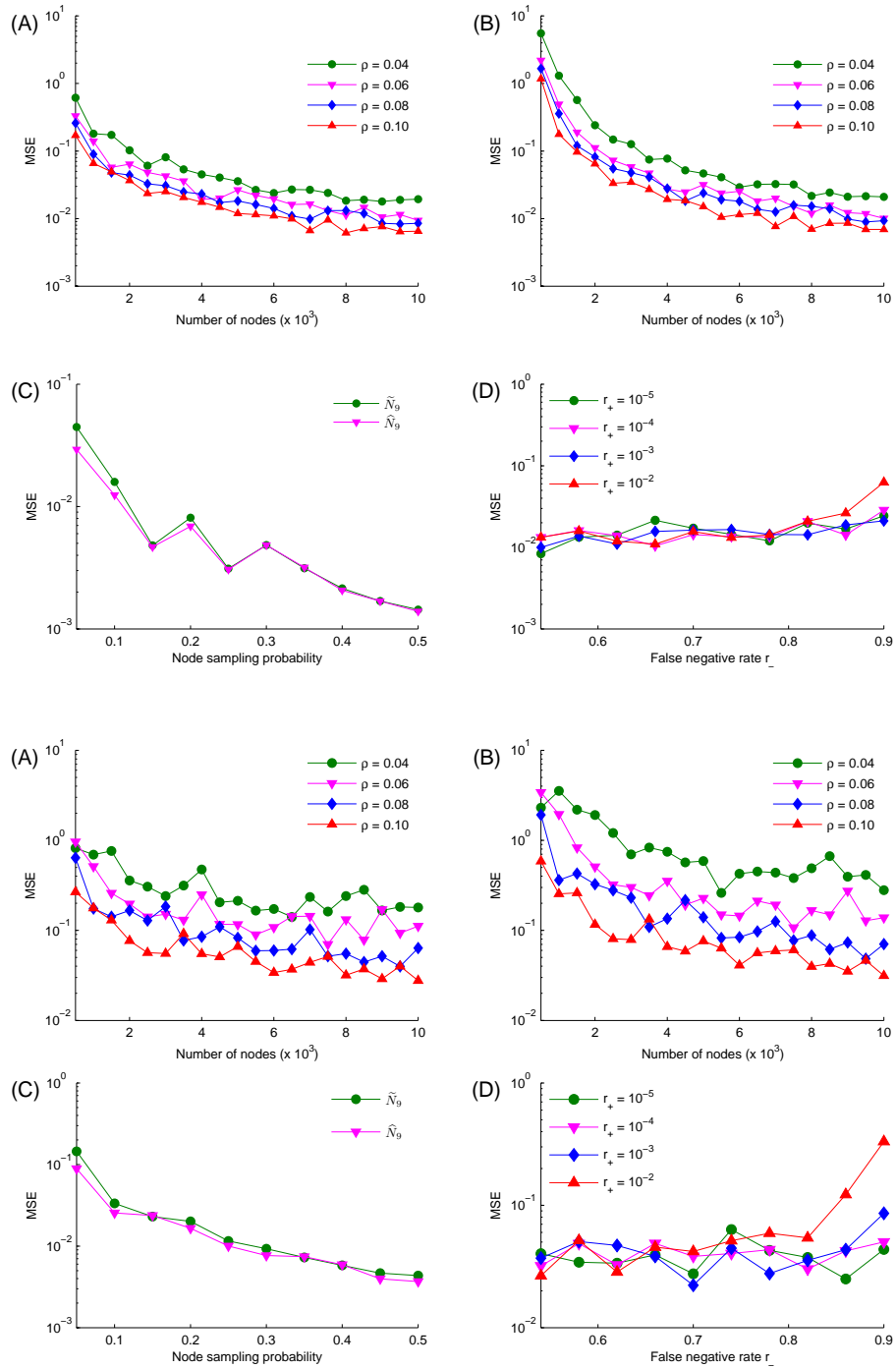


Figure 3.2: MSE of the estimators  $\widehat{N}_9$  and  $\widetilde{N}_9$  for the number of occurrences of FFL motif in networks generated from the preferential attachment (upper) and the duplication (lower) models. Notations are similar to Fig. 3.1.

of the bias-corrected estimator  $\tilde{N}_9$  is higher than that of  $\hat{N}_9$ . This is not surprising since  $\tilde{N}_9$  suffers from extra fluctuation of the random noise added into the sampled subnetworks.

Next, we study how the accuracy of the estimators  $\hat{N}_9$  and  $\tilde{N}_9$  depends on the parameters of the sampling scheme. We first generate a network with the number of nodes  $n = 5,000$  and the link density  $\rho = 0.1$ , and repeat the sampling-estimating process with increasing node sampling probability  $p = 0.05, 0.1, \dots, 0.5$ , while the error rates are fixed as  $r_- = 0.85$  and  $r_+ = 0.00001$ . As shown in panel (C) of Fig. 3.1, the accuracy of the estimators increases (MSE decreases) as the node sampling probability increases. This is to be expected because the larger the sampled subnetworks are, the smaller the variation of the estimators is. Similar to panels (A) and (B), it can also be seen clearly from panel (C) that the MSE of  $\tilde{N}_9$  is higher than that of  $\hat{N}_9$ .

We also study the effect of the error rates  $r_+$  and  $r_-$  on the estimators. We fix the number of nodes  $n = 5,000$ , the link density  $\rho = 0.1$ , the sampling probability  $p = 0.1$ , and repeat the sampling-estimating process with different values of the false positive rate  $r_+$  and the false negative rate  $r_-$ . Since the false negative rate  $r_-$  is  $\sim 0.85$  for real datasets, whereas the false positive rate  $r_+$  is  $\sim 10^{-5}$ , which is much smaller than  $r_-$ , we consider different values of  $r_-$  ranging from 0.5 to 0.9, and different values of  $r_+$  ranging from  $10^{-5}$  to  $10^{-2}$ . Panel (D) of Fig. 3.1 shows that the MSE increases with respect to the false negative rate  $r_-$ , whereas the false positive rate  $r_+$  has little effect on the MSE because it is too low. There is almost no difference in the MSE for the three cases  $r_+ = 10^{-5}, 10^{-4}, 10^{-3}$ .

Fig. 3.2 shows similar results for networks generated from the preferential attachment and the duplication models. It can be seen from panels (A) and (B) that the MSE for networks generated from the ER model is lower than for those generated from the preferential attachment and the duplication models, especially for the estimator  $\hat{N}_9$ . As

mentioned in section 1.1.4 in chapter 1, networks generated from the ER model have a symmetric architecture. Thus, subnetworks which are sampled uniformly at random from the original network tend to have similar structure. As a result, the variation of the estimators calculated from sampled subnetworks will be smaller for the ER model than for the other two models. This observation is similar to the theoretical results of the variation of the estimator for the number of links in undirected networks,  $\widehat{N}_1$ , which have been proved in Proposition 1 in chapter 2.

On the other hand, if we look at the difference between the MSE of the two estimators  $\widehat{N}_9$  and  $\widetilde{N}_9$  illustrated in panel (C) of Fig. 3.2, it can be seen that there is almost no difference for the preferential attachment and the duplication models. In contrast, there is a consistent difference between  $\widehat{N}_9$  and  $\widetilde{N}_9$  for the ER model in panel (C) of Fig. 3.1. This could be explained by the robustness of networks generated from the preferential attachment and the duplication models. As mentioned in section 1.1.4 in chapter 1, networks generated from these two models are scale-free, and hence, are robust against random attacks. As a result, the bias-corrected estimator is less affected by the random noise added into the sampled subnetworks.

Similar simulation validation results for other motif-network combinations can be found in the Appendix. In general, the simulation results confirm the theoretical results obtained in Theorem 1, 2, and 3 that our proposed estimators are asymptotically unbiased and consistent for different types of motifs and regardless of the topological structure of the underlying network  $\mathcal{G}$ . Thus, they can be easily applied to any complex network from diverse fields, including biological networks, social networks, the World-Wide-Web, engineering and electrical circuitries, etc.

Last but not least, we show the simulation validation results to confirm the convergence rates derived in Proposition 1 in chapter 2. Using each of the four random graph models, we generate networks with the number of nodes  $n$  up to 100,000 to obtain a bet-

ter evaluation of the convergence rates. Other parameters are fixed as  $\rho = 0.1$ ,  $p = 0.1$ , and the number of sampled subnetworks is 100. We then calculate the  $\text{Var}\left(\frac{\widehat{N}_1}{N_1}\right)$  and the dominated term  $\frac{N_2}{N_1^2}$  (see equation (2.11)). Fig. 3.3 and 3.4 show the results obtained for the preferential attachment model. As shown in Fig. 3.3, the convergence rate of  $\text{Var}\left(\frac{\widehat{N}_1}{N_1}\right)$  is very close to that of the term  $\frac{N_2}{N_1^2}$  as we have seen in equation (2.11). Moreover, Fig. 3.4 shows that the convergence rate of  $\frac{N_2}{N_1^2}$  is bounded by  $\frac{\log(n)}{n}$  as derived in Proposition 1. Similar results for the ER, geometric, and duplication models can be found in the Appendix. All simulation results confirm the accuracy of the convergence rates derived in Proposition 1 in chapter 2.

### 3.1.2 Simulation from real network data

Table 3.1: Number of nodes and links in the observed PPI subnetworks of *S. cerevisiae*, *C. elegans*, *H. sapiens*, and *A. thaliana*.

	<i>S. cerevisiae</i>	<i>C. elegans</i>	<i>H. sapiens</i>	<i>A. thaliana</i>
Number of proteins	1,278	9,906	7,194	8,429
Number of links	1,641	1,816	2,754	5,664

Most random graph models can only explain some, but not all characteristics of any complex network in the real-world. Thus, in addition to random networks, we also examine the performance of the estimators in “real” networks. To do so, we consider each observed subnetwork from real datasets as the entire complete network from which we draw smaller subnetworks, that is, sub-subnetworks, and do the estimation and validation. In particular, we use the observed PPI subnetworks of *S. cerevisiae*, *C. elegans*, *H. sapiens*, and *A. thaliana*. We obtained those subnetworks from the following four datasets which were produced from high-throughput Y2H experiments recently by the Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute: CCSB-HI1 [25, 27], CCSB-YI1 [28], CCSB-WI-2007 [29], and CCSB-AI1-Main [30]. The number

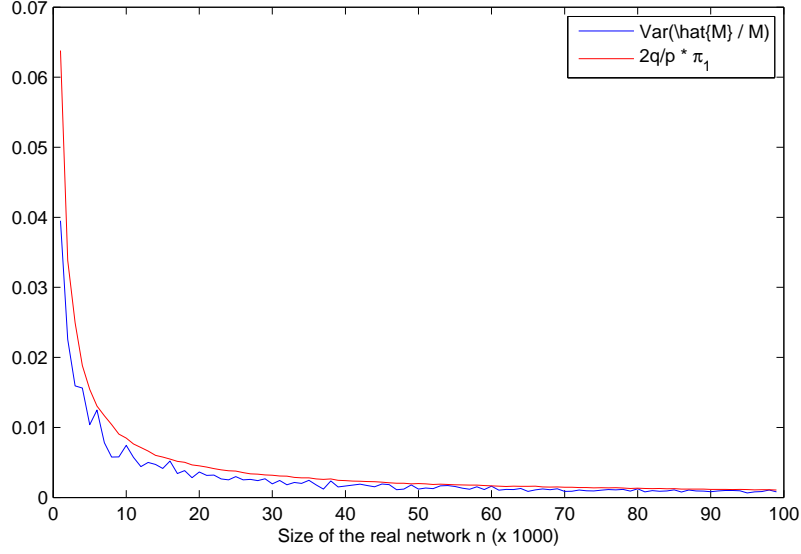


Figure 3.3: The convergence rate of  $\text{Var} \left( \frac{\hat{N}_1}{N_1} \right)$  in equation (2.11) and the dominated term  $\frac{N_2}{N_1^2}$  (denoted by  $\pi_1$  in the legend) for the preferential attachment model.

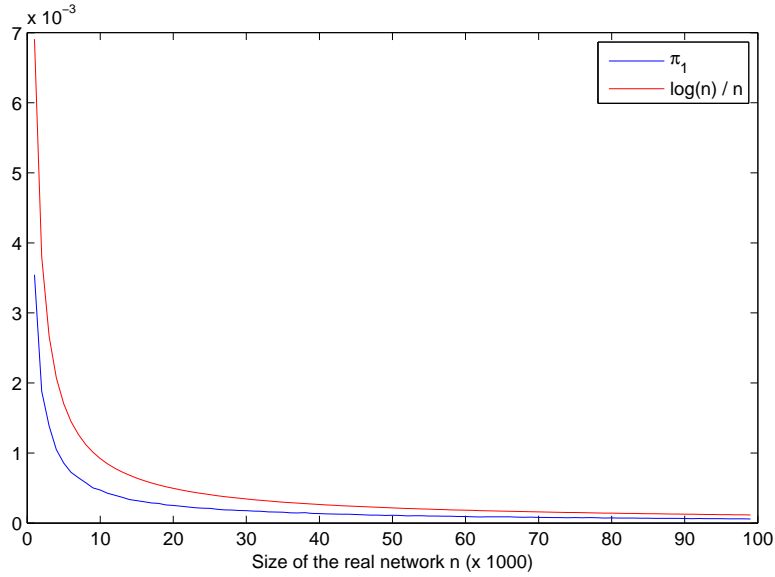


Figure 3.4: The convergence rate of  $\frac{N_2}{N_1^2}$  (denoted by  $\pi_1$  in the legend) is bounded by  $\frac{\log(n)}{n}$  as shown in Proposition 1 for the preferential attachment model.

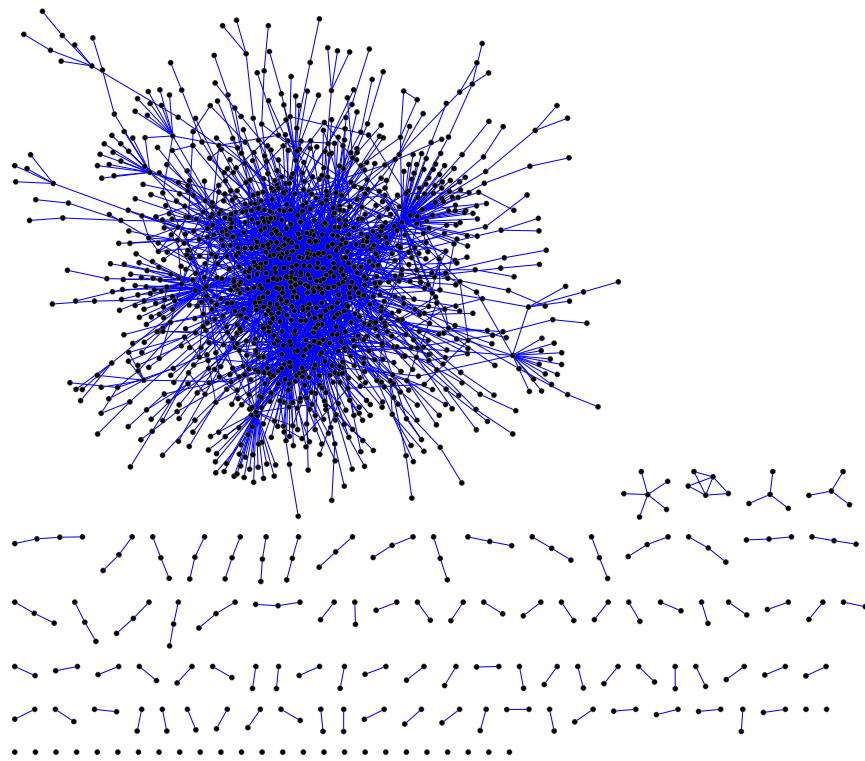


Figure 3.5: Observed PPI subnetwork of *H. sapiens* from Y2H experiment.

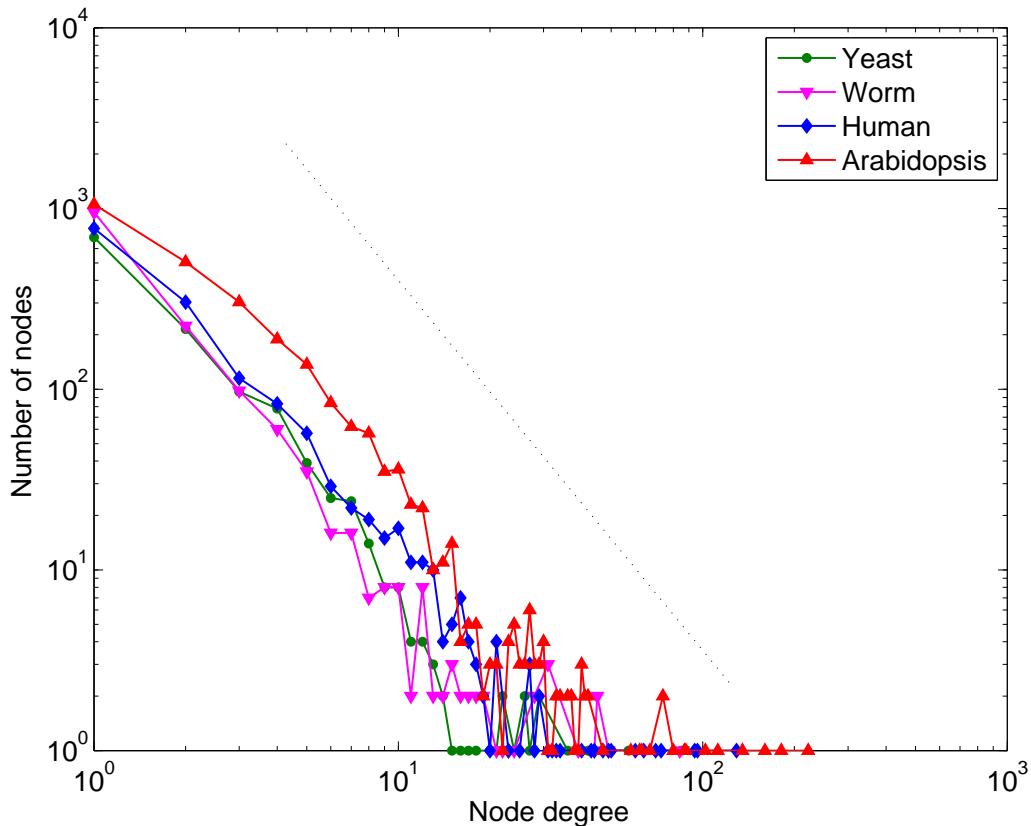


Figure 3.6: Degree distribution of four observed <sup>65</sup> PPI subnetworks of *S. cerevisiae*, *C. elegans*, *H. sapiens*, and *A. thaliana* (log-log scale). The linear pattern between the log of the number of nodes and the log of the node degree implies the scale-free structure of these subnetworks.

of proteins and interactions in those four subnetworks are presented in Table 3.1.

The PPI subnetwork of *H. sapiens* is shown in Fig. 3.5, the other three PPI subnetworks can be found in the Appendix. It can be seen from the figures that those subnetworks have more complicated structures than networks generated from random graph models. They are not symmetric, not hierarchical, etc. There is one large connected component and a lot of isolated islands, indicating a significant number of missing links between them. Moreover, those subnetworks just represent a part of the whole pictures since the number of proteins in each subnetwork is only about 1/6 to half of the entire proteomes. Interestingly, as shown in Fig. 3.6, the linear pattern between the log of the number of nodes and the log of the node degree suggests that those four complex subnetworks share similar scale-free structures, a prominent feature of biological networks as mentioned earlier in the chapter 1. More details of those datasets will be discussed in the next sections.

From each of those four networks, we repeat similar simulations as what have been done for random networks. Note that in this case, the number of nodes,  $n$ , and the link density,  $\rho$ , are fixed. Hence, we first study how the MSE of the estimators change with respect to the node sampling probability  $p$ . In particular, we try different values of the node sampling probability  $p = 0.05, 0.1, \dots, 0.5$ , while fixing the error rates as  $r_+ = 0.00001, r_- = 0.85$ .

Figure 3.7 shows the performance of the estimator  $\hat{N}_{\mathcal{M}}$  in the PPI network of *S. cerevisiae* for different undirected motifs. Despite the small number of nodes and the complex structure of the PPI network of *S. cerevisiae*, the estimator  $\hat{N}_{\mathcal{M}}$  still performs well, especially for small motifs and large values of the sampling probability  $p$ . In general, the MSE decreases as the sampling probability  $p$  increases. This is as expected, since the larger the sampling probability  $p$  is, the larger the sampled subnetworks are, and the more information they contain.

We also observe that larger motifs, that is, those with higher number of nodes and links, have worse estimation than the smaller ones. This is due to their low frequencies of appearance in the original network, making it difficult to estimate their number of occurrences from sampled subnetworks. In some cases, when the sampling probability is too low, we cannot get good estimation for large motifs. For example, as shown in Figure 3.7, the estimation for motifs  $u_4, u_5, u_6$  can be done only for sampling probability  $p \geq 0.1$ . Similarly, we need  $p \geq 0.15$  for motif  $u_7$ , and  $p \geq 0.2$  for motif  $u_8$ .

Figure 3.8 shows the performance of the estimator  $\tilde{N}_{\mathcal{M}}$ . As expected, its performance is not as good as the estimator  $\hat{N}_{\mathcal{M}}$  due to the noise. The bias-corrected estimation for motifs  $u_3, u_4, u_5$  can be done only for sampling probability  $p \geq 0.2$ , while we need  $p \geq 0.3$  for motif  $u_6$ . Other larger motifs cannot be estimated with sampling probability  $p \leq 0.5$ . Similar simulation results for the other three PPI networks of *C. elegans*, *H. sapiens*, and *A. thaliana* can be found in the Appendix.

We also study how the MSE of the bias corrected estimator  $\tilde{N}_{\mathcal{M}}$  depends on the false positive rate  $r_+$  and the false negative rate  $r_-$ . We fix the node sampling probability as  $p = 0.5$ , that is, 50% of the proteins in the PPI network of *S. cerevisiae* will be sampled. The error rates of Y2H assay are estimated as  $r_+ \simeq 10^{-5}$  and  $r_- \simeq 0.85$  (more details are given in the next sections). Thus, we try different values of the false positive rate  $r_+ = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$  and false negative rate  $r_- = 0.1, 0.2, \dots, 0.9$ .

Figures 3.9 and 3.10 show the MSE of the estimator  $\tilde{N}_{\mathcal{M}}$  for estimating the number of links and triangles. Similar results for other motif types can be found in the Appendix. In general, when the false positive rate  $r_+$  is fixed, the MSE is increased with respect to the false negative rate  $r_-$  as expected. There is no significant difference in the MSE for small values of the false positive rate  $r_+ = 10^{-5}, 10^{-4}, 10^{-3}$ . However, the MSE for  $r_+ = 10^{-2}$  is remarkably higher than that for the other three cases. We also notice that the MSE is higher for larger motifs, as previously observed.



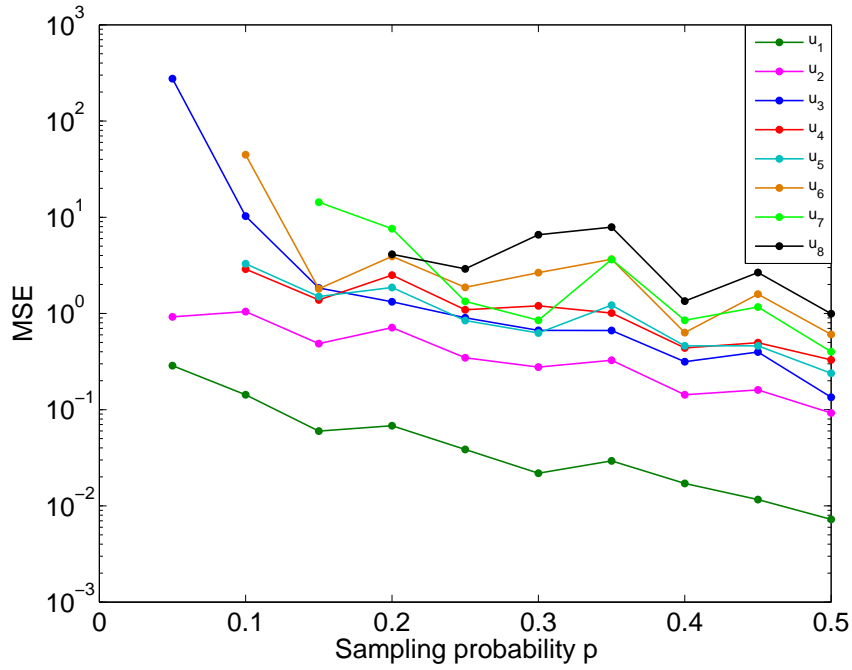


Figure 3.7: Performance of the estimator  $\hat{N}_M$  with respect to the node sampling probability  $p$  in the PPI network of *S. cerevisiae* for different undirected motifs. The motif IDs,  $u_1, u_2, \dots, u_8$  correspond to those in Figure 2.1.

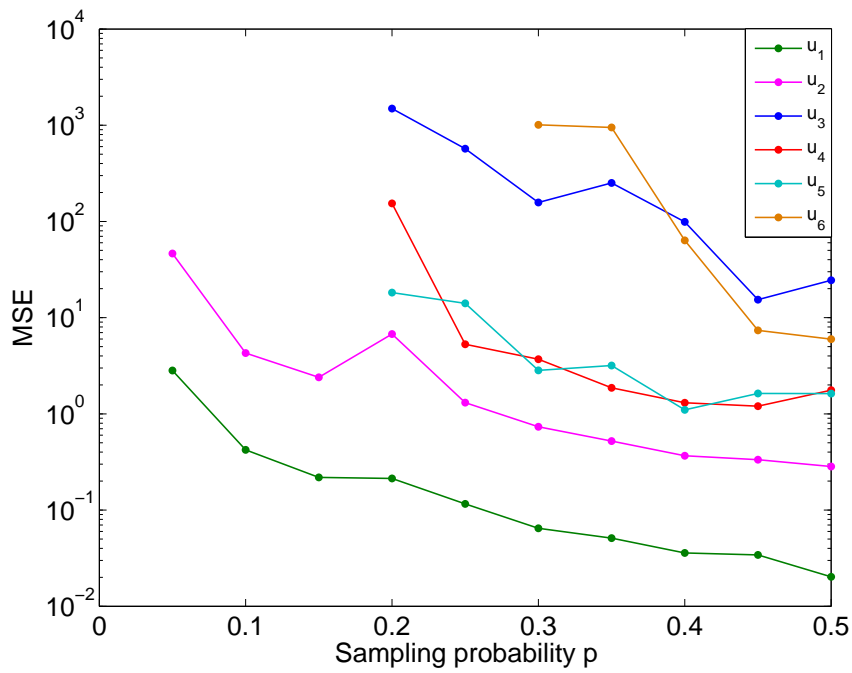


Figure 3.8: Performance of the estimator  $\tilde{N}_M$  with respect to the node sampling probability in the PPI network of *S. cerevisiae*.

In summary, our proposed estimators also show good performance in “real” PPI networks, despite their very complex structures. However, estimating the number of occurrences of large motifs requires moderate sampling probability, i.e., sufficient information from sampled subnetworks.

### 3.2 Computational time efficiency of the sampling-estimating approach

As a side result of the study, we found that even if the real network  $\mathcal{G}$  is fully known, and hence, one can directly count the real number of motif occurrences by exhaustive enumeration, sampling a sufficient number of small subnetworks and estimating the number of occurrences of a motif  $\mathcal{M}$  using the corresponding estimator  $\widehat{N}_{\mathcal{M}}$  is much more efficient in terms of computational time than exhaustive counting, while still can provide comparably accurate results. This is especially useful for huge networks for which exhaustively enumerating all motifs is impossible.

Take the triangle motif for instance. Consider a naive method, for example, the link iteration algorithm, which travels over all links of the network  $\mathcal{G}$  and counts the number of common neighbors of the two ends of each link. For counting the number of triangles in a network with  $n$  nodes and link density  $\rho$ , the algorithm will travel over  $\sim \binom{n}{2}\rho$  links, and for each link check  $(n - 2)$  nodes for common neighbors. Hence the total number of iterations is approximately  $(n - 2)\binom{n}{2}\rho$ . However, under the sampling-estimating approach with node sampling probability  $p$ , the number of nodes in a sampled subnetwork is  $\sim np$  nodes. Hence, the number of iterations can be reduced by a factor of  $\sim (p^3 \times \text{number of sampled subnetworks})$ , which may be less than one by suitably choosing small enough  $p$  and a sufficient number of sampled subnetworks.

We define the computational time efficiency as the ratio between the sampling-

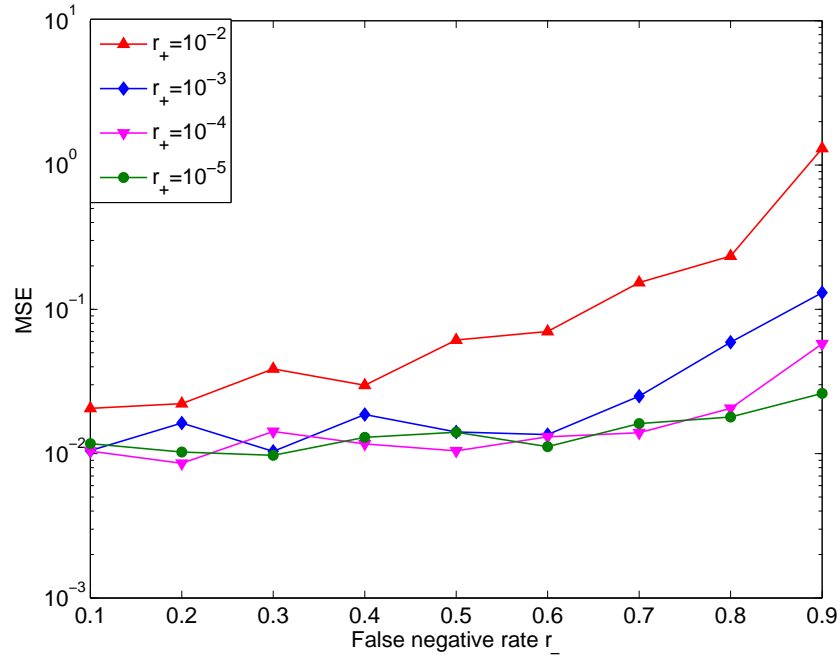


Figure 3.9: Performance of the estimator  $\tilde{N}_{\mathcal{M}}$  of the number of links with respect to false positive and false negative rates in the PPI network of *S. cerevisiae*.

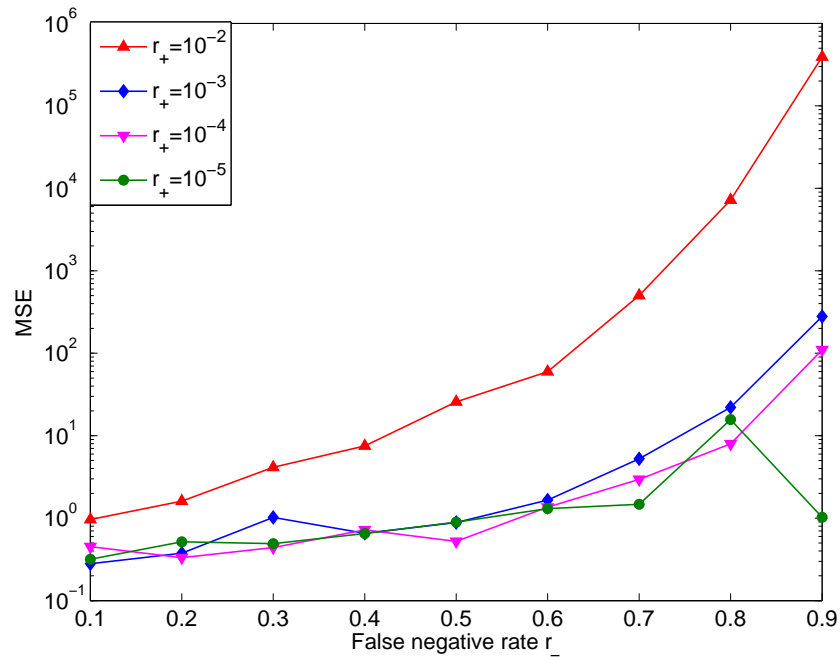


Figure 3.10: Performance of the estimator  $\tilde{N}_{\mathcal{M}}$  of the number of triangles with respect to false positive and false negative rates in the PPI network of *S. cerevisiae*.

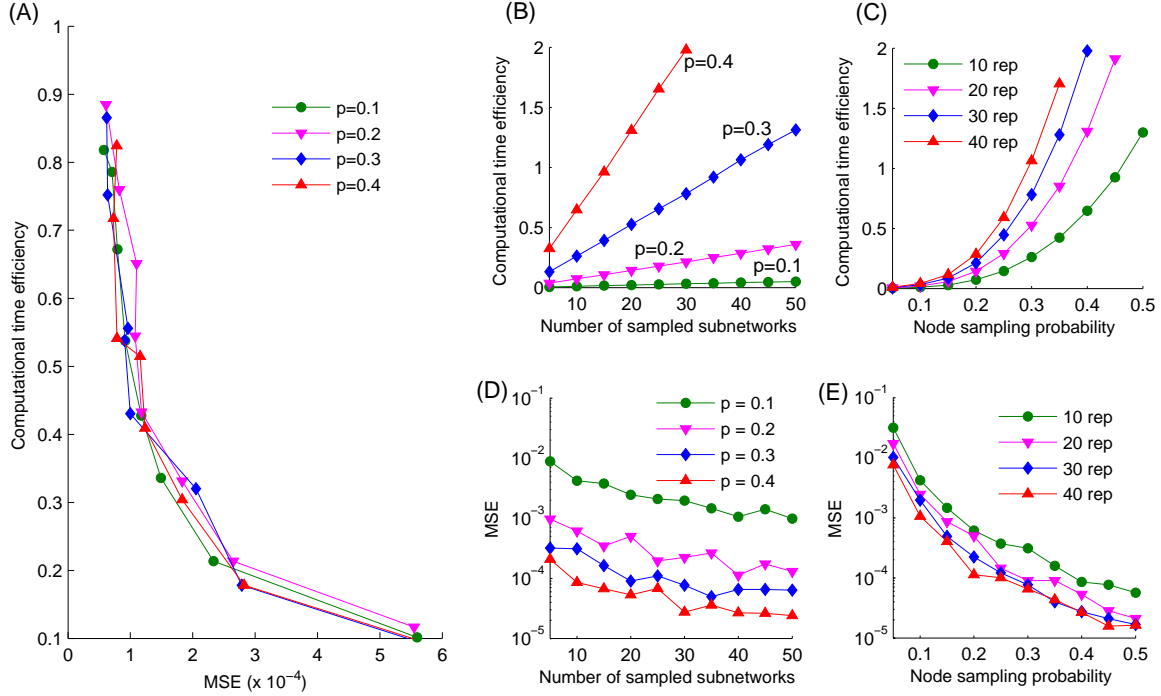


Figure 3.11: Computational time efficiency and MSE of the estimator  $\hat{N}_3$  for estimating the number of triangles in an example network of  $n = 5,000$  nodes and link density  $\rho = 0.1$ . **(A)** Computational time efficiency versus MSE for four values of the node sampling probability  $p$ . When  $p = 0.1, 0.2, 0.3, 0.4$ , the number of sampled subnetworks was set to  $125k, 25k, 5k, 2k$  ( $1 \leq k \leq 8$ ), respectively, so that the computational time efficiency is less than one. Using the sampling-estimation approach, we can achieve an  $\text{MSE} \sim 10^{-4}$ , that is, an estimate within  $\sim 1\%$  deviation from the real number, while using no more than  $\sim 50\%$  of the computational time of the exhaustive enumeration approach. **(B)** When  $p$  is fixed, the computational time efficiency increases as a linear function of the number of sampled subnetworks. **(C)** When the number of sampled subnetworks (that is, “**rep**”) is fixed, the computational time efficiency increases as a cubic function of  $p$ . **(D)** MSE decreases as the number of sampled subnetworks increases. **(E)** MSE decreases as  $p$  increases.

estimating time and the exhaustive enumeration time. Fig. 3.11 shows an example of how the computational time efficiency and the MSE of  $\widehat{N}_3$ , that is, our estimator for the number of triangles, depend on the node sampling probability  $p$  and the number of sampled subnetworks (or the number of repetitions “**rep**”). In this simulation, given a random network with  $n = 5,000$  nodes and the link density  $\rho = 0.1$ , we first count the number of triangles and record the counting time. Then for each value of the node sampling probability  $p = 0.05, 0.1, \dots, 0.5$ , we draw a sufficient number of subnetworks from each of which we calculate the estimator  $\widehat{N}_3$ . We take average of  $\widehat{N}_3$  over all sampled subnetworks as an estimate for the number of triangles in the original network. Finally, we calculate the sampling-estimating time and the MSE of the estimation.

As expected, panel (B) of Fig. 3.11 shows that the computational time efficiency increases as a linear function of the number of sampled subnetworks when the node sampling probability  $p$  is fixed. Similarly, panel (C) of Fig. 3.11 shows that the computational time efficiency increases as a cubic function of the node sampling probability  $p$  when the number of sampled subnetworks, “**rep**”, is fixed. Panels (D) and (E) show that the MSE decreases with respect to the number of sampled subnetworks and the node sampling probability  $p$ . Finally, in panel (A), we show all possible combinations of the number of sampled subnetworks and the node sampling probability  $p$  for which the computational time efficiency is less than one, that is, when the sampling-estimating approach is more efficient than the exhaustive enumeration approach. Using the sampling-estimation approach, we can achieve an  $\text{MSE} \sim 10^{-4}$ , that is, an estimate within  $\sim 1\%$  deviation from the real number, while using no more than  $\sim 50\%$  of the computational time of the exhaustive enumeration approach.

Most importantly, this sampling-estimating approach is a friend rather than a competitor to other counting algorithms. Any method which is able to count the number of motif occurrences in a network can also be applied to its subnetworks in combination

with this sampling-estimating approach. This is especially useful for huge networks for which exhaustively enumerating all motifs is impossible.

### 3.3 Estimating motif counts in PPI networks

Together with the theoretical results obtained in chapter 2, the simulation validation results in the previous sections have confirmed the accuracy of our proposed estimators. In this section we apply the biased-corrected estimator  $\tilde{N}_1$  and  $\tilde{N}_3$  (see Table 2.3) to estimate the number of links and triangles in the PPI networks of four species: *S. cerevisiae* (Yeast), *C. elegans* (Worm), *H. sapiens* (Human), and *A. thaliana* (Arabidopsis).

Table 3.2: Observed PPI subnetworks of *S. cerevisiae*, *C. elegans*, *H. sapiens*, & *A. thaliana*, and their quality parameters.

	<i>S. cerevisiae</i>	<i>C. elegans</i>	<i>H. sapiens</i>	<i>A. thaliana</i>
Total no. of proteins	6,000	20,065	22,500	27,029
No. of proteins screened	3,676	9,906	7,194	8,429
No. of links detected	1,130	1,816	2,754	5,664
<b>Quality parameters</b>				
Precision	0.9400	0.8600	0.7940	0.8030
False discovery rate	0.0600	0.1400	0.2060	0.1970
Sensitivity	0.1700	0.0496	0.0950	0.1570
False negative rate ( $r_-$ )	0.8300	0.9504	0.9050	0.8430
False positive rate ( $r_+$ )	$0.8 \times 10^{-5}$	$0.5 \times 10^{-5}$	$2 \times 10^{-5}$	$3 \times 10^{-5}$

We obtain their PPI subnetworks from the following four datasets which have been published recently by the Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute: CCSB-HI1 [25, 27], CCSB-YI1 [28], CCSB-WI-2007 [29], and CCSB-AI1-Main [30]. These datasets were produced from high-throughput Y2H experiments and their quality parameters were accurately estimated by the authors from CCSB. Table 3.2 summarizes the information of these datasets and their quality parameters.

Take Yeast for instance. It is estimated that the total number of proteins in Yeast

is about 6,000. In this Y2H experiment, only 3,676 proteins available in hands were tested. All of them were used both as bait and prey, so the test space is a  $3,676 \times 3,676$  square matrix. There are 1,130 interactions detected between these 3,676 proteins. To examine the accuracy of Y2H experiments and the quality of detected interactions, the traditional approach is comparing the observed data with some gold-standard datasets [67, 68]. However, as mention in chapter 1, all biological network data are incomplete and so are the gold-standard datasets. Hence, it is not reliable to use the traditional approach to assess the quality of the observed data. To overcome these limitations, in [27] the authors from CCSB have proposed an empirical framework which uses multiple cross-assay experiments to validate the detected interactions, and then, accurately estimate the quality parameters of the observed data, including precision, false discovery rate, sensitivity, false negative and false positive rates.

Given the observed PPI subnetworks and their link error rates, we are now ready to apply the bias-corrected estimators  $\tilde{N}_1$  and  $\tilde{N}_3$  to estimate the number of links and triangles in the entire PPI networks.

### 3.3.1 Comparison of our estimator $\tilde{N}_1$ and CCSB estimator

$$\tilde{N}^{\text{CCSB}}$$

We first re-estimate the size, i.e. the number of links, of these four interactomes using the bias-corrected estimator  $\tilde{N}_1$ . As mentioned in chapter 1, this task is of critical importance in Network Biology because the number of interactions together with the number of genes/proteins may reflect the biological complexity of living organisms. Moreover, the estimates also show us how complete the current biological network data are and how much work still to be done. There are two approaches for this task of estimation. The first approach is to model the intersection of two observed PPI subnetworks of the same species using the hypergeometric distribution [70, 72].

The second approach, which is applied in our work and [27, 28, 29, 30, 73, 74], is to extrapolate the number of links in the observed subnetwork to estimate that in the entire network.

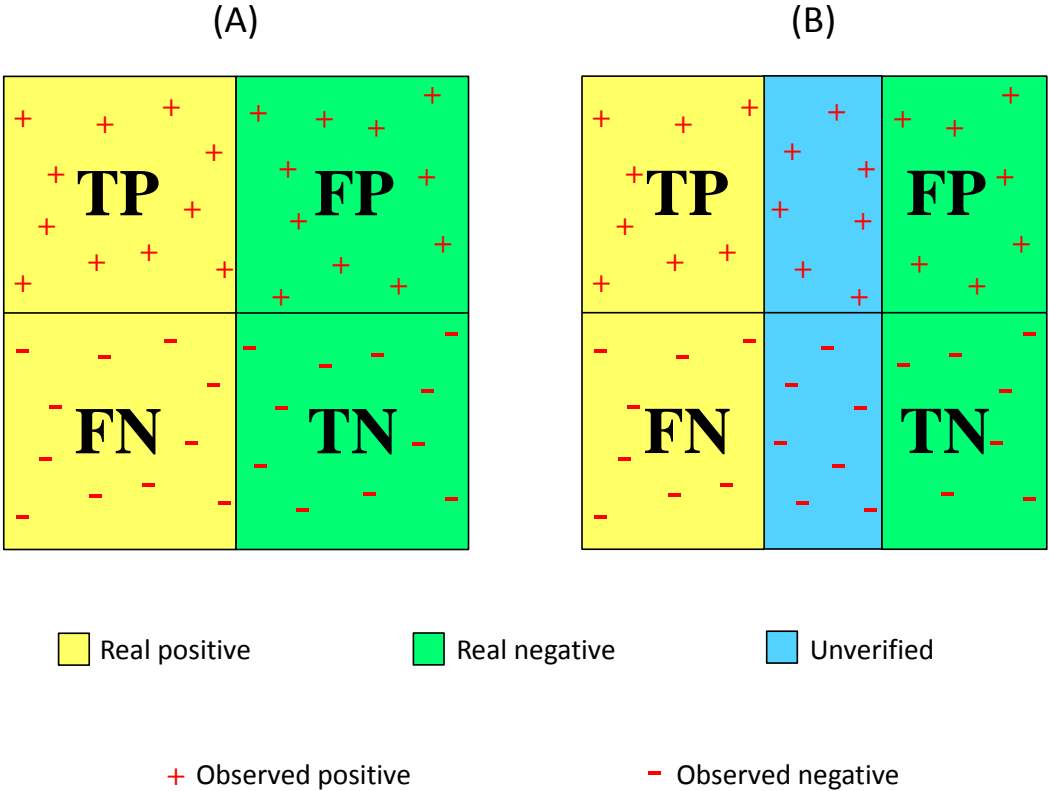


Figure 3.12: Limitation of gold-standard datasets. (A) Ideal classification: gold-standard positive and negative sets are completely known. Hence, we can verify if a detected interaction is a real one (TP) or a spurious link (FP), and if an undetected interaction is a missing link (FN) or it really does not exist in the real network (TN). (B) Gold-standard sets for biological networks data are currently limited and biased. As a result, many detected as well as undetected interactions cannot be verified.

In [73] and CCSB papers [27, 28, 29, 30], the size of an interactome was estimated from observed subnetwork data using the same “extrapolation” approach, but in a slightly different way from ours. In particular, in both studies the number of links in the observed subnetwork  $\mathcal{G}^{\text{obs}}$  is first scaled up by the factor  $\frac{\binom{n}{2}}{\binom{n^{\text{obs}}}{2}}$  (which is referred



to as  $\frac{1}{\text{completeness}}$  in [27, 28, 29, 30]) to obtain the estimator  $\widehat{N}_1$ . The bias caused by the missing and spurious links, however, is handled in different ways: we make use of the false positive and false negative rates, whereas they consider the precision and sensitivity.

Let  $\widetilde{N}^{\text{CCSB}}$  denote the CCSB estimator. We have:

$$\begin{aligned}\widetilde{N}_1 &= \frac{\widehat{N}_1 - \binom{n}{2}r_+}{1 - r_+ - r_-}, \\ \widetilde{N}^{\text{CCSB}} &= \frac{\widehat{N}_1 \times \text{precision}}{\text{sensitivity}}.\end{aligned}\tag{3.2}$$

Recall that the quality parameters are defined as follows:

- TP: true positives; FP: false positives; FN: false negatives; TN: true negatives;
- precision =  $\frac{\text{TP}}{\text{TP}+\text{FP}}$ ;
- false discovery rate  $r_d = \frac{\text{FP}}{\text{TP}+\text{FP}} = 1 - \text{precision}$ ;
- sensitivity =  $\frac{\text{TP}}{\text{TP}+\text{FN}}$ ;
- false negative rate  $r_- = \frac{\text{FN}}{\text{TP}+\text{FN}} = 1 - \text{sensitivity}$ ;
- false positive rate  $r_+ = \frac{\text{FP}}{\text{FP}+\text{TN}}$ .

In an ideal classification problem (panel (A), Fig. 3.12), the gold-standard positive and negative sets are completely known. Hence, we can verify if a detected interaction is a real one (TP) or a spurious link (FP), and if an undetected interaction is a missing link (FN) or it really does not exist in the real network (TN). In this ideal classification,

we have the following equalities:

$$\widehat{N}_1 = \text{TP} + \text{FP}, \quad (3.3)$$

$$N_1 = \text{TP} + \text{FN}, \quad (3.4)$$

$$\binom{n}{2} - N_1 = \text{FP} + \text{TN}. \quad (3.5)$$

Since

$$N_1 = \text{TP} + \text{FN} = \widehat{N}_1 \times \text{precision} + N_1 \times (1 - \text{sensitivity}), \quad (3.6)$$

this gives rise to the CCSB estimator  $\widetilde{N}^{\text{CCSB}}$ .

On the other hand, we have

$$\widehat{N}_1 = \text{TP} + \text{FP} = N_1(1 - r_-) + \left( \binom{n}{2} - N_1 \right) r_+, \quad (3.7)$$

and this gives rise to our estimator  $\widetilde{N}_1$ .

Thus, in the ideal classification problem, our estimator  $\widetilde{N}_1$  and the CCSB estimator  $\widetilde{N}^{\text{CCSB}}$  are mathematically equivalent.

However, for biological networks data, gold-standard positive and negative sets are currently limited and biased (panel (B) Fig. 3.12). As a result, many detected as well as undetected interactions cannot be verified. For example, even if a detected interaction has not been reported previously in any gold-standard positive set, it is not necessary a spurious link (FP) because it can also represent a novel interaction as well. Thus, using gold-standard sets to infer the quality parameters is not a reliable approach. In particular, equations in (3.3, 3.4, 3.5) no longer hold, and neither is the equivalent relationship between the two estimators  $\widetilde{N}_1$  and  $\widetilde{N}^{\text{CCSB}}$ .

The empirical framework using multiple cross-assay validations proposed by CCSB in [27] offers more accurate estimates of the quality parameters than traditional com-

parison approaches [27]. In such cases, it can be shown that our estimator  $\tilde{N}_1$  and the CCSB estimator  $\tilde{N}^{\text{CCSB}}$  are still almost the same. In particular, using the error rates  $r_-$ ,  $r_+$ , and  $r_d$ , which are estimated from empirical experiments, the two estimators can be rewritten as

$$\tilde{N}_1 = \frac{1}{1 - r_+ - r_-} \left( \frac{\binom{n}{2}}{\binom{n^{\text{obs}}}{2}} N_1^{\text{obs}} - \binom{n}{2} r_+ \right), \quad (3.8)$$

$$\tilde{N}^{\text{CCSB}} = \frac{1}{1 - r_-} \frac{\binom{n}{2}}{\binom{n^{\text{obs}}}{2}} N_1^{\text{obs}} (1 - r_d). \quad (3.9)$$

Since  $r_+$  is much smaller than  $r_-$ , we have

$$1 - r_+ - r_- \simeq 1 - r_-. \quad (3.10)$$

Then

$$\tilde{N}_1 - \tilde{N}^{\text{CCSB}} \simeq \frac{\binom{n}{2}}{1 - r_-} \left( \frac{N_1^{\text{obs}} r_d}{\binom{n^{\text{obs}}}{2}} - r_+ \right) = \frac{\binom{n}{2}}{1 - r_-} \left( \frac{\text{FP}^{\text{obs}}}{\binom{n^{\text{obs}}}{2}} - \frac{\text{FP}^{\text{obs}}}{\binom{n^{\text{obs}}}{2} - \text{P}^{\text{obs}}} \right), \quad (3.11)$$

where  $\text{FP}^{\text{obs}}$  and  $\text{P}^{\text{obs}}$  respectively denote the number of false positive links and real positive links in the observed subnetwork  $\mathcal{G}^{\text{obs}}$ . It should be noted that biological networks are quite sparse, and so,  $\text{P}^{\text{obs}} \ll \binom{n^{\text{obs}}}{2}$ . Hence, our estimator  $\tilde{N}_1$  and the CCSB estimator  $\tilde{N}^{\text{CCSB}}$  are almost the same.

However using multiple cross-assay experiments to validate detected interactions and to accurately estimate the quality parameters is not always possible due to time and financial constraints. As a result, one must use gold-standard sets to make inference although they are currently limited and biased (Fig. 3.12). In such cases, we believe that our estimator is better than the CCSB estimator because estimating the false positive and false negative rates is more reliable than estimating the precision.

### 3.3.2 Estimating the number of links in PPI networks

In Table 3.3, we present our estimates of the interactome size of four species together with the CCSB estimates. As our method only provides a point estimate, but not the variation of the estimator, we need to use an empirical approach to quantify the uncertainty in the estimation. However, there is only one observed subnetwork data set available for each species. Hence, we use the re-sampling strategy to assess the variation of the estimator. In particular, we sample 100 smaller subnetworks from each observed PPI subnetwork, that is, sub-subnetworks, by using the same uniformly random node sampling process. Each sub-subnetwork is used to estimate the interactome size, and those estimates are then used to estimate the standard deviation of the estimator. We set the node sampling probability to be small,  $p = 0.1$ , as we think that it is the worst case and it gives the upper bound for the standard deviation. Larger values of  $p$  give smaller estimates of the standard deviation. As shown in Table 3.3, the estimates calculated from sampled sub-subnetworks agree well with the estimates calculated from the observed PPI subnetworks, indicating that our estimator  $\tilde{N}_1$  is consistent. Moreover, our estimates agree well with those of CSSB for three species Yeast, Worm, and Arabidopsis, demonstrating again the accuracy of our estimator. For Human, our estimate is  $\sim 50\%$  higher than CCSB estimate, suggesting that the interactome of Human is larger than previously expected.

Since the total number of proteins of these four species are quite different, it is reasonable to compare the link density rather than the total number of interactions of these four PPI networks. The link density is defined as the ratio between the number of interactions and the number of all possible pairs of proteins. Interestingly, as shown in Table 3.3, the estimated link density of these four species are surprisingly quite similar, around  $6-9 \times 10^{-4}$ . This observation has also been reported previously in [30]. We notice that Yeast has the highest link density, but this may be due to the fact that

Yeast is the most well-studied model organism. More importantly, the proteome size and interactome size of Human are not as large as previously expected. For instance, Arabidopsis has even larger proteome and interactome than Human does. Thus, it seems that the number of proteins and their interactions simply cannot explain the nature of biological diversity of living organisms, and we need to look at more complex and informative features, in particular, network motifs.

### 3.3.3 Estimating the number of triangles in PPI networks

Table 3.3: The interactome size and the number of triangles in the PPI networks of *S. cerevisiae*, *C. elegans*, *H. sapiens*, and *A. thaliana*, estimated based on recently published datasets from the Center for Cancer Systems Biology (CCSB).

	<i>S. cerevisiae</i>	<i>C. elegans</i>	<i>H. sapiens</i>	<i>A. thaliana</i>
Total no. of proteins	6,000	20,065	22,500	27,029
No. of proteins screened	3,676	9,906	7,194	8,429
No. of links detected	1,130	1,816	2,754	5,664
<b>Quality parameters</b>				
Precision	0.9400	0.8600	0.7940	0.8030
False discovery rate	0.0600	0.1400	0.2060	0.1970
Sensitivity	0.1700	0.0496	0.0950	0.1570
False negative rate $r_-$	0.8300	0.9504	0.9050	0.8430
False positive rate $r_+$	$0.8 \times 10^{-5}$	$0.5 \times 10^{-5}$	$2 \times 10^{-5}$	$3 \times 10^{-5}$
<b>Interactome size</b>				
CCSB estimate	$18,000 \pm 4,500$	$116,000 \pm 26,400$	$130,000 \pm 32,600$	$299,000 \pm 79,200$
Our estimate <sup>a</sup>	17,000	121,000	210,000	289,000
Mean $\pm$ SD <sup>b</sup>	$18,000 \pm 2,800$	$122,000 \pm 16,600$	$214,000 \pm 32,200$	$295,000 \pm 33,400$
Link density	$9 \times 10^{-4}$	$6 \times 10^{-4}$	$8 \times 10^{-4}$	$8 \times 10^{-4}$
<b>No. of triangles</b>				
Our estimate <sup>a</sup>	82,000	6,263,000	10,270,000	7,381,000
Mean $\pm$ SD <sup>b</sup>	$75,000 \pm 38,400$	$5,971,000 \pm 3,593,800$	$11,255,000 \pm 4,717,100$	$7,720,000 \pm 3,132,700$
Triangle density	$2 \times 10^{-6}$	$5 \times 10^{-6}$	$5 \times 10^{-6}$	$2 \times 10^{-6}$

<sup>a</sup> estimates calculated from observed PPI subnetworks

<sup>b</sup> mean and standard deviation (SD) of the estimates calculated by sampling 100 sub-subnetworks from each observed PPI subnetwork using the node sampling probability  $p = 0.1$

We proceed further to estimate the number of triangles in these four interactomes using the corresponding bias-corrected estimator  $\tilde{N}_3$ . For each interactome, we estimate the number of triangles directly from the observed subnetwork data and from the 100 sampled sub-subnetworks. Again, the two estimates agree well for all species, indicating that the bias-corrected estimator  $\tilde{N}_3$  is consistent (Table 3.3). We noted that the

estimates of the standard deviation are relatively high, approximately half of the mean estimates. However, recall that we set the node sampling probability to be relatively small,  $p = 0.1$ , which we think is the worst case, hence those estimates can be considered as upper bounds of the standard deviation. Larger values of  $p$  should give smaller estimates of the standard deviation. Similar to the link density, we also look at the triangle density, which is defined as the ratio between the estimated number of triangles and the number of all possible combinations of three proteins chosen from the entire proteome. Surprisingly, we found that the triangle density of these four interactomes are also in the same range  $2-5 \times 10^{-6}$ . However, this time we notice that triangle density of the Human and Worm interactomes are 2.5 times as large as that of the interactome of Arabidopsis and Yeast, whereas as reported in the previous section, the link density of the later are higher than that of the formers. This may indicate a highly clustering and well-connected structure of the PPI network of Human and Worm.

### **3.3.4 Gene Ontology (GO) analysis of triangles in the observed PPI subnetwork of Yeast**

To further explore the biological meaning of triangles in PPI networks, we analyze the Gene Ontology (GO) annotations of triangles in the observed PPI subnetwork of Yeast. GO annotations of Yeast were downloaded from the Gene Ontology website (July, 2012). There are totally 60,982 unique annotations, corresponding to 6,383 genes and 4,705 GO terms. In the observed PPI subnetwork of Yeast, we found 112 triangles formed by 155 proteins, which have been annotated to 616 GO terms via 1,671 annotations. Among these 112 triangles, we found 61 triangles in which all three proteins share at least one GO term. There are 53 GO terms that have been assigned to all three nodes of at least one triangle.

For each of these 53 GO terms:

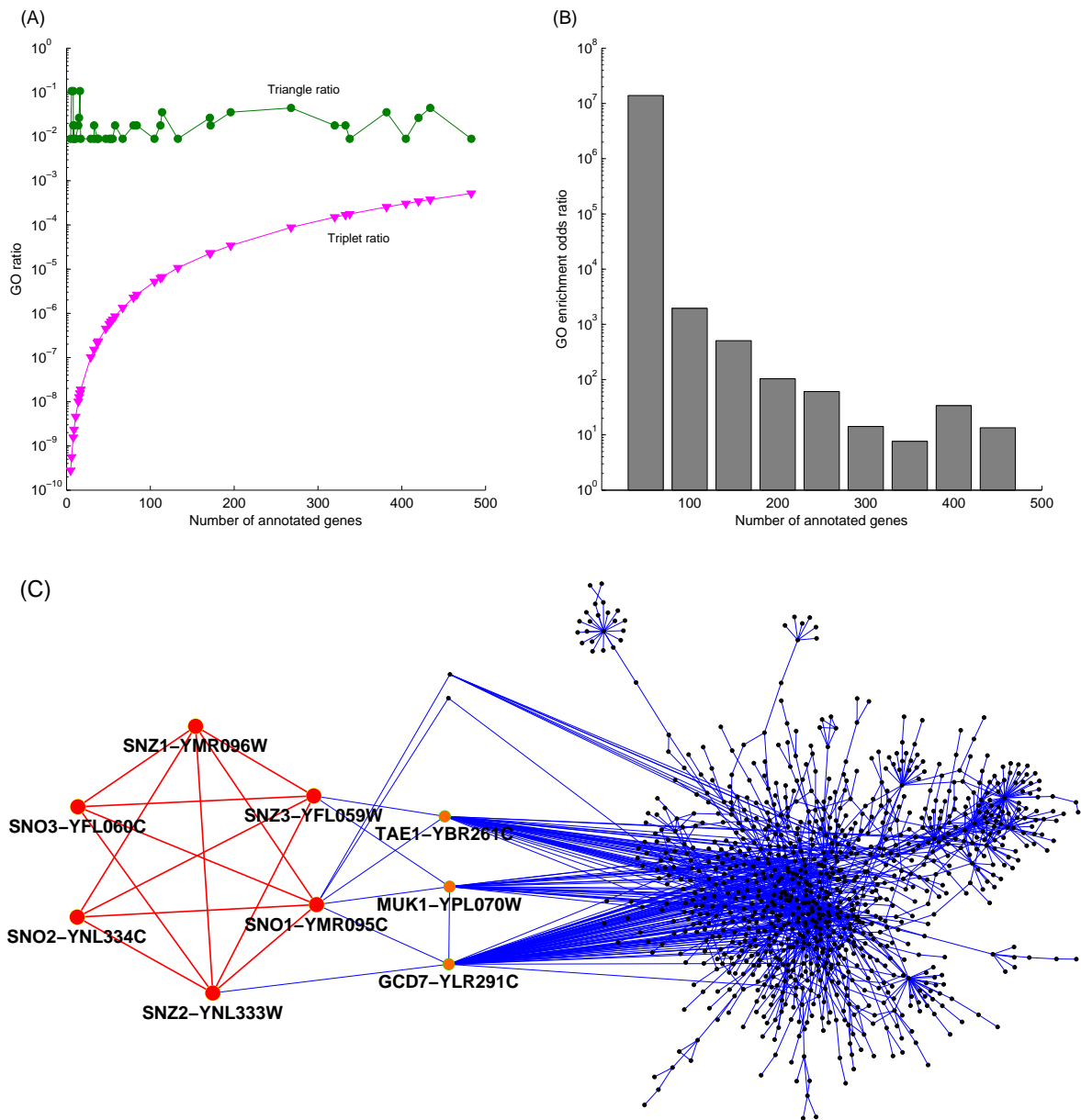


Figure 3.13: The enrichment in shared GO annotations of triangles in the observed PPI subnetwork of *S. cerevisiae*. **(A)** The triangle ratio and the triplet ratio of 43 GO terms that have less than 500 annotated genes and have been assigned to all three nodes of at least one triangle. **(B)** The odds ratio between the triangle ratio and the triplet ratio. The GO terms are grouped according to their number of annotated genes. **(C)** Three GO terms GO:0008614 (pyridoxine metabolic process, 6 annotated proteins), GO:0008615 (pyridoxine biosynthetic process, 8 annotated proteins), and GO:0009228 (thiamine biosynthetic process, 16 annotated proteins) are assigned to 12 triangles of a complex formed by 6 proteins SNZ1-YMR096W, SNZ2-YNL333W, SNZ3-YFL059W, SNO1-YMR095C, SNO2-YNL334C, and SNO3-YFL060C. These proteins are connected to the remaining part of the subnetwork via three hubs: TAE1-YBR261C, MUK1-YPL070W, and GCD7-YLR291C, where GCD7-YLR291C is the most highly connected protein in the observed PPI subnetwork of *S. cerevisiae*.

- We count the number of triangles in which all three proteins are annotated to that GO term:  $X_{GO}$ .
- We count the number of proteins in the whole genome of Yeast that are annotated to that GO term:  $n_{GO}$ .
- We calculate the triplet ratio,  $\frac{\binom{n_{GO}}{3}}{\binom{n}{3}}$ , where  $n = 6,000$  is the approximate genome size of Yeast.
- We calculate the triangle ratio:  $\frac{X_{GO}}{112}$ .
- We calculate the odds ratio between the triangle ratio and the triplet ratio.

The triplet ratio represents the probability that any three randomly chosen proteins of Yeast are all annotated to that particular GO. The triangle ratio represents the probability that all three proteins of a randomly chosen triangle in the observed PPI subnetwork are annotated to that particular GO. Thus, the odds ratio reflects the enrichment of triangles in shared GO annotations, that is, how more likely three proteins will share a particular GO if they come from a triangle rather than being chosen randomly.

We use the number of annotated proteins,  $n_{GO}$ , as a measure of the specificity of a GO term: the smaller  $n_{GO}$  is, the more specific the GO term is. We consider only those GO terms with  $n_{GO} \leq 500$  (43 out of 53). In Fig. 3.13, panel (A) shows the triangle ratio and the triplet ratio, and panel (B) shows the odds ratio of those 43 GO terms with respect to their number of annotated proteins. It can be seen that the triangle ratio is orders of magnitude larger than the triplet ratio, suggesting that the triangles in the observed PPI subnetwork of Yeast are enriched in shared GO annotations. Moreover, panel (B) shows that the odds ratio decreases with respect to the number of annotated proteins, indicating that the enrichment is more significant for more specific GO terms.



Panel (C) of Fig 3.13 shows an example of the three GO terms GO:0008614 (pyridoxine metabolic process, 6 annotated proteins), GO:0008615 (pyridoxine biosynthetic process, 8 annotated proteins), and GO:0009228 (thiamine biosynthetic process, 16 annotated proteins), which are assigned to 12 triangles of a complex formed by 6 proteins SNZ1-YMR096W, SNZ2-YNL333W, SNZ3-YFL059W, SNO1-YMR095C, SNO2-YNL334C, and SNO3-YFL060C. These proteins are connected to the remaining part of the subnetwork via three hubs: TAE1-YBR261C, MUK1-YPL070W, and GCD7-YLR291C, where GCD7-YLR291C is the most highly connected protein in the observed PPI subnetwork of Yeast.

We also calculated the  $p$ -value for each GO term using the hypergeometric distribution with parameters:

- population size:  $\binom{n}{3}$ ; total number of successes:  $\binom{n_{GO}}{3}$ ;
- number of draws: 112; number of successes in draws:  $X_{GO}$ .

and found that for 52 out of 53 GO terms, the  $p$ -value is less than 0.01, indicating again that the triangles found in the observed PPI subnetwork of Yeast are significantly enriched in shared GO annotations.

Details of the 53 GO terms and their odds ratio,  $p$ -value can be found in the Appendix.

### 3.4 Estimating motif counts in gene regulatory networks

In the previous section, we have demonstrated how to apply our proposed method to real PPI networks, which are undirected examples. In this section, we shall apply the method to the directed case with transcription factor (TF) regulatory networks.

Recently, the TF regulatory network of forty-one Human cell and tissue types were obtained from genome-wide in vivo DNaseI footprints map [38]. In these networks, the nodes are 475 TFs and the regulations of one TF by another are represented by network directed links. There are totally 38,393 unique regulatory interactions detected, with an average of 11,193 links per cell type.

### 3.4.1 Significant enrichment of motifs

Table 3.4: The estimated network size and the estimated counts of triad and quadriad motifs (in thousands).

	No. of links	No. of feedback loop	No. of FFL	No. of bi-parallel	No. of bi-fan
<b>Epithelia</b>	412 ± 68	2,836 ± 1,106	30,629 ± 10,512	3,261,403 ± 1,506,055	5,680,572 ± 2,363,616
<b>Stroma</b>	412 ± 38	2,727 ± 705	29,155 ± 6,290	3,052,803 ± 883,160	5,094,576 ± 1,337,401
<b>Blood</b>	434 ± 97	3,687 ± 1,699	37,884 ± 15,241	4,379,527 ± 2,320,472	7,359,970 ± 3,421,025
<b>Endothelia</b>	447 ± 40	3,160 ± 695	35,314 ± 6,567	3,844,161 ± 948,207	6,877,606 ± 1,540,212
<b>Cancer</b>	380 ± 7	2,378 ± 111	30,122 ± 710	2,862,215 ± 91,628	6,267,987 ± 99,444
<b>Fetal Cells</b>	426 ± 70	3,088 ± 998	33,782 ± 9,955	3,660,840 ± 1,500,838	6,498,027 ± 2,284,014
<b>ES Cell<sup>a</sup></b>	485	2,766	32,400	3,282,473	6,436,708

<sup>a</sup> There is only one embryonic stem (ES) cell TF network.

Given that Human has  $\sim 2886$  TFs [75], we want to estimate the number of occurrences of five motifs for each of seven functionally related classes of cells. The five motifs include link, feedback loop, feed-forward loop, bi-fan, and bi-parallel (Fig. 2.2). As mentioned in chapter 1, these motifs have been highlighted in previous studies [59] as functional units or building blocks of real-world complex networks. Forty-one Human cell and tissue types are classified into seven functionally related classes: Epithelia, Stroma, Blood, Endothelia, Cancer, Fetal, and Embryonic Stem cells [38]. Since there is no information about the quality assessment of this dataset, we simply set the false positive and negative rates to zero.

Table 3.4 shows our estimates of the five motifs counts in each of the seven functional classes. For each class, the mean and standard deviation of the motif counts are reported. We found that the feed-forward loop motif is significantly enriched in these

forty-one Human cell-specific TF regulatory networks. In particular, the number of occurrences of the feed-forward loop is an order of magnitude larger than that of the feedback loop, whereas in a random network this ratio is approximately 3:1. Table 3.4 also suggests that bi-fan is relatively abundant in these forty-one networks as the ratio of the bi-fan count to the bi-parallel count is 2:1, whereas in a random network this ratio should be 1:2.

To accurately assess the enrichment of motifs in these forty-one Human cell-specific TF regulatory networks, we need to compare the motif counts in the observed networks with those in random networks. It is important to ensure that the random networks must have similar topological structures as the observed ones. Hence, we consider the following link-rewiring process:

- At each iteration, a pair of links “a-b” and “c-d” in the observed network are randomly chosen.
- Rewire “a-b” and “c-d” to “a-d” and “c-b”
- From each of forty-one observed networks, 100k iterations are done to generate one randomly rewired network.
- 50 randomly rewired networks are generated for each of forty-one observed networks.

After the rewiring process, the out-degree & in-degree distributions remain unchanged, as well as the number of nodes and the number of links.

For each of the forty-one Human cell-specific TF regulatory networks, we calculate the mean and the standard deviation of the motif counts from the corresponding 50 randomly required networks, and compare them with the observed ones. Fig. 3.14 and 3.15 show the simulation results for the feedback loop and feed-forward loop. It

can be seen clearly from the figures that the feed-forward loop is significantly enriched in forty-one observed networks, since the observed count is more than  $3\times$  standard deviation higher than the mean count from 50 randomly rewired networks. In contrast, the feedback loop is not enriched as its observed count is within  $3\times$  standard deviation from the mean count. Similar results have been reported previously in [38]

### 3.4.2 Linear correlation of motif counts

Beside the enrichment of some functional motifs, we also found another surprising feature of the motif counts in these forty-one Human cell-specific TF regulatory networks. In particular, there is a very strong linear correlation between the counts of triad and quadriad motifs, although these motifs are topologically very different and the TF networks are from very different cell types (Fig. 3.16). The linear patterns can be seen clearly from the scatter plots in the upper diagonal panels of Fig. 3.16. The lower diagonal panels also show remarkably high correlation coefficients of the motif counts in these forty-one networks. This linear correlation is an interesting phenomenon, and further analysis is needed to explore its biological meaning. More details of the linear correlation of motif counts will be given in the next chapter.

We notice that the TF regulatory networks of blood cells (red dots) have very different motif counts. Specifically, for all triad and quadriad motifs, the promyelocytic leukemia cell TF network has the largest number of occurrences whereas the erythroid cell TF network has the smallest number of occurrences. The two cancer cells (green dots) always stick to each other in all scatter plots, indicating that their networks should have very similar motif counts, and perhaps, other topological structures as well.

Finally, the scatter plots in the first row of Fig. 3.16 show that the embryonic stem cell network (black dot) seems to be far away from other networks and the regression line. Our further analysis of the regression residuals confirms that its residual is more

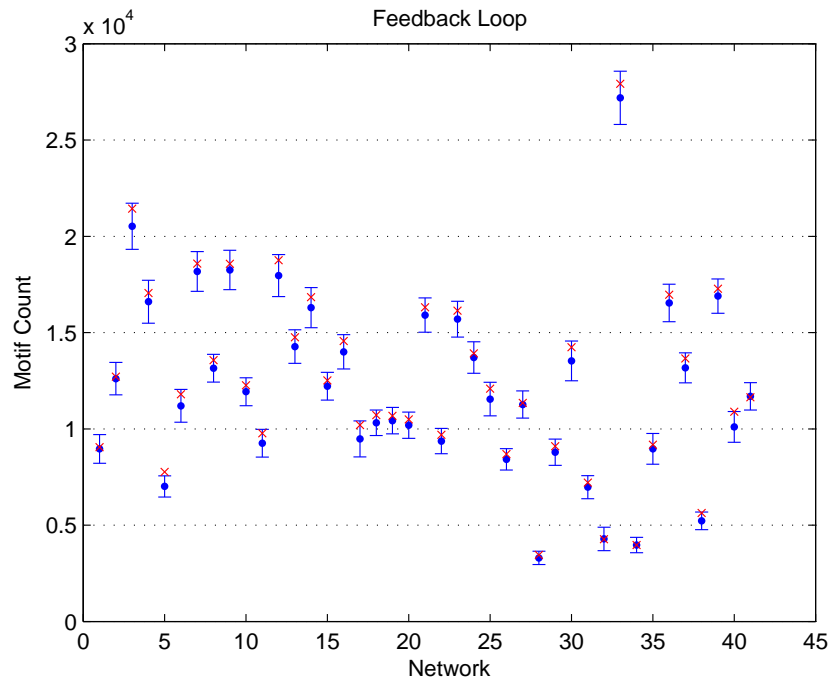


Figure 3.14: Motif count of feedback loop in forty-one observed networks (red “x”) and in their randomly rewired replicates ( $\mu \pm 3\sigma$  from 50 replicates for each observed network).

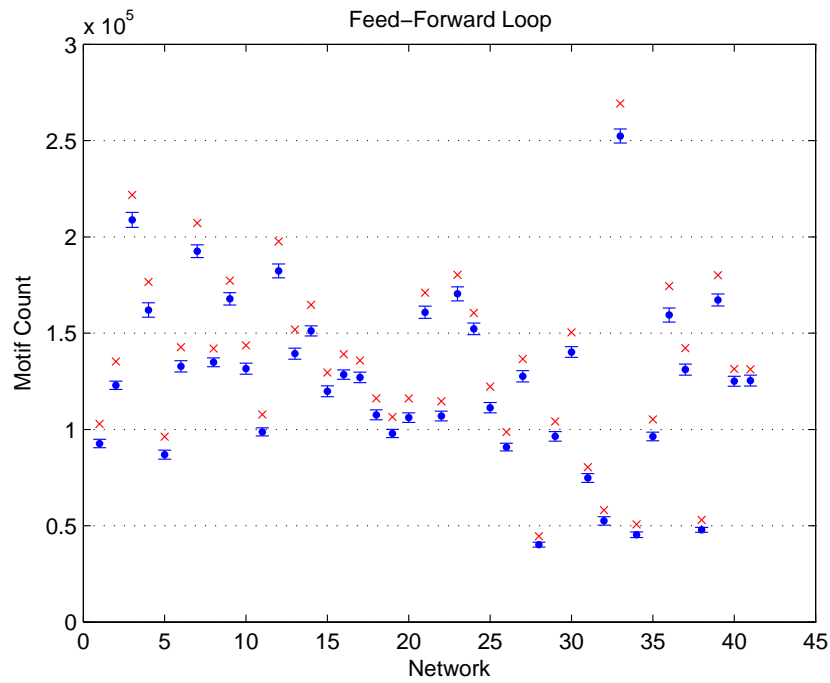


Figure 3.15: Motif count of feed-forward loop in forty-one observed networks (red “x”) and in their randomly rewired replicates ( $\mu \pm 3\sigma$  from 50 replicates for each observed network).

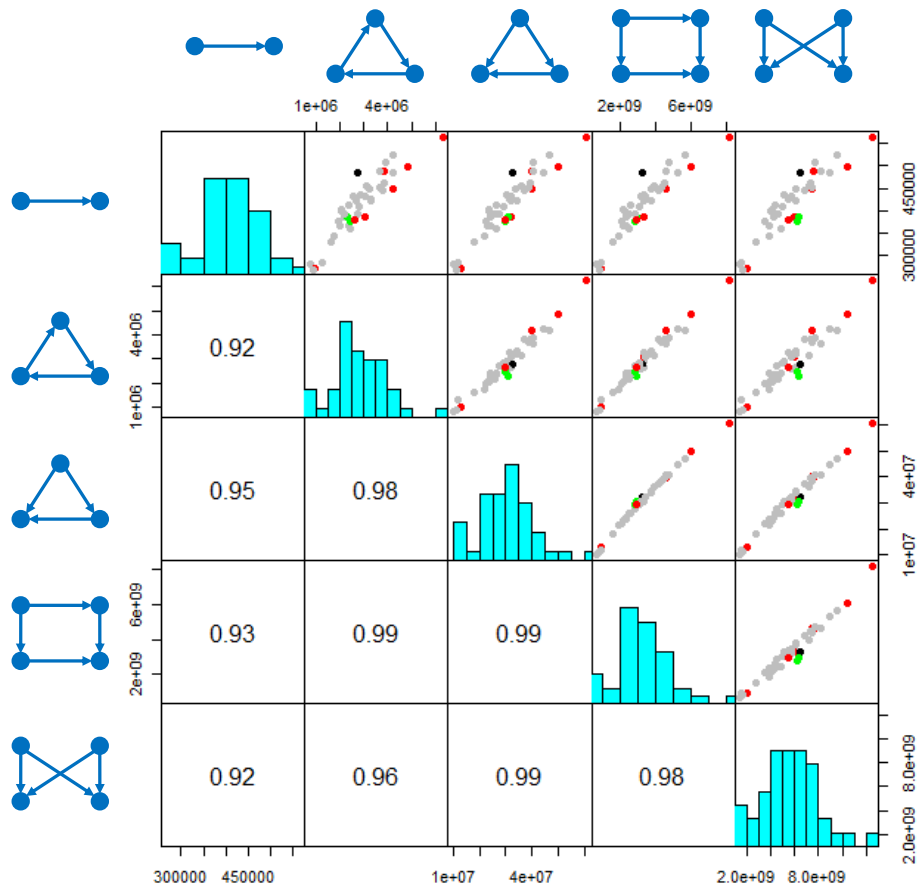


Figure 3.16: Correlation of motif counts in forty-one Human cell-specific transcription factor (TF) regulatory networks. The TF regulatory networks were obtained from genome-wide in vivo DNaseI footprints map [38]. The upper diagonal panels are scatter plots of the motif counts in these forty-one networks, including 1 embryonic stem cell (black), 7 blood cell types (red), 2 cancer cell types (green), 31 other cell and tissue types (grey). The diagonal panels show the distribution of the motif counts. Their correlation coefficients are given in the lower diagonal panels.

than  $3\times$  standard deviation from other networks’ residual. This means that the embryonic stem cell TF network has the smallest number of motif counts relative to its network size, suggesting that the embryonic stem cell TF network is less well-structured than other “mature” networks

### 3.5 Summary

In the first half of this chapter, we have presented the simulation results to validate the theoretical results obtained in chapter 2. We have studied the performance of our proposed estimators  $\widehat{N}_{\mathcal{M}}$  and  $\widetilde{N}_{\mathcal{M}}$  on random networks generated from four random graph models: Erdos-Renyi, geometric, preferential attachment, and duplication models. We have also performed simulation validation on “real” PPI networks. For each combination of network and motif, we assess the accuracy the estimators with respect to the network parameters  $n$ ,  $\rho$ , and the sampling parameters  $p$ ,  $r_-$ ,  $r_+$ . All simulation results confirm the asymptotically unbiased and consistent properties of our proposed estimators, as have been proved in Theorem 1, Theorem 2, and Proposition 1 in chapter 2.

In the second half of this chapter, we first applied our proposed estimators to the PPI network of four species: Yeast, Worm, Human, and Arabidopsis. Our estimates of the number of interactions, i.e. the size of the interactomes, are consistent with the results from the Center for Cancer System Biology [27, 28, 29, 30]. Our estimate of the number of triangles indicates that the PPI network of Human and Worm have a higher clustering and well-connected structure than that of Yeast and Arabidopsis. The GO annotation analysis further shows that triangles may play important roles in PPI networks as they are significantly enriched in shared GO annotations.

We also applied our method to estimate the number of occurrences of link, feed-

back loop, feed-forward loop, bi-fan and bi-parallel in forty-one Human cell-specific TF regulatory networks. Our estimation shows that the feed-forward loop and bi-fan are significantly enriched in these forty-one networks. We also found a striking feature that there is a strong linear correlation between the motif counts in the networks. Finally, our estimation suggests that cell-specific network motif counts are associated with the functional class of the cell.



# Chapter 4

## Discussion

In this chapter we discuss some limitations of our study and some potential solutions to address those problems. We then summarize all results obtained in the study and their contributions to Network Biology, as well as other fields of Network Sciences. We conclude this thesis with discussion on some further perspectives for future research.

### 4.1 Networks with different types of nodes

#### 4.1.1 Baits and Preys in PPI networks

As discussed in chapter 1, a protein-protein interaction is detected in Y2H experiments when one protein is used as “bait” and the other is used as “prey”. In high-throughput Y2H experiments, in order to achieve the best sensitivity, all proteins in the experiments are tested in two rounds, that is, both as bait and prey. In other words, the test space is a square matrix. However, there are some limitations that could cause the set of bait proteins slightly different from the set of prey proteins. For example, some proteins are auto-activators and they cannot be used as bait. This issue violates our assumption that the adjacency matrix of the observed subnetwork is a square  $n^{obs} \times n^{obs}$  matrix,

because if  $n^{bait} \neq n^{prey}$ , how should we choose  $n^{obs}$ .

Actually, this problem can be considered as part of the sensitivity of the Y2H assay: this type of experiment cannot detect interactions for auto-activating proteins. Hence, an accurate estimate of the assay sensitivity, as what has been done in [27], already takes this issue into account. Thus, one can simply use the total number of proteins involved in the Y2H experiment as  $n^{obs}$ , and there is no need to care about this bait and prey problem. And this is actually what we have done for the PPI networks in chapter 3.

Another more conservative approach is to consider the subnetwork  $\mathcal{G}^{obs}$  that is induced only by proteins in the intersection of the set of bait proteins and the set of prey proteins. In this way, the assumption  $n^{obs} = n^{bait} = n^{prey}$  holds in this induced subnetwork. Mathematically, this approach should not cause any problem.

First, to form the set of bait proteins, we assume that each of  $n$  proteins in the real network  $\mathcal{G}$  is uniformly selected at random with probability  $p_{bait}$ . Similarly, this process is repeated again with probability  $p_{prey}$  to obtain the set of prey proteins. We can use Bernoulli random variables  $X_i$  and  $Y_i$ ,  $1 \leq i \leq n$ , to indicate if node  $i$  is selected as bait ( $X_i = 1$ ) and prey ( $Y_i = 1$ ). Then we use  $Z_i = X_i \times Y_i$  to indicate if node  $i$  belongs to the intersection of the set of bait proteins and the set of prey proteins.

The only assumption we need is that  $X$ 's and  $Y$ 's are independent. Then we have  $Z_i$  follows Bernoulli distribution with probability  $p_{bait} \times p_{prey}$ . Thus, the intersection of bait proteins and prey proteins can be constructed by uniformly sampling nodes from the real network  $\mathcal{G}$  with probability  $p_{bait} \times p_{prey}$ . In other words, the observed subnetwork  $\mathcal{G}^{obs}$  is obtained from the real network  $\mathcal{G}$  via the uniform node sampling scheme: each node from  $\mathcal{G}$  is randomly selected with probability  $p_{bait} \times p_{prey}$ ,  $0 < p_{bait}, p_{prey} \leq 1$ . Hence, we only need to replace the parameter  $p$  by the product of the two parameters  $p_{bait}$ ,  $p_{prey}$ , and all theorems still hold.

Among the four PPI networks of Yeast, Worm, Human, and Arabidopsis in chapter 3, the condition  $n^{bait} = n^{prey}$  holds for Human and Worm, as confirmed by the authors from CCSB (personal communication via email). For Yeast [28], 5,487 proteins have been tested, 3,917 of them were used as bait, 5,246 of them were used as prey, and 3,676 were used both as bait and prey. For Arabidopsis [30], 8,430 proteins have been tested, 7,645 of them were used as bait, 7,896 of them were used as prey, and 7,108 were used both as bait and prey. We have re-estimated the number of links and triangles in Yeast and Arabidopsis from the subnetworks induced only by the intersection of bait proteins and prey proteins. Table 4.1 shows that the new estimates are not very different from the old ones, especially the link density and the triangle density. The overall message still remains the same.

Table 4.1: Re-estimation of the interactome size and the number of triangles in the PPI networks from the intersection of the set of bait proteins and the set of prey proteins.

	<i>S. cerevisiae</i>	<i>C. elegans</i>	<i>H. sapiens</i>	<i>A. thaliana</i>
<b>Interactome size</b>				
Estimate <sup>a</sup>	17,000	121,000	210,000	289,000
Mean $\pm$ SD <sup>a</sup>	18,000 $\pm$ 2,800	122,000 $\pm$ 16,600	214,000 $\pm$ 32,200	295,000 $\pm$ 33,400
Link density <sup>a</sup>	$9 \times 10^{-4}$	$6 \times 10^{-4}$	$8 \times 10^{-4}$	$8 \times 10^{-4}$
Estimate <sup>b</sup>	14,000	(same)	(same)	377,000
Mean $\pm$ SD <sup>b</sup>	15,000 $\pm$ 2,700	(same)	(same)	376,000 $\pm$ 45,600
Link density <sup>b</sup>	$8 \times 10^{-4}$	(same)	(same)	$10 \times 10^{-4}$
<b>No. of triangles</b>				
Estimate <sup>a</sup>	82,000	6,263,000	10,270,000	7,381,000
Mean $\pm$ SD <sup>a</sup>	75,000 $\pm$ 38,400	5,971,000 $\pm$ 3,593,800	11,255,000 $\pm$ 4,717,100	7,720,000 $\pm$ 3,132,700
Triangle density <sup>a</sup>	$2 \times 10^{-6}$	$5 \times 10^{-6}$	$5 \times 10^{-6}$	$2 \times 10^{-6}$
Estimate <sup>b</sup>	53,000	(same)	(same)	10,697,000
Mean $\pm$ SD <sup>b</sup>	61,000 $\pm$ 33,800	(same)	(same)	10,158,000 $\pm$ 4,289,000
Triangle density <sup>b</sup>	$1 \times 10^{-6}$	(same)	(same)	$3 \times 10^{-6}$

<sup>a</sup> Estimates from the full datasets

<sup>b</sup> Estimates from the subnetworks induced only by the intersection of bait proteins and prey proteins

## 4.1.2 Transcription factors and target genes in gene regulatory networks

There are typically two types of nodes in gene regulatory networks: transcription factors and target genes. A transcription factor (TF) can regulate as well as can be regulated

by other transcription factors, whereas a target gene (TG) is regulated by transcription factors and cannot regulate any other gene. As a result, for the same motif type, there may be different variants depending on the type of nodes. For example, there are two types of regulatory interaction:  $\text{TF} \rightarrow \text{TF}$  and  $\text{TF} \rightarrow \text{TG}$ . This problem has been studied in [74] in which the authors estimated the number of three different types of regulatory interactions separately:  $\text{TF} \rightarrow \text{TF}$ , TF self-interaction, and  $\text{TF} \rightarrow \text{TG}$ . Similarly, a feed-forward loop may consist of three TFs, or two TFs and one TG. Thus, one may want to distinguish these two types of feed-forward loop and estimate the number of occurrences of each type. The first case, i.e. three-TF feed-forward loop, is exactly the same as what we have done in chapter 3 for forty-one Human cell-specific TF regulatory networks which only contain TFs, no TG. For the second case, one can apply the same “extrapolation” approach, but with a different scaling factor:  $\frac{\binom{n_{TF}}{2} \times n_{TG}}{\binom{n_{TF}^{obs}}{2} \times n_{TG}^{obs}}$ . This can be further generalized for any arbitrary motif  $\mathcal{M}$ . However, theoretical analysis of the estimators and how to deal with noisy data are challenging problems for future research.

## 4.2 Effects of sampling schemes on the estimation

In order to make inference of the properties of a network  $\mathcal{G}$  from its observed subnetwork  $\mathcal{G}^{\text{obs}}$ , we need a suitable model to describe how the subnetwork can be obtained from the entire network. In chapter 2 and chapter 3, we consider the uniformly random node sampling scheme that independently select nodes from the entire network  $\mathcal{G}$  with some given probability  $p$ ,  $0 < p < 1$ , and then form the subnetwork  $\mathcal{G}^{\text{obs}}$  from the sampled nodes. More specifically, independent and identically distributed Bernoulli random variables  $X_i$  with parameter  $0 < p < 1$  are used to denote the event whether the node  $i$  is sampled ( $X_i = 1$ ) or not ( $X_i = 0$ ),  $i = 1, 2, \dots, n$ .

In practice, biologists, however, do not randomly choose genes or proteins for inves-

tigation but follow certain underlying hypotheses. In this section, we further explore different random sampling schemes, which can be used to draw a subnetwork from a real network, and their effects on the estimation of motif counts. For instance, highly connected proteins are believed to be more essential than others, and thus are more likely to be the target of biological studies. Hence, a random node sampling scheme with probabilities proportional to node degrees also represents a good model to study PPI subnetworks data.

First, we still keep the assumption of independence, but allow nodes in  $\mathcal{G}$  to be sampled with different probabilities:

$$X_i \sim \text{Bernoulli}(p_i), 0 < p_i < 1, i = 1, 2, \dots, n. \quad (4.1)$$

To keep the average proportion of sampled nodes to be  $p$ ,  $0 < p < 1$ , we consider the following two distributions of  $p_i$ :

1. Each  $p_i$  is randomly drawn from a uniform distribution

$$p_i \sim \text{Uniform}\left(\frac{p}{2}, \frac{3p}{2}\right). \quad (4.2)$$

2. Each  $p_i$  is randomly drawn from a normal distribution

$$p_i \sim \text{Normal}\left(\mu = p, \sigma^2 = \frac{p^2}{16}\right). \quad (4.3)$$

Here, the variance of the distributions is chosen so that randomly generated values of  $p_i$  are not likely to be negative and the mean of  $p_i$  is  $p$  in each sampling scheme. When  $p_i \geq 1$ , we take it to be 1.

Biological networks are often modeled as scale-free, that is, most of the nodes have low degrees, whereas a small number of nodes have significantly high degrees. As

mentioned earlier, highly connected proteins are more likely to play many important functions through their vast repertoire of interactions with other proteins. Thus, one may select proteins with probabilities proportional to their degrees in the corresponding PPI network. To take such bias selection into account, we also consider the following degree-based sampling scheme:

3. Each  $p_i$  is calculated from the node degrees as follows

$$p_i = \min\left\{np \frac{d_i}{\sum_{j=1}^n d_j}, 1\right\}, \quad (4.4)$$

where  $d_i$  is the degree of the node  $i$  in the corresponding network  $\mathcal{G}$ . Such a sampling scheme also has the average sampling proportion  $p$ .

Next, we perform simulation validation to examine the effects of the above three non-uniform sampling schemes on the motif counts estimation. As the degree-based sampling scheme (scheme 3) is likely to have more significant impacts when applied to scale-free networks, we generate network  $\mathcal{G}$  from the preferential attachment model, the best random graph model that can capture the scale-free property of real-world networks. We choose the number of nodes,  $n$ , from 5,000 to 30,000, as most species have their genome size in that range. The link density  $\rho$  is fixed at 0.001 so that the resulting networks are similar to real PPI networks studied in this work. Another important feature of scale-free networks is the exponent  $\gamma$  of the power-law degree distribution. We choose  $\gamma = 1.5, 2, 2.5, 3, 3.5$ , which is the common range for most biological networks.

We also consider different values of the average sampling proportion  $p = 0.1, 0.2, \dots, 0.9$ . For each  $p$  and each non-uniform sampling scheme, we first generate a vector of sampling probabilities  $(p_1, p_2, \dots, p_n)$ . For each sampling vector, we sample 50 subnetworks and estimate the number of links and triangles using the estimators  $\widehat{N}_1$  and  $\widehat{N}_3$ , respec-

tively. Here we only consider the error-free case, that is, there is no link error in the sampled subnetworks.

Fig. 4.1 demonstrates how the MSE of the estimator  $\widehat{N}_3$  for triangle count depends on the sampling scheme and the average sampling proportion  $p$ . For each of the four sampling schemes considered (uniform and three non-uniform schemes), the MSE decreases when  $p$  increases as expected. The first two non-uniform sampling schemes are not significantly different from the uniform node sampling scheme. However, when the degree-bias sampling scheme is used, the MSE is significantly higher than that of the other three. This is not surprising because bias selection towards highly connected nodes should lead to over-estimation.

Beside the preferential attachment model, we also study how the effects of the degree-bias sampling scheme when applied to the Erdos-Renyi, geometric and duplication models. Fig. 4.2 further confirms the over-estimation bias since the mean of the ratio of estimate to real count (i.e.,  $\widehat{N}_3/N_3$ ) is larger than one. Moreover, it can be seen clearly from the figure that the over-estimation bias is significantly higher for networks generated from the preferential attachment and duplication models because of their scale-free degree distribution.

Fig. 4.3 shows how the over-estimation bias of the degree-bias sampling scheme depends on the power-law exponent  $\gamma$  of scale-free networks. The mean of the ratio  $\widehat{N}_3/N_3$  slightly decreases as the power-law exponent  $\gamma$  increases. Interestingly, the ratio mean does not change much when we change  $\gamma$ . This shows that estimation is robust against different choices of exponents in the power law. Fig. 4.3 also demonstrates how the mean of the ratio  $\widehat{N}_3/N_3$  depends on the average sampling proportion  $p$ . In particular, when more than 60% of nodes in  $\mathcal{G}$  are sampled, the estimate is less than five times the real count.

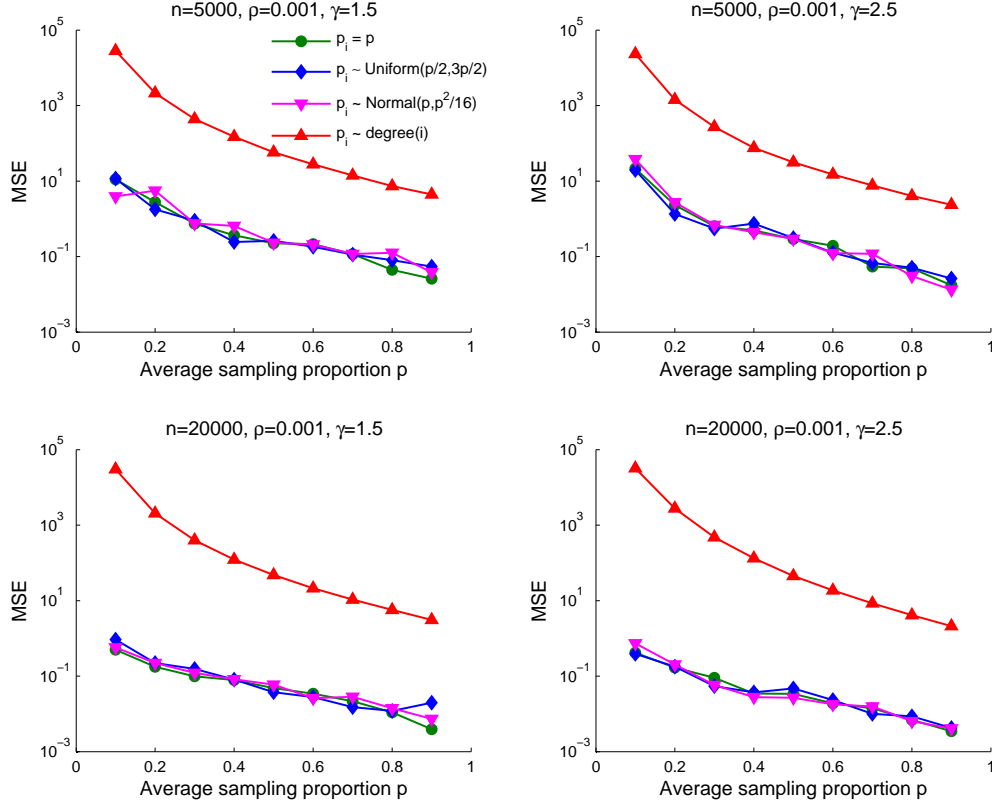


Figure 4.1: Plots of the MSE of the estimator  $\widehat{N}_3$  for triangle count with respect to four different sampling schemes and average sampling proportion  $p$ . Random networks were generated in the preferential attachment model with different parameters  $n, \rho, \gamma$ . Subnetworks are drawn using four sampling schemes:  $p_i = p$  (uniform node sampling),  $p_i \sim \text{Uniform}(\frac{p}{2}, \frac{3p}{2})$ ,  $p_i \sim \text{Normal}(\mu = p, \sigma^2 = \frac{p^2}{16})$ ,  $p_i = np \frac{d_i}{\sum_{j=1}^n d_j}$ ,  $i = 1, 2, \dots, n$ .

### 4.3 Linear correlation of motif counts

In chapter 3 we have reported a striking feature of the motif counts in forty-one Human cell-specific TF regulatory networks: the strong linear correlation between the counts of triad and quadriad motifs, although these motifs are topologically very different and the TF networks are from very different cell types (Fig. 3.16). To further explore if this observation has any biological meaning, we generate random replicates from these real networks using the link rewiring process. Surprisingly, we still observe a strong linear



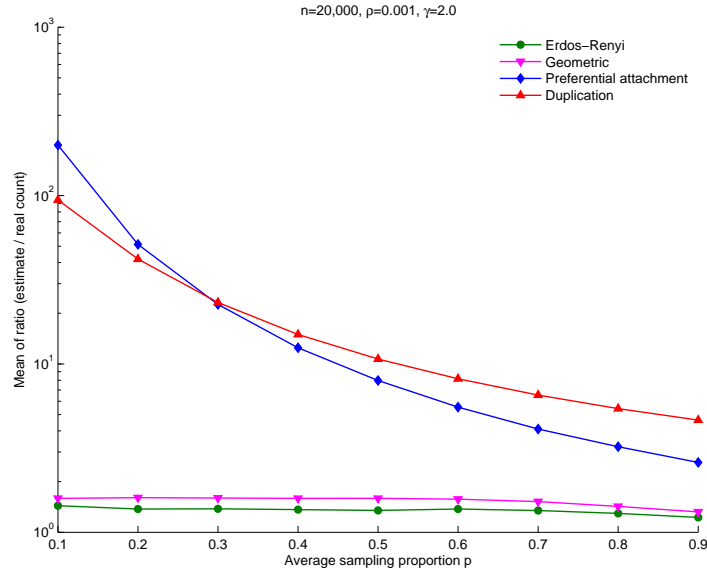


Figure 4.2: Plots of the mean of the ratio  $\frac{\widehat{N}_3}{N_3}$  for triangle count with respect to four network models and increasing average sampling proportion  $p$ . Random network is generated using the Erdos-Renyi, geometric, preferential attachment, and duplication models, where  $n = 20,000$  nodes, link density  $\rho = 0.001$ , power-law exponent  $\gamma = 2.0$  (for preferential attachment and duplication models). Subnetworks are drawn using the degree-bias sampling scheme with  $p_i = np \frac{\text{degree}(i)}{\sum_{j=1}^n \text{degree}(j)}$ ,  $i = 1, 2, \dots, n$ .

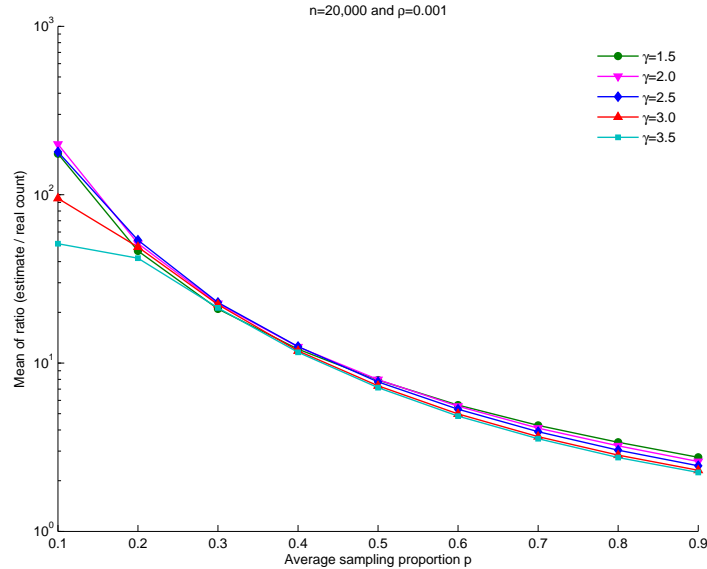


Figure 4.3: Plots of the mean of the ratio  $\frac{\widehat{N}_3}{N_3}$  for triangle count with respect to the power-law exponent  $\gamma$  and average sampling proportion  $p$ . Random networks were generated in the preferential attachment model, which have 20,000 nodes and link density  $\rho = 0.001$ . Subnetworks were drawn using the degree-bias sampling scheme with  $p_i = np \frac{d_i}{\sum_{j=1}^n d_j}$ ,  $i = 1, 2, \dots, n$ .

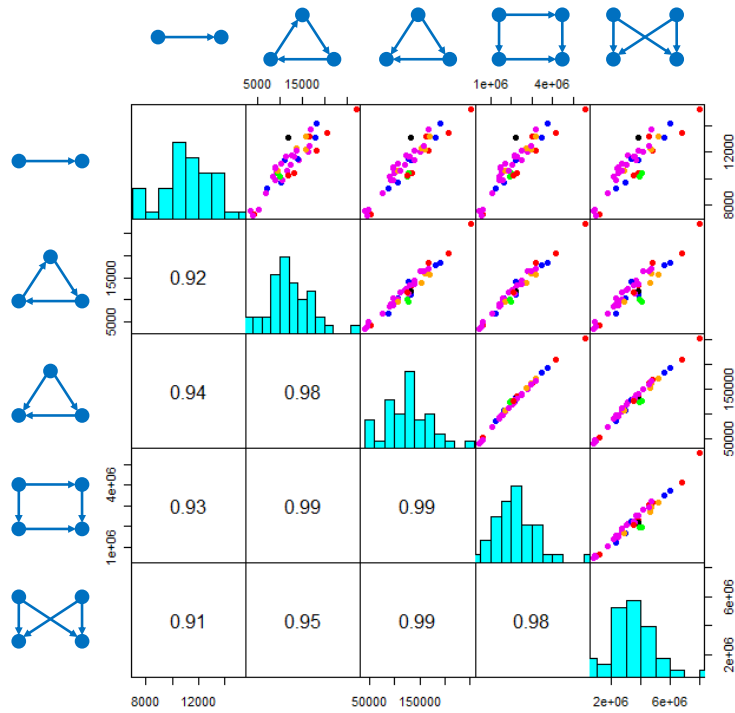


Figure 4.4: Linear correlation of the motif counts in random networks which are generated from the forty-one Human cell-specific TF regulatory networks using the link rewiring process. Different colors in the scatter plots correspond to different functional classes of the cells.

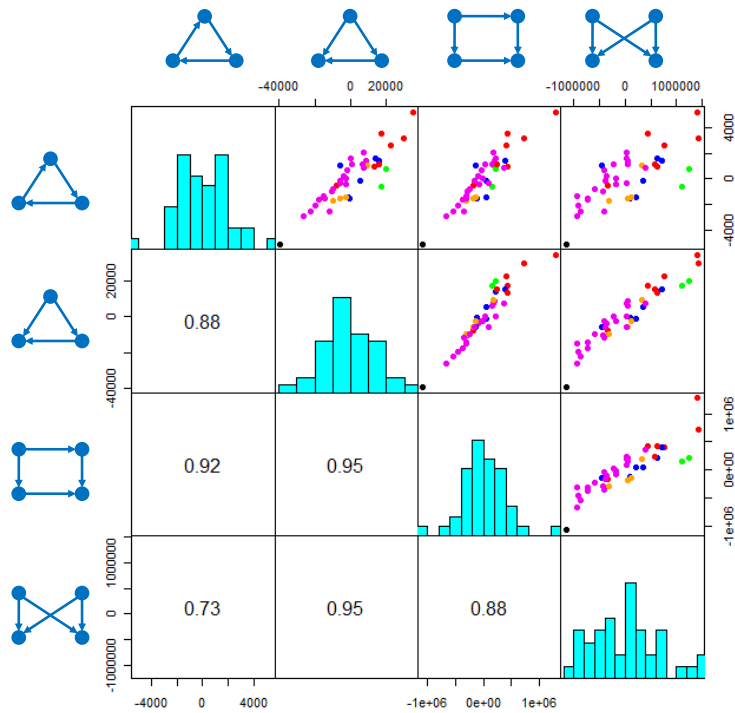


Figure 4.5: Linear correlation of the residuals of the motif counts' regression with respect to the number of links in forty-one Human cell-specific TF regulatory networks.

correlation between the motif counts in these randomly rewired networks (Fig. 4.4). One possible explanation could be: “more links, more motifs”. Hence, we perform linear regression of the motif counts on the number of links, and then study the residuals. Again, the linear pattern is still there, but this time with lower correlation coefficients. We also perform similar analysis with networks generated from the four random graph models and found the same correlation. We are currently still studying this interesting phenomenon to figure out reasonable explanations as well as some possible biological meanings.

## 4.4 Conclusion

In this thesis, we have proposed a simple, yet powerful method to estimate the number of occurrences of any arbitrary motif in biological networks from their observed subnetworks. More importantly, we have addressed the problem of the inaccuracy and incompleteness of data, and developed bias-corrected estimators to account for noise, that is, spurious and missing links, in observed subnetworks.

In chapter 2 of the thesis, we have performed rigorous theoretical analysis on the properties of our proposed estimators and proved that the estimators are asymptotically unbiased and consistent for any type of motifs and regardless of the structure of the underlying network. Thus, the method can be further applied to any network across diverse fields of Network Sciences, including, but not limited to, biological networks, social networks, the World-Wide-Web, engineering and electrical circuitries, etc. In particular, the simulation results presented in chapter 3 further showed that the proposed estimators performed well in networks generated from four random graph models: Erdos-Renyi, preferential attachment, duplication, and geometric, which are most commonly used to study real-world networks. For each combination of model and

motif, we carefully examined the accuracy of the estimators with respect to different network parameters and sampling parameters. Our simulation results also showed that even if the network is completely known, the sampling-estimating approach is more efficient than exhaustive enumeration approach, while still providing comparably accurate results.

In chapter 3 we applied the method to estimate the number of links and triangles in four protein-protein interaction networks of Yeast, Worm, Human, and Arabidopsis. Our estimates of the interactome size of these four species agree well with previous studies, indicating the accuracy of the proposed estimators. We found that the estimated triangle density in Human and Worm were 2.5 times larger than that in Yeast and Arabidopsis, whereas the later have higher link density than the formers, indicating a higher clustering and well-connected structure of the PPI network of Human and Worm. We also performed Gene Ontology (GO) annotation analysis of triangles in the observed PPI subnetwork of Yeast. Our results showed that those triangles are significantly enriched in shared GO annotations, suggesting that triangles may also play important roles in PPI networks.

For the case of gene regulatory networks, we found a strong positive linear correlation between the number of occurrences of different three-node and four-node motifs in forty-one Human cell-specific transcription factor regulatory networks. Our estimation also shows that the feed-forward loop and bi-fan are significantly enriched in these forty-one networks, and the motif counts are highly associated with the functional class of the cell. For instance, the embryonic stem cell exhibits a larger number of links, but a smaller number of motifs, when compared to other more “mature” cell types. The properties of protein-protein interaction and gene regulatory networks uncovered in our study are consistent with our biological intuition about the complexity of living organisms.

The problem of estimating the size of interactomes, i.e. the number of links, has been the target of several studies. One of the first attempt was proposed in [72] in which the author estimated the average degree in the protein-protein interaction network of Yeast by modelling the overlap between two independent datasets from [23, 24]. The authors in [70] further extended the method in [72] by incorporating the false discovery rate. However, this method requires that the two datasets must be generated from identical, or at least similar experimental conditions, but this is rarely the case for biological networks data. Moreover, this approach cannot be generalized to the case of larger motifs.

The second approach to address this problem is the “extrapolation” idea. In [73] the authors scaled up the number of interactions in observed protein-protein interaction subnetworks to estimate the size of real interactomes, assuming that the link density in a real interactome can be approximated by the link density in its observed subnetworks. However, their proof of the unbiasedness property of the estimator was not complete. In Theorem 1 and 2, chapter 2, we have showed both asymptotically unbiased and consistent properties of the estimators. Moreover, in Theorem 3, we generalized the results to any arbitrary motif. Most importantly, all results hold regardless of the topological structures of the underlying network.

Taking into account noise in observed subnetworks, the authors from CCSB [27, 28, 29, 30] provided accurate estimates of the interactome size by carefully considering different quality parameters of Y2H datasets. They handled the link errors in a slightly different way from ours. In particular, our bias-corrected estimator  $\tilde{N}^{(1)}$  requires the false positive rate  $r_+$  and the false negative rate  $r_-$ , whereas the CCSB estimator  $\tilde{N}^{\text{CCSB}}$  requires the precision and the sensitivity. We have proved that the two estimators are mathematically equivalent, and indeed, our estimates of the interactome size from real datasets agree well with those of CSSB for all species. However, in the field of Network

Biology, gold-standard sets, which are usually constructed from literature curation, are limited and biased [27, 28]. In such cases, the estimated precision and false discovery rate are less reliable than the sensitivity, the false positive rate  $r_+$ , and the false negative rate  $r_-$  (as discussed in chapter 3). Most importantly, unlike the CCSB estimator, our method can be easily generalized to larger motifs such as triangles, feed-forward loop, bi-fan, etc.

The key assumption in our study is the uniformly random node sampling scheme. In practice, biologists do not randomly choose genes or proteins for investigation but follow certain underlying hypotheses. For example, highly connected proteins are believed to be more essential than others, and thus are more likely to be the target of biological studies. Hence, a random node sampling scheme with probabilities proportional to node degrees also represents a good model to study protein-protein interaction subnetworks data. In the previous sections, we have explored three non-uniform sampling schemes, especially the degree-bias sampling scheme. We study how the over-estimation bias depends on the scale-free structure of the network as well as the node sampling proportion. How to correct the bias still remains as a challenging problem for future research. The results we obtained here for the uniformly random node sampling scheme thus serve as a guideline to study more complicated, yet interesting sampling schemes.

In the previous sections we have also discussed some other limitations the study. For example, the huge amount of data in Network Biology eventually leads to the integration of different types of datasets. Thus, the network of interest may consider different types of nodes and links, and hence, even more complicated types of motifs which make the motif count estimation problem more challenging. Another question that we haven't been able to answer is the linear correlation of the motif counts in forty-one Human TF regulatory networks: what are the possible explanations, and what implications from that observation? There are also some other limitations that we haven't discussed, for

example, the assumption of the independence and homogeneity of link errors. There is always a trade-off between the two error rates, false positive and false negative, and thus, a bivariate distribution may be better to model the link errors. All of these limitations are potential topics for future research to fully explore all characteristics of biological networks from noisy and incomplete data.

The approach of inferring properties of a real network from its observed subnetworks is especially useful in Network Biology because of the current limitation of biological networks data. Motivated by the estimation of the number of motif occurrences, further research questions may focus on other network properties such as degree distribution, clustering coefficient, shortest path length, etc., which also can be inferred from observed subnetworks. However, handling the experimental errors, that is, spurious and missing interactions, especially the high false negative rate, still remains as the most challenging problem to the research community.

# Appendix

## Proof of Theorem 2

**Theorem.** Let  $\mathcal{G}$  be an arbitrary undirected network of  $n$  nodes, and  $\mathcal{G}^{obs}$  be a sub-network obtained from  $\mathcal{G}$  via a uniformly random node sampling process that selects a node with probability  $p$ ,  $0 < p < 1$ . Let  $N_1$  denote the number of links in  $\mathcal{G}$ , and  $N_1$  is estimated by the estimator  $\hat{N}_1$  defined in Equation (2.3). Let  $N_2$  denote the number of three-node paths in  $\mathcal{G}$  (that is, the number of pairs of links that share exactly one common node, see motif ID  $u_2$  in Figure 2.1). We have:

$$\text{Var} \left( \frac{\hat{N}_1}{N_1} \right) = \frac{2q}{p} \frac{N_2}{N_1^2} [1 + O(n^{-1})] + \frac{(1+p)q}{p^2 N_1} [1 + O(n^{-1})] + O(n^{-1}),$$

where  $q = 1 - p$ .

*Proof.* First recall that the estimator  $\hat{N}_1$  has the following form:

$$\hat{N}_1 = \frac{n(n-1)}{n^{obs}(n^{obs}-1)} \sum_{1 \leq i < j \leq n} a_{i,j} X_i X_j.$$



Hence, we compute the variance of  $\widehat{N}_1$  as follows:

$$\begin{aligned}
n^{-2}(n-1)^{-2}\text{Var}\left(\widehat{N}_1\right) &= \text{Var}\left(\sum_{1 \leq i < j \leq n} a_{i,j} \frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}}-1)}\right) \\
&= \sum_{1 \leq i < j \leq n} a_{i,j} \text{Var}\left(\frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}}-1)}\right) \\
&\quad + \sum_{(i,j,k,l) \in \Gamma_1} a_{i,j} a_{k,l} \text{Cov}\left(\frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}}-1)}, \frac{X_k X_l}{n^{\text{obs}}(n^{\text{obs}}-1)}\right) \\
&\quad + \sum_{(i,j,k,l) \in \Gamma_0} a_{i,j} a_{k,l} \text{Cov}\left(\frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}}-1)}, \frac{X_k X_l}{n^{\text{obs}}(n^{\text{obs}}-1)}\right),
\end{aligned} \tag{5.1}$$

where

$$\Gamma_0 = \{(i, j, k, l) : 1 \leq i < j \leq n; 1 \leq k < l \leq n; \{i, j\} \cap \{k, l\} = 0\}, \tag{5.2}$$

$$\Gamma_1 = \{(i, j, k, l) : 1 \leq i < j \leq n; 1 \leq k < l \leq n; \{i, j\} \cap \{k, l\} = 1\}. \tag{5.3}$$

Note that in the second equality we use the fact that  $a_{i,j}^2 = a_{i,j}$  for any  $1 \leq i < j \leq n$ .

Since random variables  $X_i$  are independent and identically distributed (*i.i.d*),  $1 \leq i \leq n$ , we have:

$$\begin{aligned}
\text{Var}\left(\frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}}-1)}\right) &= \text{Var}\left(\frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}}-1)}\right), \text{ for } 1 \leq i < j \leq n, \\
\text{Cov}\left(\frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}}-1)}, \frac{X_k X_l}{n^{\text{obs}}(n^{\text{obs}}-1)}\right) &= \text{Cov}\left(\frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}}-1)}, \frac{X_2 X_3}{n^{\text{obs}}(n^{\text{obs}}-1)}\right), \text{ for } (i, j, k, l) \in \Gamma_1, \\
\text{Cov}\left(\frac{X_i X_j}{n^{\text{obs}}(n^{\text{obs}}-1)}, \frac{X_k X_l}{n^{\text{obs}}(n^{\text{obs}}-1)}\right) &= \text{Cov}\left(\frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}}-1)}, \frac{X_3 X_4}{n^{\text{obs}}(n^{\text{obs}}-1)}\right), \text{ for } (i, j, k, l) \in \Gamma_0.
\end{aligned}$$

Thus, Equation (5.1) can be rewritten as follows:

$$\begin{aligned}
n^{-2}(n-1)^{-2}\text{Var}\left(\widehat{N}_1\right) &= \text{Var}\left(\frac{X_1X_2}{n^{\text{obs}}(n^{\text{obs}}-1)}\right) \sum_{1 \leq i < j \leq n} a_{i,j} & (5.4) \\
&+ \text{Cov}\left(\frac{X_1X_2}{n^{\text{obs}}(n^{\text{obs}}-1)}, \frac{X_2X_3}{n^{\text{obs}}(n^{\text{obs}}-1)}\right) \sum_{(i,j,k,l) \in \Gamma_1} a_{i,j}a_{k,l} \\
&+ \text{Cov}\left(\frac{X_1X_2}{n^{\text{obs}}(n^{\text{obs}}-1)}, \frac{X_3X_4}{n^{\text{obs}}(n^{\text{obs}}-1)}\right) \sum_{(i,j,k,l) \in \Gamma_0} a_{i,j}a_{k,l}.
\end{aligned}$$

Note that

$$E\frac{X_1X_2}{n^{\text{obs}}(n^{\text{obs}}-1)} = E\frac{X_2X_3}{n^{\text{obs}}(n^{\text{obs}}-1)} = E\frac{X_3X_4}{n^{\text{obs}}(n^{\text{obs}}-1)}, \quad (5.5)$$

$$E\left(\frac{X_1X_2^2X_3}{(n^{\text{obs}})^2(n^{\text{obs}}-1)^2}\right) = E\left(\frac{X_1X_2X_3}{(n^{\text{obs}})^2(n^{\text{obs}}-1)^2}\right), \quad (5.6)$$

$$E\left(\frac{X_1^2X_2^2}{(n^{\text{obs}})^2(n^{\text{obs}}-1)^2}\right) = E\left(\frac{X_1X_2}{(n^{\text{obs}})^2(n^{\text{obs}}-1)^2}\right), \quad (5.7)$$

where in the first equation we use the fact that random variables  $X_i$  are *i.i.d.*, and in the last two equations we use the fact that  $X_i^2$  and  $X_i$  have the same Bernoulli distribution for any  $1 \leq i \leq n$ .

Hence, we let

$$\mu = E\frac{X_1X_2}{n^{\text{obs}}(n^{\text{obs}}-1)}, \quad (5.8)$$

$$\alpha_0 = E\left(\frac{X_1X_2X_3X_4}{(n^{\text{obs}})^2(n^{\text{obs}}-1)^2}\right), \quad (5.9)$$

$$\alpha_1 = E\left(\frac{X_1X_2X_3}{(n^{\text{obs}})^2(n^{\text{obs}}-1)^2}\right), \quad (5.10)$$

$$\alpha_2 = E\left(\frac{X_1X_2}{(n^{\text{obs}})^2(n^{\text{obs}}-1)^2}\right). \quad (5.11)$$

Then, we have

$$\alpha_0 - \mu^2 = \text{Cov} \left( \frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)}, \frac{X_3 X_4}{n^{\text{obs}}(n^{\text{obs}} - 1)} \right), \quad (5.12)$$

$$\alpha_1 - \mu^2 = \text{Cov} \left( \frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)}, \frac{X_2 X_3}{n^{\text{obs}}(n^{\text{obs}} - 1)} \right), \quad (5.13)$$

$$\alpha_2 - \mu^2 = \text{Var} \left( \frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)} \right), \quad (5.14)$$

and Equation (5.4) can be rewritten as follows:

$$\begin{aligned} n^{-2}(n-1)^{-2} \text{Var} \left( \widehat{N}_1 \right) &= (\alpha_0 - \mu^2) \sum_{1 \leq i < j \leq n} a_{i,j} \\ &+ (\alpha_1 - \mu^2) \sum_{(i,j,k,l) \in \Gamma_1} a_{i,j} a_{k,l} \\ &+ (\alpha_2 - \mu^2) \sum_{(i,j,k,l) \in \Gamma_0} a_{i,j} a_{k,l}. \end{aligned} \quad (5.15)$$

Since we have shown in Theorem 1 that

$$\mu = E \frac{X_1 X_2}{n^{\text{obs}}(n^{\text{obs}} - 1)} = \frac{1 - q^n - npq^{n-1}}{n(n-1)}, \quad (5.16)$$

it remains to calculate  $\alpha_0$ ,  $\alpha_1$ , and  $\alpha_2$ .

By conditioning on the event that  $X_1 = X_2 = X_3 = X_4 = 1$ , we have

$$\begin{aligned} \alpha_0 &= E \left( \frac{X_1 X_2 X_3 X_4}{(n^{\text{obs}})^2 (n^{\text{obs}} - 1)^2} \right), \\ &= p^4 E \left( (Z + 4)^{-2} (Z + 3)^{-2} \right), \end{aligned} \quad (5.17)$$

where  $Z \sim \text{Binomial}(n-4, p)$ .

Similarly, by conditioning on the event that  $X_1 = X_2 = X_3 = 1$ , we have

$$\begin{aligned}
\alpha_1 &= E \left( \frac{X_1 X_2 X_3}{(n^{\text{obs}})^2 (n^{\text{obs}} - 1)^2} \right) \\
&= p^3 E [(Z + \xi + 3)^{-2} (Z + \xi + 2)^{-2}] \\
&= p^4 E [(Z + 4)^{-2} (Z + 3)^{-2}] + p^3 q E [(Z + 3)^{-2} (Z + 2)^{-2}], \tag{5.18}
\end{aligned}$$

where  $\xi \sim \text{Bernoulli}(p)$ ,  $Z$  and  $\xi$  are independent.

And finally, by conditioning on the event that  $X_1 = X_2 = 1$ , we have

$$\begin{aligned}
\alpha_2 &= E \left( \frac{X_1 X_2}{(n^{\text{obs}})^2 (n^{\text{obs}} - 1)^2} \right) \\
&= p^2 E [(Z + \xi_1 + \xi_2 + 2)^{-2} (Z + \xi_1 + \xi_2 + 1)^{-2}] \\
&= p^4 E [(Z + 4)^{-2} (Z + 3)^{-2}] + 2p^3 q E [(Z + 3)^{-2} (Z + 2)^{-2}] \\
&\quad + p^2 q^2 E [(Z + 2)^{-2} (Z + 1)^{-2}], \tag{5.19}
\end{aligned}$$

where  $\xi_1, \xi_2 \sim \text{Bernoulli}(p)$ ,  $Z$ ,  $\xi_1$  and  $\xi_2$  are independent.

For simplicity, we define  $\gamma_m = E[(Z + m)^{-2} (Z + m + 1)^{-2}]$ , for  $m = 1, 2, 3$ . It follows that

$$\alpha_0 = p^4 \gamma_3, \tag{5.20}$$

$$\alpha_1 = p^4 \gamma_3 + p^3 q \gamma_2, \tag{5.21}$$

$$\alpha_2 = p^4 \gamma_3 + 2p^3 q \gamma_2 + p^2 q^2 \gamma_1. \tag{5.22}$$

On the other hand, let  $\beta_0$  be the number of pairs of edges in  $\mathcal{G}$  which have no common nodes. Recall that  $N_2$  is the number of pairs of edges in  $\mathcal{G}$  that have exactly

one common neighbor. Hence, we have

$$2\beta_0 = \sum_{(i,j,k,l) \in \Gamma_0} a_{i,j} a_{k,l}, \quad (5.23)$$

$$2N_2 = \sum_{(i,j,k,l) \in \Gamma_1} a_{i,j} a_{k,l}, \quad (5.24)$$

$$\beta_0 + N_2 = \binom{N_1}{2}. \quad (5.25)$$

Thus, Equation (5.15) can be rewritten as

$$\begin{aligned} & n^{-2}(n-1)^{-2} \text{Var}(\widehat{N}_1) \\ &= (\alpha_2 - \mu^2) \sum_{1 \leq i < j \leq n} a_{i,j} + (\alpha_0 - \mu^2) \sum_{(i,j,k,l) \in \Gamma_0} a_{i,j} a_{k,l} + (\alpha_1 - \mu^2) \sum_{(i,j,k,l) \in \Gamma_1} a_{i,j} a_{k,l} \\ &= N_1(\alpha_2 - \mu^2) + 2\beta_0(\alpha_0 - \mu^2) + 2N_2(\alpha_1 - \mu^2) \\ &= N_1(\alpha_2 - \alpha_0) + 2N_2(\alpha_1 - \alpha_0) + N_1^2(\alpha_0 - \mu^2) \\ &= N_1(2p^3q\gamma_2 + p^2q^2\gamma_1) + 2N_2(p^3q\gamma_2) + N_1^2(p^4\gamma_3 - \mu^2) \\ &= 2p^3q\gamma_2N_2 + p^2q(2p\gamma_2 + q\gamma_1)N_1 + (p^4\gamma_3 - \mu^2)N_1^2, \end{aligned} \quad (5.26)$$

where in the second equality we use (5.23) and (5.24), in the third equality we use (5.25), and in the last two equalities we used (5.20), (5.21), and (5.22).

Dividing both sides by  $N_1^2$  and using the fact that  $n^2(n-1)^2 = n^4 [1 + O(n^{-1})]$ , we have

$$\text{Var}\left(\frac{\widehat{N}_1}{N_1}\right) = n^4 [1 + O(n^{-1})] \left\{ \frac{2p^3q\gamma_2N_2}{N_1^2} + \frac{p^2q(2p\gamma_2 + q\gamma_1)}{N_1} + (p^4\gamma_3 - \mu^2) \right\}. \quad (5.27)$$

We need the following lemma to study the asymptotic behavior of  $\text{Var}\left(\frac{\widehat{N}_1}{N_1}\right)$  in Equation (5.27) as  $n \rightarrow \infty$ .

**Lemma.** *Recall the notations  $Z$  and the  $\gamma_m$ 's. We have*

- (i)  $n^4\gamma_2 = p^{-4} [1 + O(n^{-1})]$ ,
- (ii)  $n^4(2p\gamma_2 + q\gamma_1) = (1 + p)p^{-4} [1 + O(n^{-1})]$ ,
- (iii)  $n^4(p^4\gamma_3 - \mu^2) = O(n^{-1})$ .

*Proof.* Since  $\varphi(x) = (x + 2)^{-2}(x + 3)^{-2}$  is convex, we apply Jensen's inequality (i.e.,  $E[\varphi(X)] \geq \varphi(E(X))$ ) to obtain a lower bound on

$$\begin{aligned}
\gamma_2 &= E [(Z + 2)^{-2}(Z + 3)^{-2}] \geq [E(Z) + 2]^{-2}[E(Z) + 3]^{-2} \\
&= [(n - 4)p + 2]^{-2}[(n - 4)p + 3]^{-2} \\
&= [np + (2 - 4p)]^{-2}[np + (3 - 4p)]^{-2} \\
&= [(np)^2 + (5 - 8p)(np) + (2 - 4p)(3 - 4p)]^{-2} \\
&= (np)^{-4} - 2(5 - 8p)(np)^{-5} + O(n^{-6}),
\end{aligned}$$

where in the second equality we used the fact that  $Z \sim \text{Binomial}(n - 4, p)$ , and hence,  $E(Z) = (n - 4)p$ .

For upper bound, we proceed as

$$\begin{aligned}
\gamma_2 &= E [(Z + 2)^{-2}(Z + 3)^{-2}] \\
&\leq E \left[ \frac{1}{(Z + 1)(Z + 2)(Z + 3)(Z + 4)} \right] \\
&= E \left[ \frac{1}{6} \int_0^1 (1 - t)^3 t^Z dt \right] \\
&= \frac{1}{6} \int_0^1 (1 - t)^3 E(t^Z) dt \\
&= \frac{1}{6} \int_0^1 (1 - t)^3 (q + pt)^{n-4} dt \\
&= \frac{1 - q^n - npq^{n-1} - \binom{n}{2}p^2q^{n-2} - \binom{n}{3}p^3q^{n-3}}{p^4n(n-1)(n-2)(n-3)} \\
&= (np)^{-4} + 6p^4n^{-5} + O(n^{-6}),
\end{aligned}$$

where the third equality is obtained from integration by parts, the fourth equality is obtained from interchange of expectation & integral, the fifth equality is obtained from the fact that  $Z \sim \text{Binomial}(n-4, p)$ , and the sixth equality is obtained from integration by parts.

From these lower and upper bounds, we have  $\gamma_2 = (np)^{-4} + O(n^{-5})$ , and part (i) immediately follows.

Similarly we can show that  $\gamma_1 = (np)^{-4} + O(n^{-5})$  and  $\gamma_3 = (np)^{-4} + O(n^{-5})$ . Thus, part (ii) also immediately follows.

Since  $\mu^2 = \left(\frac{1-q^n-npq^{n-1}}{n(n-1)}\right)^2 = n^{-4}[1 + O(n^{-1})]$ , part (iii) also follows immediately from these bounds. □

From the above Lemma and Equation (5.27) we have

$$\text{Var} \left( \frac{\widehat{N}_1}{N_1} \right) = \frac{2q}{p} \frac{N_2}{N_1^2} [1 + O(n^{-1})] + \frac{(1+p)q}{p^2 N_1} [1 + O(n^{-1})] + O(n^{-1}).$$

□

# Proof of Proposition 1

**Proposition.** *When  $\mathcal{G}$  is generated by the Erdos-Renyi, preferential attachment, duplication, or geometric models, the corresponding convergence rate of  $\frac{N_2}{N_1^2}$  is as following:*

- *ER model:  $O(n^{-1})$*
- *Geometric model:  $O(n^{-1})$*
- *Preferential attachment model:  $O(\frac{\log(n)}{n})$*
- *Partial duplication model: let  $\beta$  be the approximated exponent of the power-law degree distribution of  $G$ , we have*

- $\beta = 2: O(\frac{1}{(\log(n))^2})$
- $2 < \beta < 3: O(\frac{1}{n^{\beta-2}})$
- $\beta = 3: O(\frac{\log(n)}{n})$
- $\beta > 3: O(n^{-1})$ .

*Thus,  $\text{Var}(\widehat{N}_1/N_1) \rightarrow 0$  as  $n \rightarrow \infty$ .*

## Proof for the ER model

As described in the introduction, for the ER model, we first start with  $n$  singleton nodes and then for each pair of nodes, a link is placed independently with some probability  $\rho$ .  $\rho$  is also referred to as the link density of the network  $\mathcal{G}$  generated from the ER model. As a result, the number of links in  $\mathcal{G}$  is a Binomial random variable:  $N_1 \sim \text{Binomial}(\binom{n}{2}, \rho)$ . Hence, when  $n$  tends to infinity, we approximate  $N_1$  by its expectation as follows:

$$N_1 \simeq \binom{n}{2} \rho. \tag{5.28}$$



On the other hand, we can obtain an upper bound for  $N_2$  as follows:

$$N_2 \leq 3 \binom{n}{3}, \quad (5.29)$$

where  $\binom{n}{3}$  is the total number of combinations of three nodes chosen from  $n$ , and for every three nodes we can have at most three pairs of links that share exactly one common neighbor.

In summary, we have

$$\frac{N_2}{N_1^2} \leq \frac{3 \binom{n}{3}}{[\binom{n}{2} \rho]^2} \simeq \frac{n(n-1)(n-2)/2}{[n(n-1)\rho/2]^2} = \frac{2}{\rho^2} \frac{n-2}{n(n-1)} = O(n^{-1}). \quad (5.30)$$

## Proof for the geometric model

Next, we proceed to the geometric model. Similar to the ER model, we also start with  $n$  singleton nodes. However, the nodes now are placed uniformly at random in the three-dimension unit cube. Then for each pair of nodes, we record the Euclidean distance between them, and place a link if the distance is less than some given threshold  $\delta, 0 < \delta < 1, .$

For any node  $i, 1 \leq i \leq n$ , let  $k_i$  denote its degree, that is, the number of neighbors of node  $i$ , in the network  $\mathcal{G}$  generated from the geometric model. For any node  $j \neq i, 1 \leq j \leq n$ , node  $j$  is connected to node  $i$  if and only if the distance between them is less than  $\delta$ , as described in the model setting. In other words, node  $j$  must fall in the sphere centered at node  $i$  with radius  $\delta$ . On the other hand, we know that node  $j$  is placed uniformly at random in the three-dimension unit cube. Hence, the probability that node  $j$  is connected to node  $i$  is equal to the volume of the sphere, that is,  $\frac{4}{3}\pi\delta^3$ . Note that here we do not take into account the effect on the boundaries of the cube.

Since the nodes are placed independently and identically, we have  $k_i \simeq \text{Binomial}(n-$

$1, \frac{4}{3}\pi\delta^3$ ), for any  $1 \leq i \leq n$ . Thus, we can have the following approximation:

$$\frac{\sum_{i=1}^n k_i}{n} = \bar{k} \simeq E(k_i) \simeq (n-1)\frac{4}{3}\pi\delta^3. \quad (5.31)$$

On the other hand, since the number of links is actually half of the sum of node degree, we further have:

$$N_1 = \frac{\sum_{i=1}^n k_i}{2} = \frac{n}{2} \times \frac{\sum_{i=1}^n k_i}{n} \simeq \frac{n(n-1)}{2} \frac{4}{3}\pi\delta^3. \quad (5.32)$$

Then by noting that the upper bound of  $N_2$  is  $3\binom{n}{3}$ , we finally obtain:

$$\frac{N_2}{N_1^2} \leq \frac{2}{(\frac{4}{3}\pi\delta^3)^2} \times \frac{n-2}{n(n-1)} = O(n^{-1}). \quad (5.33)$$

Overall, the case of the geometric model is actually quite similar to the ER model. The only difference is the link density, which is  $\rho$  for the ER model, and  $\frac{4}{3}\pi\delta^3$  for the geometric model. Hence, not surprisingly, we also observe from the simulation results in chapter 2 that there is no significant difference in the accuracy of the estimators for these two models.

## Proof for the preferential attachment model

Recall that a random network  $\mathcal{G}$  is generated from the preferential attachment model as follows:

- First, an initial small random network  $\mathcal{G}_0$  is generated from the ER model.
- At each subsequent iteration, a new node with  $l$  incident links is added to the current network. Neighbors of the newly added node are chosen with probabilities proportional to their current degrees.

Then the number of links in  $\mathcal{G}$ ,  $N_1$ , can be approximated as

$$N_1 \simeq l \times n, \quad (5.34)$$

since the number of nodes and links in the initial network  $\mathcal{G}_0$  are neglectable.

Next, let  $n_k$  be the number of nodes of degree  $k$ , and  $p_k = n_k/n$ ,  $1 \leq k \leq n$ . It has been shown by Barabasi & Albert in [18] that networks generated by the preferential attachment model exhibit the scale free structure. In particular, the node degree follows the power-law degree distribution with exponent  $\beta = 3$ :

$$p_k \simeq Ck^{-3}, \quad (5.35)$$

where  $C$  is a normalizing constant.

On the other hand,  $N_2$  can also be written as a function of the node degree as follows:

$$\begin{aligned} N_2 &= \sum_{k=1}^n \binom{k}{2} n_k \\ &= n \sum_{k=1}^n \binom{k}{2} p_k \\ &\simeq n \sum_{k=1}^n \binom{k}{2} \frac{C}{k^3} \\ &= \frac{C}{2} n \sum_{k=1}^n \left( \frac{1}{k} - \frac{1}{k^2} \right). \end{aligned} \quad (5.36)$$

Using the fact that

$$\log(n+1) < \sum_{k=1}^n \frac{1}{k} < \log(n) + 1, \quad (5.37)$$

and

$$\sum_{k=1}^n \frac{1}{k^2} \rightarrow \frac{\pi^2}{6}, \quad (5.38)$$

as  $n \rightarrow \infty$ , we have the following approximation:

$$\frac{N_2}{N_1^2} \simeq \frac{\frac{C}{2}n(\log(n) - \pi^2/6)}{(ln)^2} = O\left(\frac{\log(n)}{n}\right). \quad (5.39)$$

It can be seen that the convergence rate of the fraction  $\frac{N_2}{N_1^2}$ , and hence the variation of the estimator  $\widehat{N}_1$ , in the preferential attachment model is slower than that in the ER and geometric models. This is because the former has the scale-free structure, whereas the latter have symmetric structure.

## Proof for the duplication model

Finally, a random network  $\mathcal{G}$  is generated from the duplication model as follows:

- An initial small random network  $\mathcal{G}_0$  is generated from the ER model.
- At each iteration, an existing node  $u$  is chosen uniformly at random. A new node  $u'$  is duplicated from  $u$ , that is,  $u'$  is connected to each neighbor of  $u$  with probability  $p_{dup}$ . The new node  $u'$  is also connected to the duplicated node  $u$ .

It has been shown by Chung *et al.* in [64] that

$$N_1 \simeq \begin{cases} \frac{n}{1-2p_{dup}} + C_0 n^{2p_{dup}} & \text{if } p_{dup} \neq \frac{1}{2}, \\ n \log(n) + C_0 n & \text{if } p_{dup} = \frac{1}{2}, \end{cases} \quad (5.40)$$

where constant  $C_0$  is determined by the initial network  $\mathcal{G}_0$ . Furthermore, the degree distribution follows a power law, that is,  $p_k \simeq Ck^{-\beta}$ , where the exponent  $\beta$  satisfying

the following equation:

$$1 + p_{dup} = p_{dup}\beta + p_{dup}^{\beta-1}. \quad (5.41)$$

Recall that  $n_k$  is the number of nodes of degree  $k$ ,  $p_k = n_k/n$ ,  $1 \leq k \leq n$ . We have

$$N_2 = \sum_{k=1}^n \binom{k}{2} n_k \simeq \sum_{k=1}^n \binom{k}{2} n p_k \simeq n \sum_{k=1}^n \binom{k}{2} C k^{-\beta} = \frac{C}{2} n \sum_{k=1}^n \left( \frac{1}{k^{\beta-2}} - \frac{1}{k^{\beta-1}} \right). \quad (5.42)$$

For  $p_{dup} = \frac{1}{2}$ , from Equation 5.41 we have  $\beta = 2$  and

$$N_2 \simeq \frac{C}{2} n \left( n - \sum_{k=1}^n \frac{1}{k} \right) < \frac{C}{2} n (n - \log(n+1)), \quad (5.43)$$

$$N_1 \simeq n \log(n) + C_0 n, \quad (5.44)$$

$$\frac{N_2}{N_1^2} < \frac{\frac{C}{2} n (n - \log(n+1))}{(n \log(n) + C_0 n)^2} = O\left(\frac{1}{(\log(n))^2}\right). \quad (5.45)$$

$$(5.46)$$

For  $p_{dup} < \frac{1}{2}$ , from Equation 5.41 we have  $\beta > 2$ . There are three possible situations as follows

- $\beta > 3$ :

– both  $\sum_{k=1}^n \frac{1}{k^{\beta-2}}$  and  $\sum_{k=1}^n \frac{1}{k^{\beta-1}}$  converge to  $O(1)$ ,

–  $N_2 \simeq \frac{C}{2} n \sum_{k=1}^n \left( \frac{1}{k^{\beta-2}} - \frac{1}{k^{\beta-1}} \right) = O(n)$ ,

–  $N_1 \simeq \frac{n}{1-2p_{dup}} + C_0 n^{2p_{dup}} = O(n)$ ,

– Hence,  $\frac{N_2}{N_1^2} = O(n^{-1})$ .

- $\beta = 3$  (similar to the case of the preferential attachment model):

–  $N_2 \simeq \frac{C}{2} n \sum_{k=1}^n \left( \frac{1}{k} - \frac{1}{k^2} \right) = O(n \log(n))$ ,

–  $N_1 \simeq \frac{n}{1-2p_{dup}} + C_0 n^{2p_{dup}} = O(n)$ ,

- Hence  $\frac{N_2}{N_1^2} = O\left(\frac{\log(n)}{n}\right)$ .
- $2 < \beta < 3$ :
  - $\sum_{k=1}^n \frac{1}{k^{\beta-1}}$  converges to  $O(1)$ ,
  - $\sum_{k=1}^n \frac{1}{k^{\beta-2}} < \int_0^n \frac{1}{x^{\beta-2}} dx = \frac{n^{3-\beta}}{3-\beta}$ ,
  - $N_2 < \frac{C}{2}n\left(\frac{n^{3-\beta}}{3-\beta} + \sum_{k=1}^n \frac{1}{k^{\beta-1}}\right) = O(n^{4-\beta})$ ,
  - $N_1 \simeq \frac{n}{1-2p_{dup}} + C_0n^{2p_{dup}} = O(n)$ ,
  - Hence,  $\frac{N_2}{N_1^2} = O\left(\frac{1}{n^{\beta-2}}\right)$ .

Thus, we have shown that the duplication model has a wide range of convergence rate, depending on the exponent of the power-law degree distribution  $\beta$ . When  $\beta = 3$ , the duplication model has the same convergence rate as the preferential attachment model since both models have the same exponent. When  $\beta > 3$ , the duplication model has the same convergence rate as the ER model and the geometric model. In general, the convergence is faster when the exponent is higher, because there will be more nodes with low degree and less nodes with high degree, which subsequently make the variation become smaller.

The simulation results shown in Figures 5.19, 5.20, 5.21, 5.22, 3.3, 3.4, 5.23, 5.24, confirm our theoretical results obtained in Theorem 2 and Proposition 1.

# Supplementary Figures

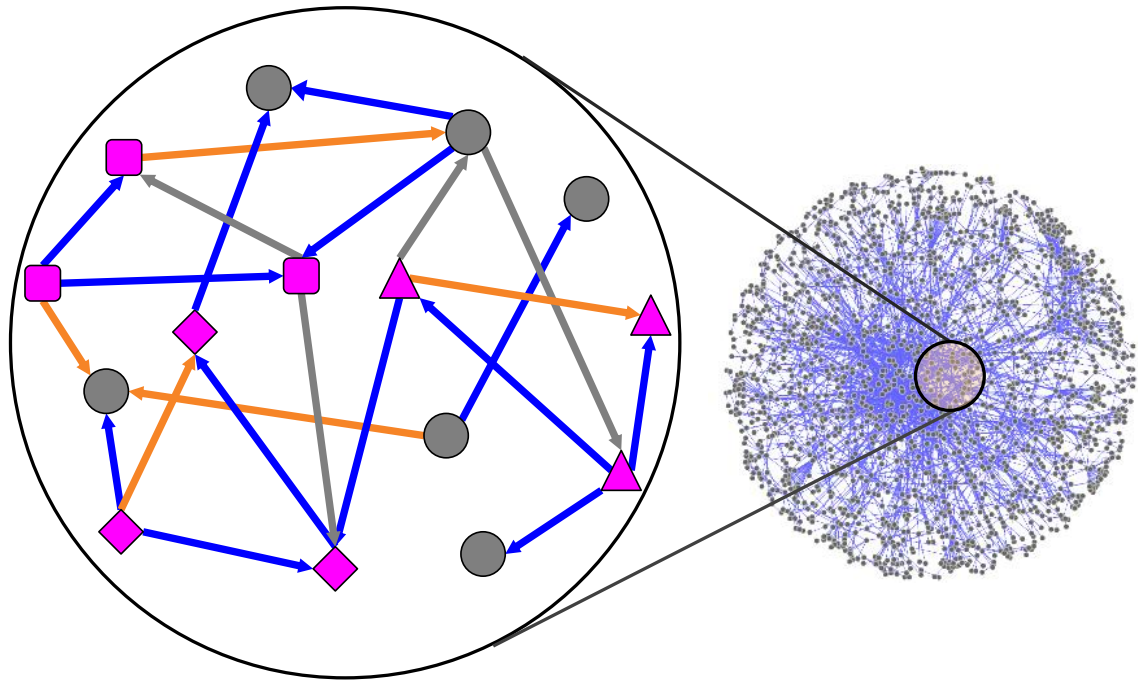


Figure 5.1: Cover art of the study.

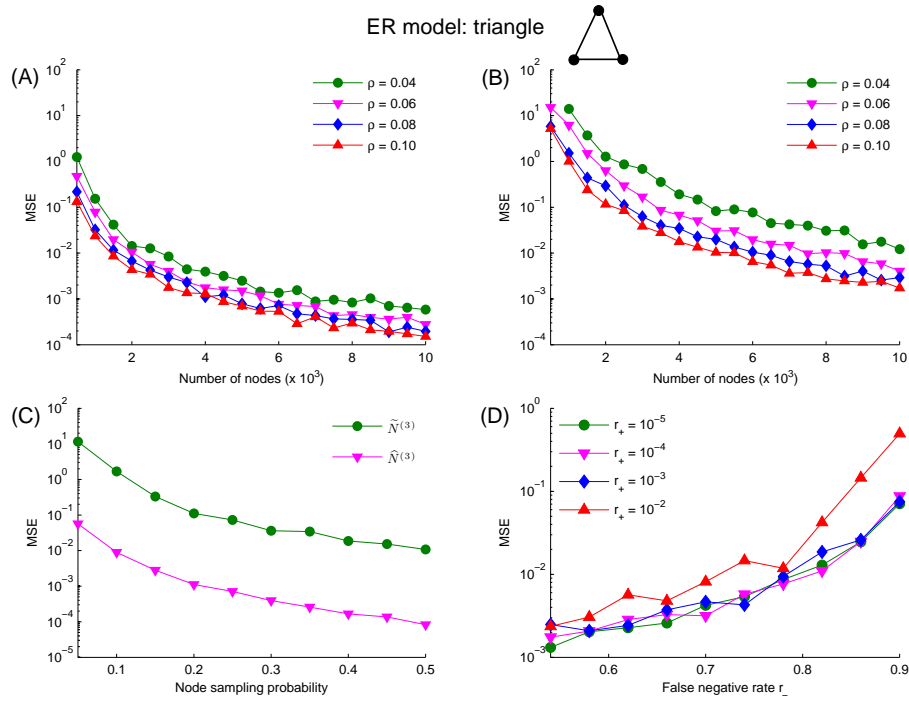


Figure 5.2: Performance of the estimators  $\hat{N}^{(3)}$  and  $\tilde{N}^{(3)}$  for estimating the number of triangles in networks generated from the ER model.

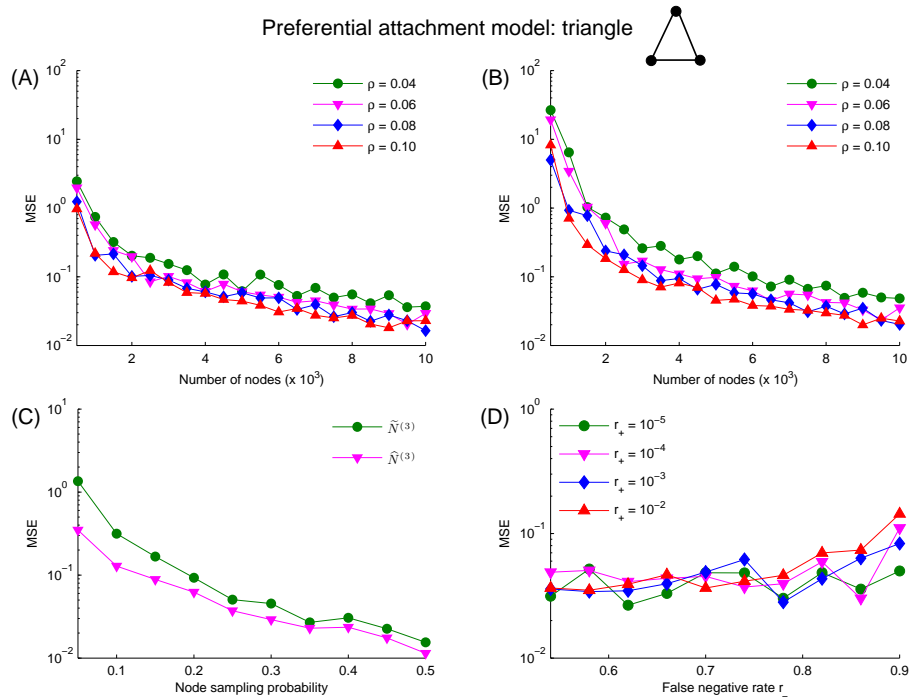


Figure 5.3: Performance of the estimators  $\hat{N}^{(3)}$  and  $\tilde{N}^{(3)}$  for estimating the number of triangles in networks generated from the preferential attachment model.



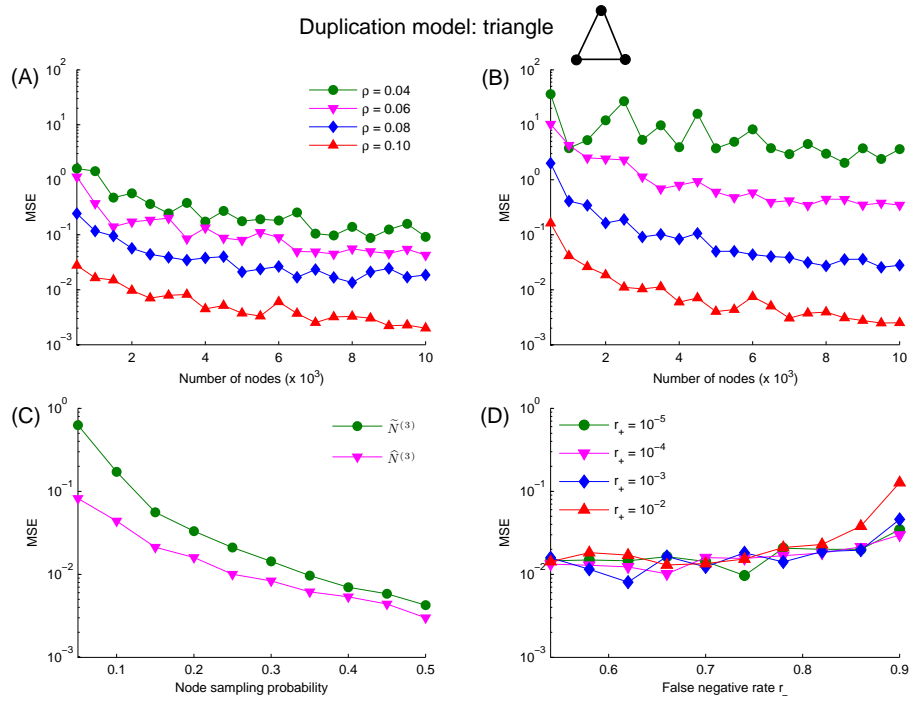


Figure 5.4: Performance of the estimators  $\hat{N}^{(3)}$  and  $\tilde{N}^{(3)}$  for estimating the number of triangles in networks generated from the duplication model.

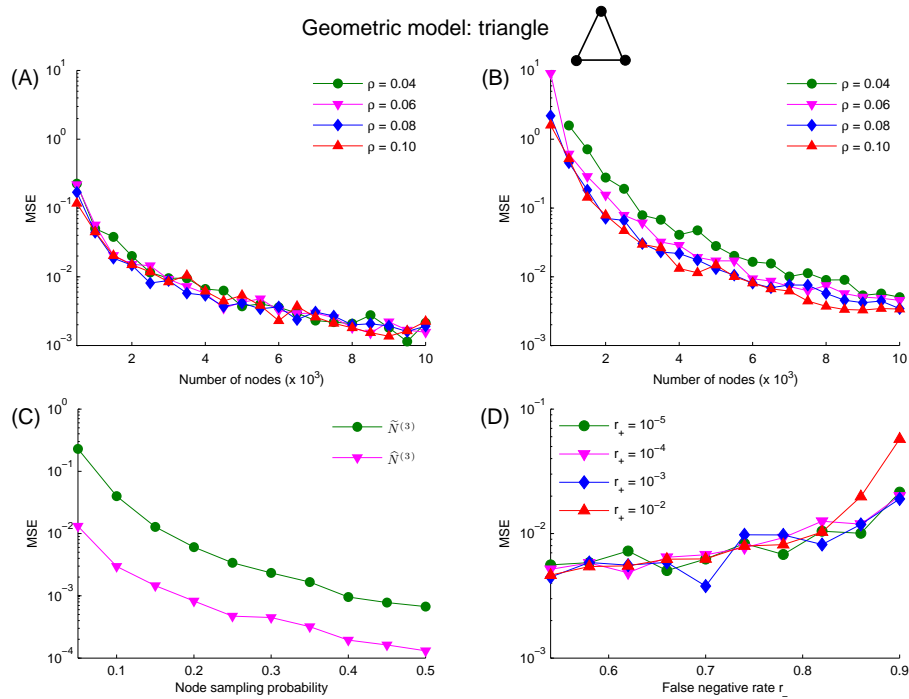


Figure 5.5: Performance of the estimators  $\hat{N}^{(3)}$  and  $\tilde{N}^{(3)}$  for estimating the number of triangles in networks generated from the geometric model.

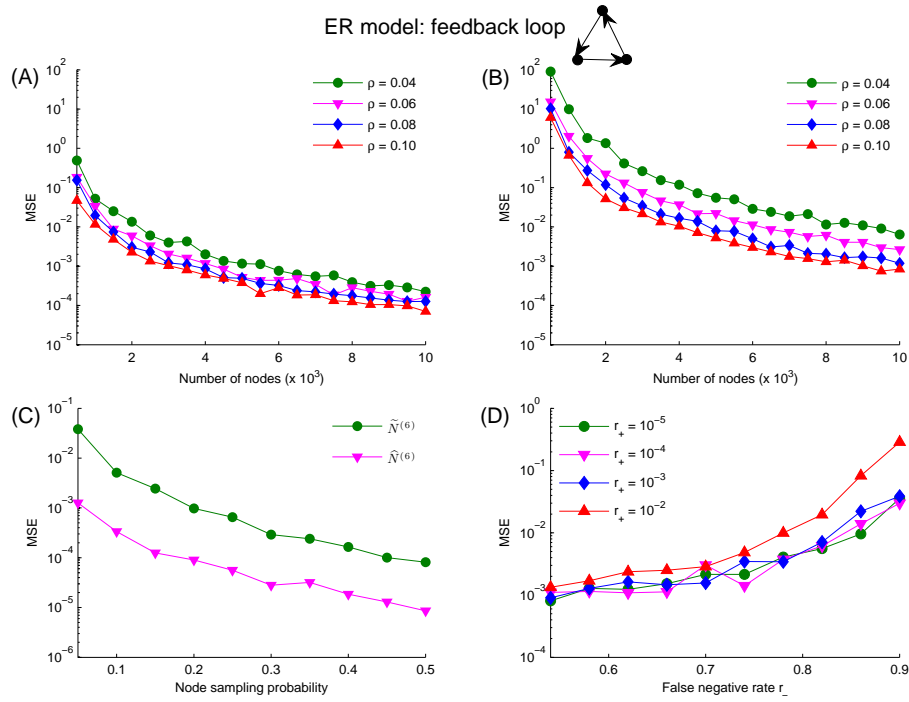


Figure 5.6: Performance of the estimators  $\hat{N}^{(8)}$  and  $\tilde{N}^{(8)}$  for estimating the number of three-node feedback loops in networks generated from the ER model.

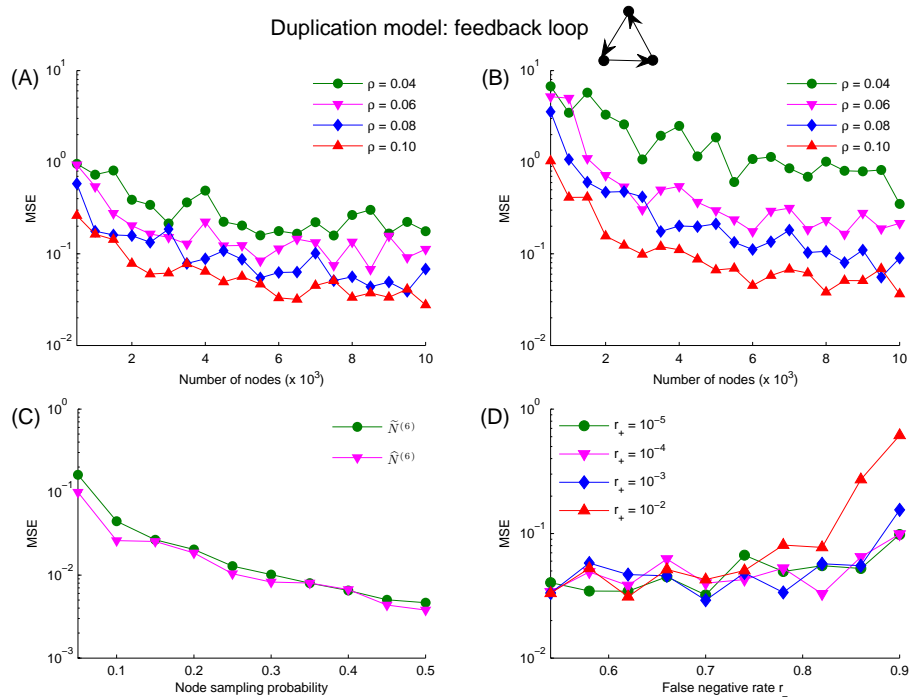


Figure 5.7: Performance of the estimators  $\hat{N}^{(8)}$  and  $\tilde{N}^{(8)}$  for estimating the number of three-node feedback loops in networks generated from the duplication model.

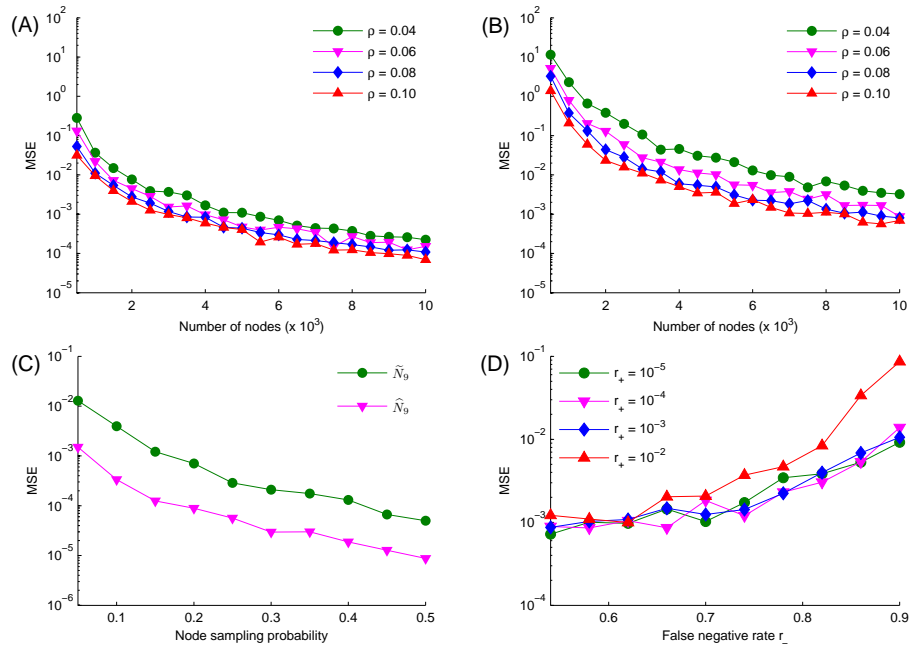


Figure 5.8: Performance of the estimators  $\hat{N}^{(9)}$  and  $\tilde{N}^{(9)}$  for estimating the number of feed-forward loops in networks generated from the ER model.

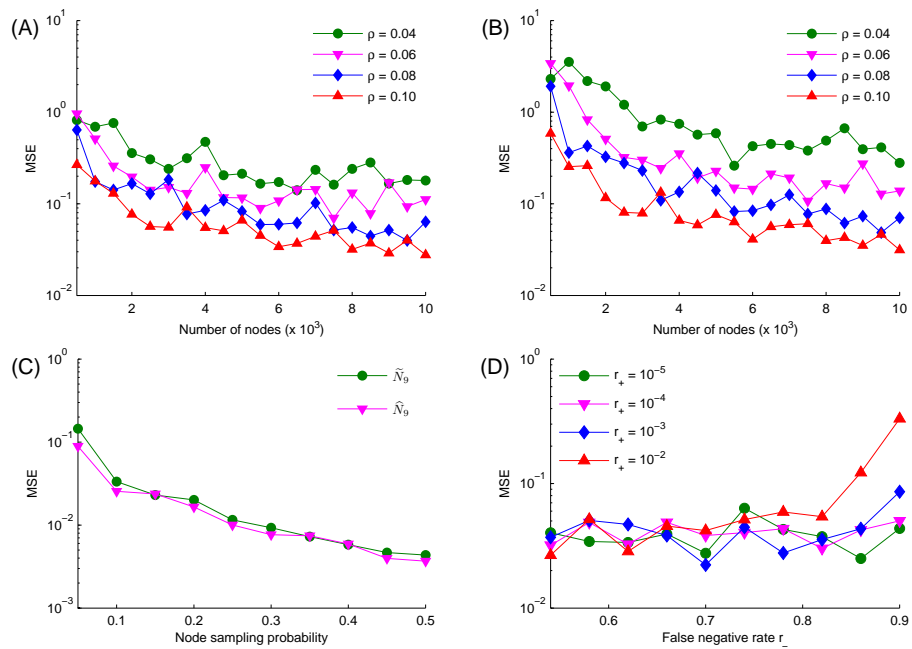


Figure 5.9: Performance of the estimators  $\hat{N}^{(9)}$  and  $\tilde{N}^{(9)}$  for estimating the number of feed-forward loops in networks generated from the duplication model.

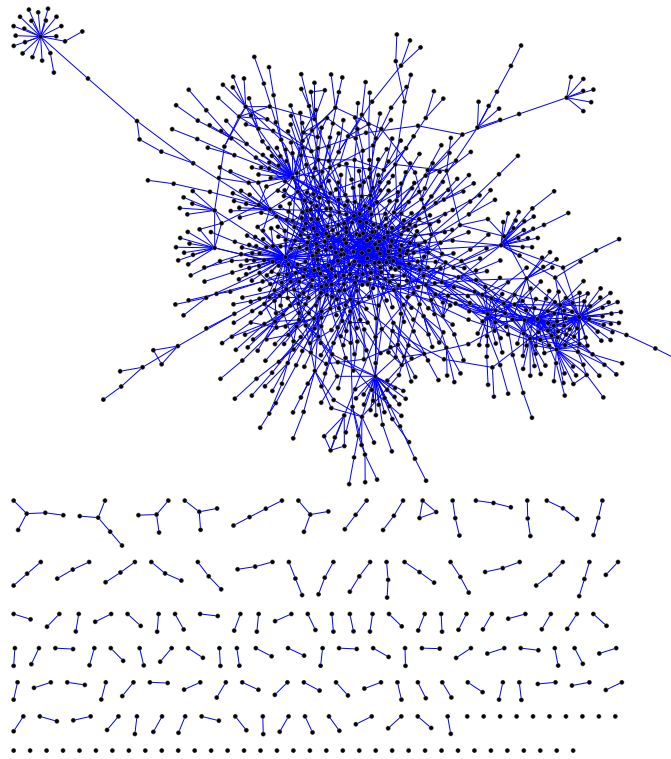


Figure 5.10: Observed PPI subnetwork of *S. cerevisiae* from Y2H experiment.

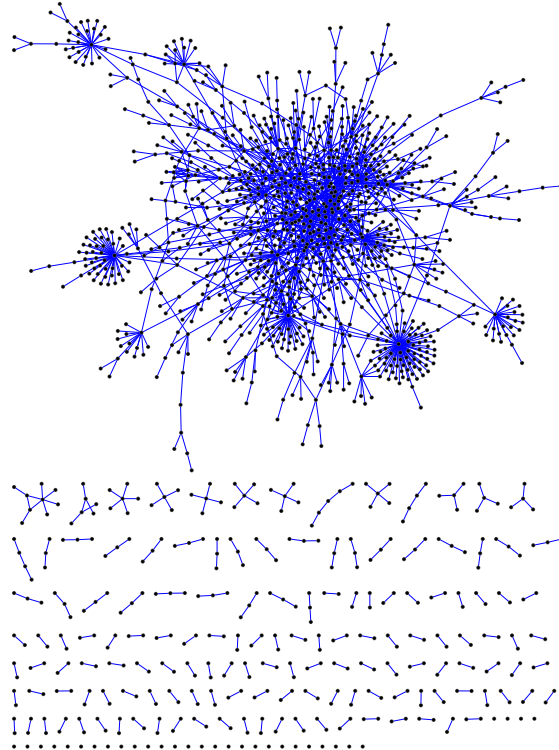


Figure 5.11: Observed PPI subnetwork of *C. elegans* from Y2H experiment.

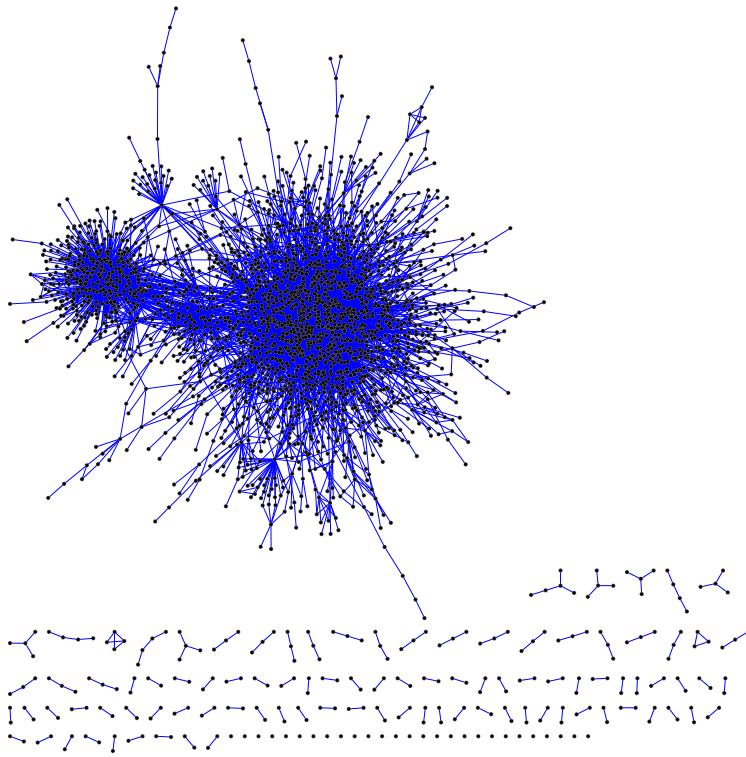
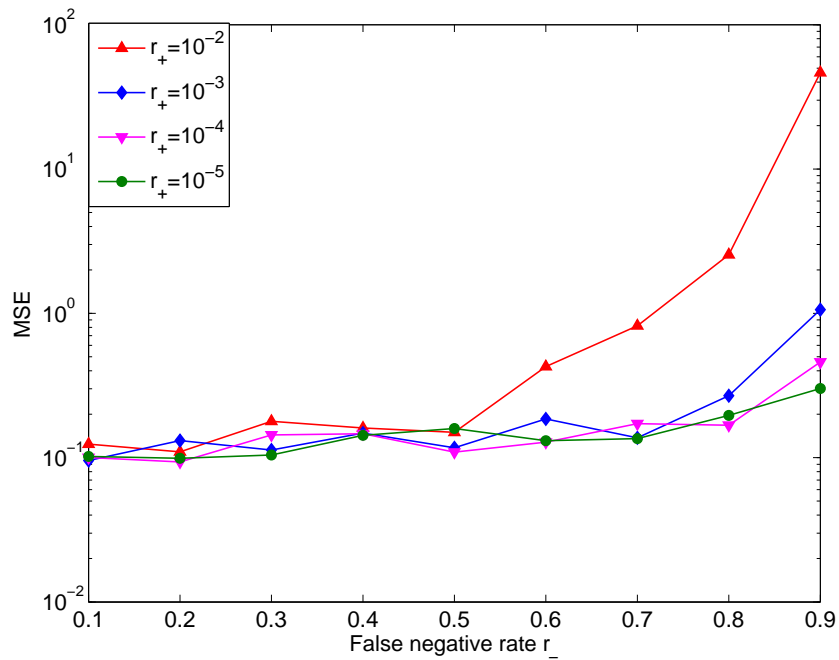


Figure 5.12: Observed PPI subnetwork of *A. thaliana* from Y2H experiment.



h]

Figure 5.13: Performance of the estimator  $\tilde{N}_{\mathcal{M}}$  for motif  $u_2$  with respect to false positive and false negative rates in the PPI network of *S. cerevisiae*.

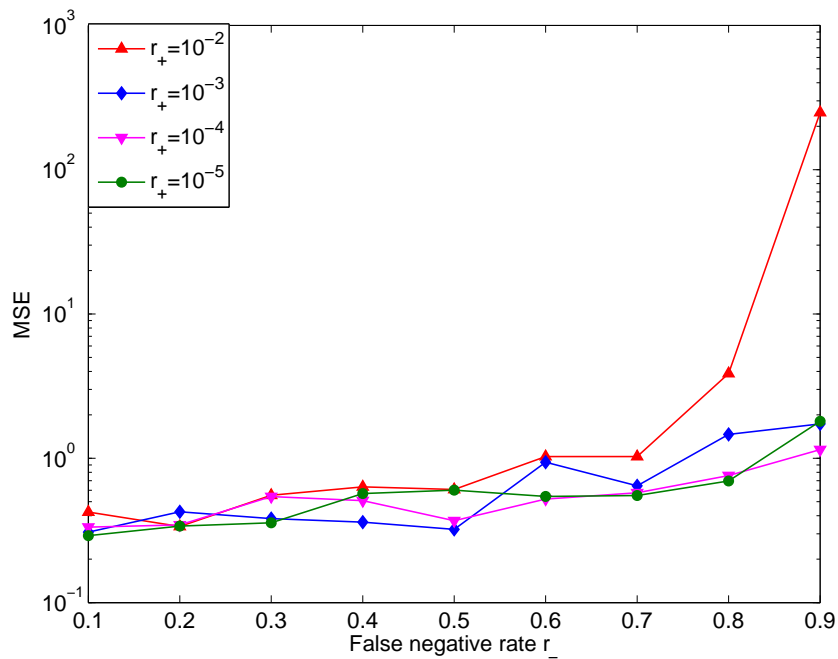
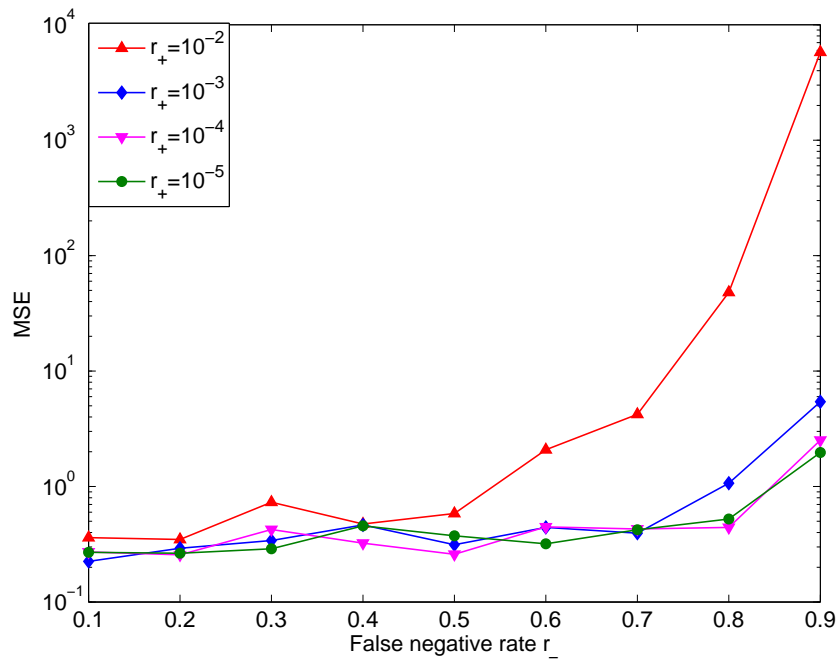


Figure 5.14: Performance of the estimator  $\tilde{N}_{\mathcal{M}}$  for motif  $u_4$  with respect to false positive and false negative rates in the PPI network of *S. cerevisiae*.



h]

Figure 5.15: Performance of the estimator  $\tilde{N}_{\mathcal{M}}$  for motif  $u_5$  with respect to false positive and false negative rates in the PPI network of *S. cerevisiae*.

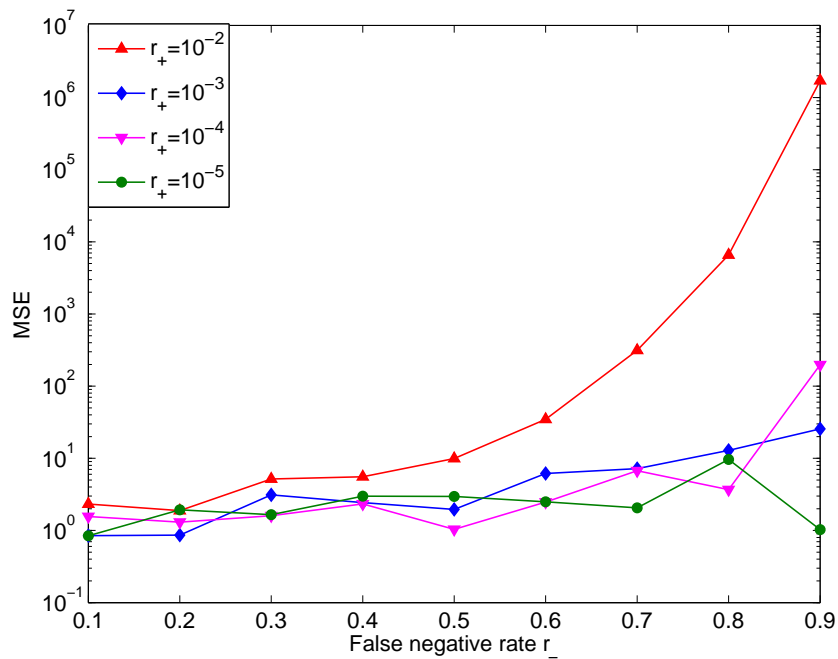
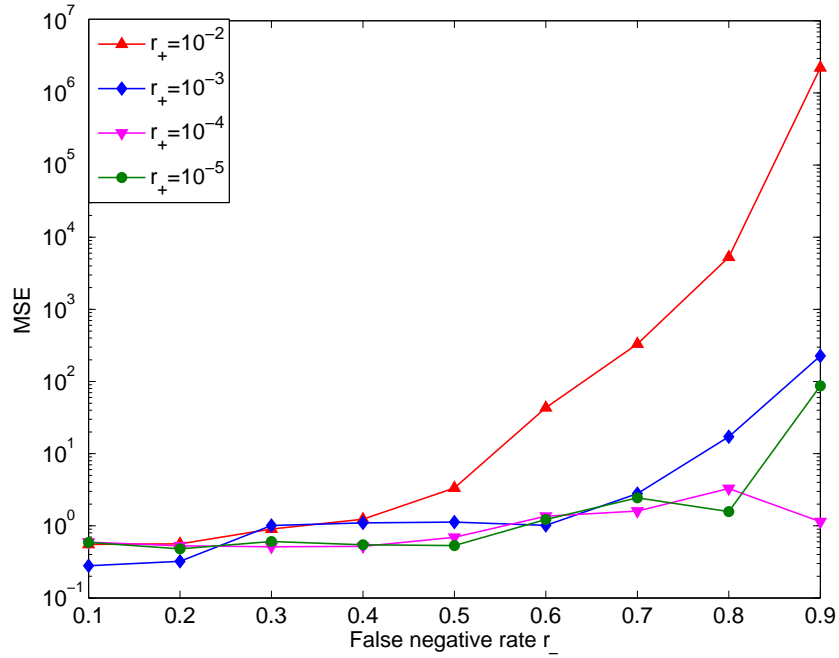


Figure 5.16: Performance of the estimator  $\tilde{N}_{\mathcal{M}}$  for motif  $u_6$  with respect to false positive and false negative rates in the PPI network of *S. cerevisiae*.



h]

Figure 5.17: Performance of the estimator  $\tilde{N}_{\mathcal{M}}$  for motif  $u_7$  with respect to false positive and false negative rates in the PPI network of *S. cerevisiae*.

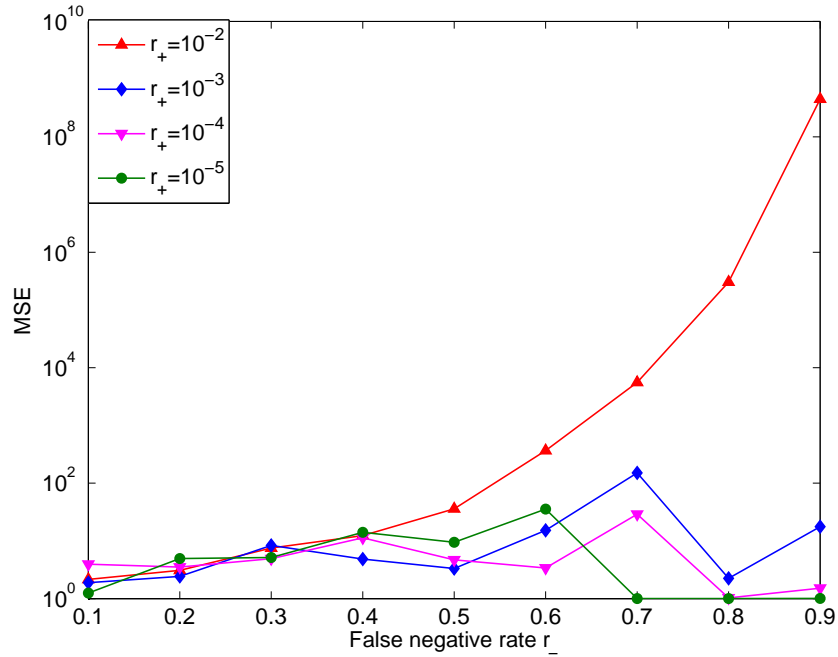
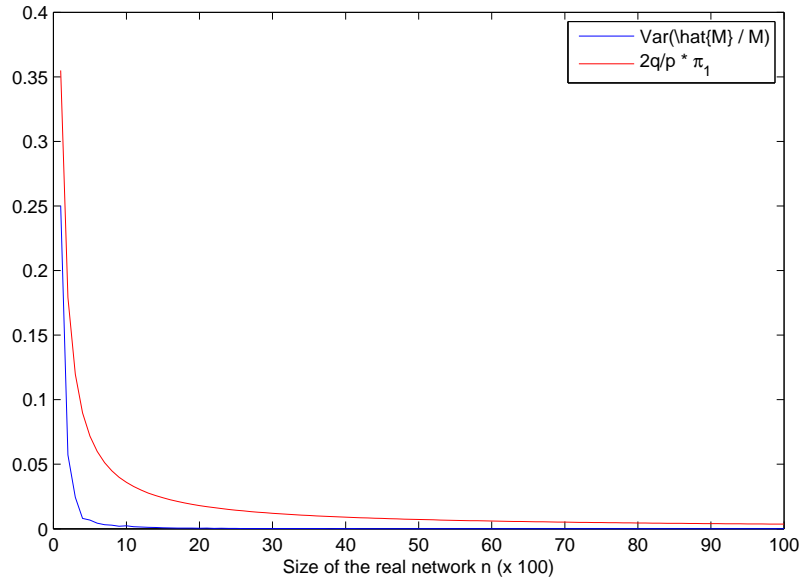


Figure 5.18: Performance of the estimator  $\tilde{N}_{\mathcal{M}}$  for motif  $u_8$  with respect to false positive and false negative rates in the PPI network of *S. cerevisiae*.





h]

Figure 5.19: The convergence rate of  $\text{Var}\left(\frac{\hat{N}_1}{N_1}\right)$  in Equation 2.11 and the dominated term  $\frac{N_2}{N_1^2}$  for the ER model.

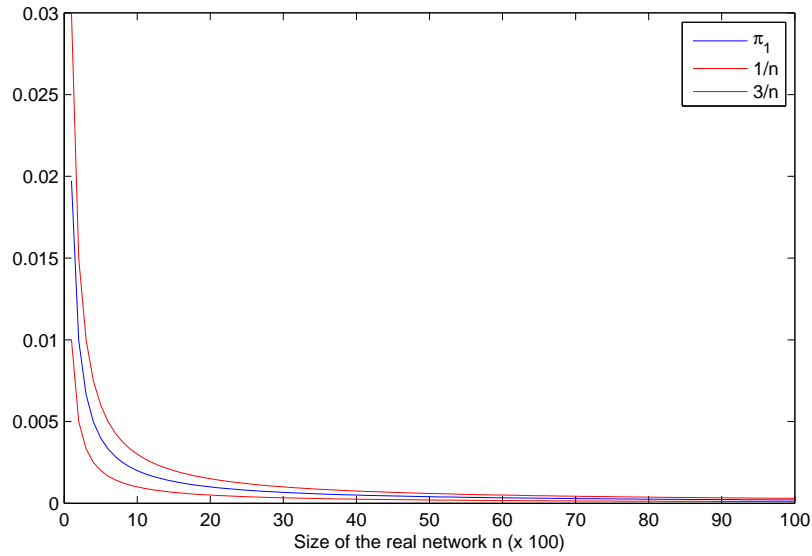
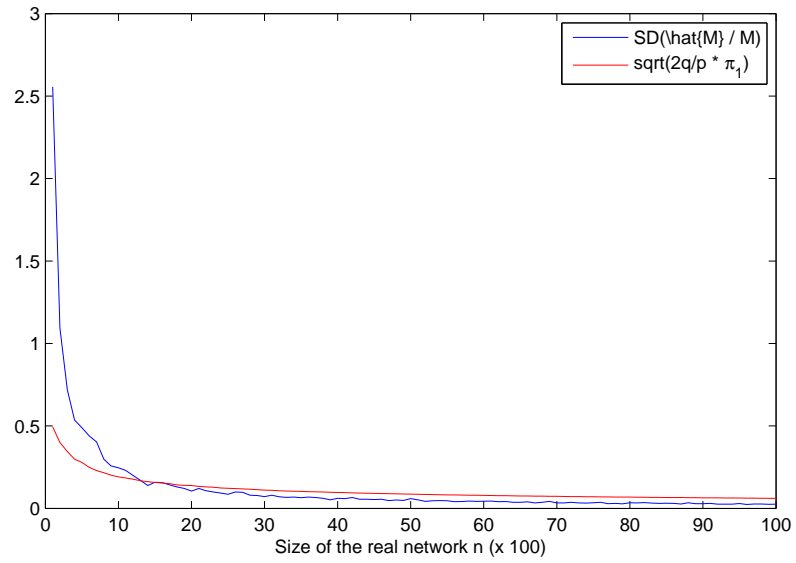


Figure 5.20: The convergence rate of  $\frac{N_2}{N_1^2}$  is bounded as shown in Proposition 1 for the ER model.



h]

Figure 5.21: The convergence rate of  $\text{Var} \left( \frac{\hat{N}_1}{N_1} \right)$  in Equation 2.11 and the dominated term  $\frac{N_2}{N_1^2}$  for the geometric model.

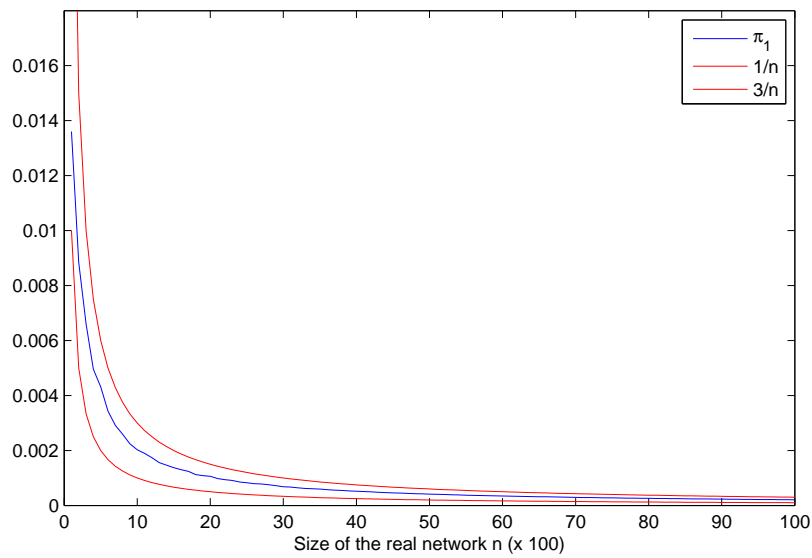
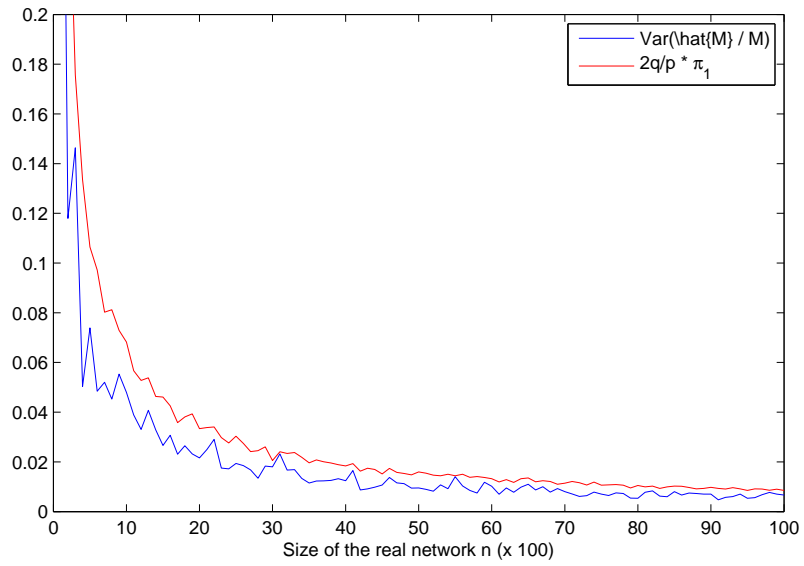


Figure 5.22: The convergence rate of  $\frac{N_2}{N_1^2}$  is bounded as shown in Proposition 1 for the geometric model.



h]

Figure 5.23: The convergence rate of  $\text{Var}\left(\frac{\hat{N}_1}{N_1}\right)$  in Equation 2.11 and the dominated term  $\frac{N_2}{N_1^2}$  for the duplication model,  $\beta = 2$ .

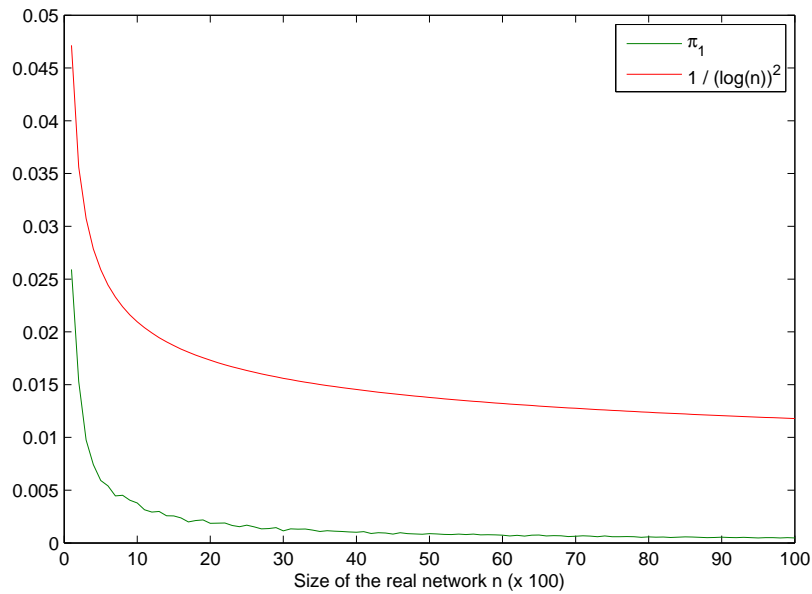


Figure 5.24: The convergence rate of  $\frac{N_2}{N_1^2}$  is bounded as shown in Proposition 1 for the duplication model,  $\beta = 2$ .

53 GO terms and their odds ratio, p-value

GO term	X_{GO}	Triangle_Ratio	n_{GO}	Triplet_Ratio	Odds ratio	p-value
GO:0000131	1	0.008928571	53	6.51048E-07	13714.15841	2.6346E-09
GO:0000398	2	0.017857143	84	2.6481E-06	6743.375068	4.23017E-12
GO:0000422	1	0.008928571	29	1.01551E-07	87922.24275	6.41081E-11
GO:0000956	2	0.017857143	14	1.01162E-08	1765208.104	5.21805E-15
GO:0001302	1	0.008928571	38	2.34451E-07	38082.9629	3.41664E-10
GO:0003674	2	0.017857143	1990	0.036447493	0.48994159	0.779097488
GO:0003723	5	0.044642857	434	0.000376032	118.7209007	6.54243E-12
GO:0003824	1	0.008928571	405	0.000305425	29.23325108	0.00056703
GO:0004679	1	0.008928571	5	2.77917E-10	32126787.5	0
GO:0005515	1	0.008928571	55	7.29115E-07	12245.77378	3.3043E-09
GO:0005575	7	0.0625	737	0.001846699	33.84417639	5.42199E-11
GO:0005634	24	0.214285714	2029	0.038633777	5.546589771	1.06337E-12
GO:0005688	2	0.017857143	9	2.3345E-09	7649235.119	3.10862E-15
GO:0005730	2	0.017857143	333	0.000169502	105.3508965	1.09468E-06
GO:0005732	2	0.017857143	9	2.3345E-09	7649235.119	3.10862E-15
GO:0005737	17	0.151785714	2091	0.042286543	3.589456657	1.1632E-06
GO:0005768	4	0.035714286	114	6.6829E-06	5344.132594	4.10783E-15
GO:0005777	1	0.008928571	67	1.33136E-06	6706.353721	1.10169E-08
GO:0005783	3	0.026785714	420	0.000340724	78.61405725	8.12797E-08
GO:0005789	2	0.017857143	320	0.00015036	118.7628922	7.65314E-07
GO:0005829	1	0.008928571	483	0.000518684	17.21390119	0.001610028
GO:0005933	1	0.008928571	47	4.50642E-07	19813.00493	1.26229E-09
GO:0006351	1	0.008928571	527	0.000674092	13.24533321	0.002688696
GO:0006355	1	0.008928571	505	0.000592998	15.05666027	0.002093047
GO:0006364	4	0.035714286	196	3.43444E-05	1039.886954	9.99201E-15
GO:0006397	3	0.026785714	171	2.2756E-05	1177.085661	1.66822E-12
GO:0006468	1	0.008928571	133	1.06527E-05	838.1498724	7.04843E-07
GO:0006810	4	0.035714286	832	0.002658067	13.43618887	1.4052E-05
GO:0006914	1	0.008928571	51	5.78762E-07	15427.02881	2.08206E-09
GO:0007118	1	0.008928571	17	1.88983E-08	472452.7574	2.21989E-12
GO:0008033	2	0.017857143	80	2.28336E-06	7820.542235	2.71516E-12
GO:0008134	1	0.008928571	11	4.58563E-09	1947078.03	1.29341E-13
GO:0008152	1	0.008928571	338	0.000177276	50.36544466	0.000192827
GO:0008380	2	0.017857143	112	6.33428E-06	2819.128422	5.79011E-11
GO:0008614	12	0.107142857	6	5.55833E-10	192760725	5.55112E-15
GO:0008615	12	0.107142857	8	1.55633E-09	68843116.07	0
GO:0009228	12	0.107142857	16	1.55633E-08	6884311.607	4.44089E-15
GO:0010008	2	0.017857143	58	8.5754E-07	20823.68907	1.4122E-13
GO:0015031	4	0.035714286	382	0.000256174	139.4140704	1.44661E-10
GO:0016020	11	0.098214286	1685	0.022120225	4.44002206	7.8109E-06
GO:0016021	5	0.044642857	1304	0.010246992	4.356679018	0.001097693
GO:0016787	1	0.008928571	633	0.001169267	7.636043672	0.007803397
GO:0030529	5	0.044642857	268	8.8164E-05	506.3617165	1.9984E-15
GO:0031120	2	0.017857143	8	1.55633E-09	11473852.68	0

GO:0031588	1	0.008928571	5	2.77917E-10	32126787.5	0
GO:0032258	1	0.008928571	37	2.15941E-07	41347.21686	2.89849E-10
GO:0034727	1	0.008928571	33	1.51631E-07	58883.40817	1.42916E-10
GO:0042254	2	0.017857143	172	2.31599E-05	771.0367317	2.82599E-09
GO:0042823	1	0.008928571	8	1.55633E-09	5736926.339	1.29896E-14
GO:0043162	3	0.026785714	15	1.26452E-08	2118249.725	0
GO:0043332	1	0.008928571	105	5.20983E-06	1713.794276	1.68652E-07
GO:0046020	1	0.008928571	9	2.3345E-09	3824617.56	3.69704E-14
GO:0046540	2	0.017857143	33	1.51631E-07	117766.8163	0

# Bibliography

- [1] Watson, J. D., Crick, F. H. C.: Molecular structure of nucleic acids. *Nature* 171, 737-738 (1953)
- [2] International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945 (2004)
- [3] Economic impact of the Human Genome Project. Battelle Technology Partnership Practice (2011)
- [4] Hartwell, L. H., Hopfield, J. J., Leibler, S., Murray, A. W.: From molecular to modular cell biology. *Nature* 402, C47-C52 (1999)
- [5] Hasty, J., McMillen, D., Collins, J. J.: Engineered gene circuits. *Nature* 420, 224-230 (2002)
- [6] Barabasi, A.-L., Oltvai, Z. N.: Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* 5, 101-113 (2004)
- [7] Vidal, M., Cusick, M. E., Barabasi, A.-L.: Interactome networks and human disease. *Nature Reviews Genetics* 12, 56-68 (2011)
- [8] Barabasi, A.-L., Gulbahce, N., Loscalzo J.: Network medicine: a network-based approach to human disease. *Cell* 144, 986-998 (2011)

- [9] Cherry et al.: Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*. 40 (Database issue), D700-D705 (2012)
- [10] Harris et al.: WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research* 38 (Database issue), D463-D467 (2010).
- [11] McQuilton, P., St. Pierre, S. E., Thurmond, J., and the FlyBase Consortium: FlyBase 101 the basics of navigating FlyBase. *Nucleic Acids Research* 40 (Database issue), D706-D714 (2012).
- [12] Swarbreck et al.: The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research* 36 (Database issue), D1009-D1014 (2008)
- [13] Wagner A.: Robustness against mutations in genetic networks of yeast. *Nature Genetics* 24, 355-361 (2000)
- [14] Sharan, R., Ideker T.: Modeling cellular machinery through biological network comparison. *Nature Biotechnology* 24, 427-433 (2006)
- [15] Mason, O., Verwoerd, M.: Graph theory and networks in biology. *IET System Biology* 1, 89-119 (2007)
- [16] Albert, R., Barabasi, A.-L.: Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 47-97 (2002)
- [17] Strogatz, S. H.: Exploring complex networks. *Nature* 410, 268-276 (2001)
- [18] Barabasi, A.-L., Albert R.: Emergence of scaling in random networks. *Science* 286, 509-512 (1999)
- [19] Watts, D. J., Strogatz, S. H.: Collective dynamics of “small-world” networks. *Nature* 393, 440-442 (1998)

- [20] Ravasz et al.: Hierarchical organization of modularity in metabolic networks. *Science* 30, 1551-1555 (2002)
- [21] Seebacher, J., Gavin, A.-C.: Snapshot: protein-protein interaction networks. *Cell* 144, 1000-1000.e1 (2011)
- [22] Field, S., Song, O.: A novel genetic system to detect protein-protein interactions. *Nature* 340, 245-246 (1989)
- [23] Uetz *et al.*: A comprehensive analysis of proteinprotein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623-627 (2000)
- [24] Ito *et al.*: A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences* 98, 4569-4574 (2001)
- [25] Rual *et al.*: Towards a proteome-scale map of the human proteinprotein interaction network. *Nature* 437, 1173-1178 (2005)
- [26] Stelz *et al.*: A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957-968 (2005)
- [27] Venkatesan *et al.*: An empirical framework for binary interactome mapping. *Nature Methods* 6, 83-90 (2009)
- [28] Yu *et al.*: High-quality binary protein interaction map of the yeast interactome network. *Science* 322, 104-110 (2008)
- [29] Simonis *et al.*: Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nature Methods* 6, 47-54 (2009)
- [30] *Arabidopsis* Interactome Mapping Consortium: Evidence for network evolution in an *Arabidopsis* interactome map. *Science* 333, 601-607 (2011)



- [31] Rigaut *et al.*: A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology* 17, 130-132 (1999)
- [32] Gavin *et al.*: Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631-636 (2006)
- [33] Krogan *et al.*: Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 440, 637-643 (2006)
- [34] Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A.J.: A gateway-compatible yeast one-hybrid system. *Genome Research* 14, 2093-2101 (2004)
- [35] Lee *et al.*: Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799-804 (2002).
- [36] Zhu *et al.*: High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Research* 19, 556566 (2009).
- [37] Gerstein *et al.*: Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91100 (2012).
- [38] Neph *et al.*: Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 150, 1274-1286 (2012).
- [39] Jeong *et al.*: The large-scale organization of metabolic networks. *Nature* 407, 651-654 (2000).
- [40] Kanehisa *et al.*: KEGG for linking genomes to life and the environment. *Nucleic Acid Research* 36 (Database issue), D480-D484 (2008).
- [41] Salwinski *et al.*: The database of interacting proteins: 2004 update. *Nucleic Acids Research* 32 (Database issue), D449-D451 (2004).

- [42] Stark *et al.*: The BioGRID interaction database: 2011 update. *Nucleic Acids Research* 39 (Database issue), D698-D704 (2011).
- [43] Szklarczyk *et al.*: The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* 39 (Database issue), D561-D568 (2011).
- [44] Matys *et al.*: TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* 31, 374-378 (2003).
- [45] Gama-Castro *et al.*: RegulonDB version 7.0: transcriptional regulation of *Escherichia coli* K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Research* 39 (Database issue), D98-D105 (2011).
- [46] Palaniswamy *et al.*: AGRIS and AtRegNet, a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiology* 140, 818829 (2006).
- [47] Pagel *et al.*: The MIPS mammalian protein-protein interaction database. *Bioinformatics* 21, 832-834 (2005).
- [48] Bader, G.D., Betel, D., Hoque, C.W.: BIND - the Biomolecular Interaction Network Database. *Nucleic Acid Research* 31, 248-250 (2003).
- [49] Caspi *et al.*: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acid Research* 38 (Database issue), D473-D479 (2010).
- [50] Croft *et al.*: Reactome: a database of reactions, pathways and biological processes. *Nucleic Acid Research* 39 (Database issue), D691-D697 (2011).

- [51] Smoot, M., Ono, K., Ruscheinski, J., Wang, P.-L., Ideker, T.: Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27, 431-432 (2011).
- [52] Carbon *et al.*: AmiGO: online access to ontology and annotation data. *Bioinformatics* 25, 288-289 (2009).
- [53] Albert, R., Jeong, H., & Barabasi, A.-L.: Error and attack tolerance of complex networks. *Nature* 406, 378-382 (2000).
- [54] Jeong, H., Mason, S. P., Barabasi, A.-L., Oltvai, Z. N.: Lethality and centrality in protein networks. *Nature* 411, 41-42 (2001).
- [55] Hahn, M. W., Kern, A. D.: Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution* 22, 803-806 (2004).
- [56] He, X., Zhang, J.: Why do hubs tend to be essential in protein networks? *PLoS Genetics* 2, e88 (2006).
- [57] Wagner, A. & Fell, D. A.: The small world inside large metabolic networks. *Proceedings of The Royal Society B* 268, 1803-1810 (2001).
- [58] Yu, H. & Gerstein, M.: Genomic analysis of the hierarchical structure of regulatory networks. *Proceedings of the National Academy of Sciences* 103, 14724-14731 (2006).
- [59] Milo *et al.*: Network motifs: simple building blocks of complex networks. *Science* 298, 824-827 (2002).
- [60] Alon, U: Network motifs: theory and experimental approaches. *Nature Reviews Genetics* 8, 450-461 (2007).

- [61] Mangan, S. & Alon, U.: Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences* 100, 11980-11985 (2004).
- [62] Kashtan, M., Itzkovitz, S., Milo, R., & Alon, U.: Efficient algorithm for estimating subgraph concentration and detecting network motifs. *Bioinformatics* 20, 1746-1758 (2004).
- [63] Erdos, P., Renyi, A.: On the strength of connectedness of a random graph. *Acta Mathematica Hungarica* 12, 261-267 (1960).
- [64] Chung, F., Lu, L., Dewey, T.G., Galas, D.J.: Duplication models for biological networks. *Journal of Computational Biology* 10, 677-687 (2003).
- [65] Ohno, S.: *Evolution by gene duplication*. Springer Verlag, New York (1970).
- [66] Przulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: scale-free or geometric? *Bioinformatics* 20, 3508-3515 (2004).
- [67] von Mering et al.: Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399-403 (2002).
- [68] Reguly et al.: Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *Journal of Biology* 5, 11 (2006).
- [69] D'haeseleer, P. & Church, G. M. : Estimating and improving protein interaction error rates. In *Proc IEEE Comput Syst Bioinform Conf: August 16-19 2004; California*. IEEE Computer Society 216-223 (2004).
- [70] Hart, G. T., Ramani, A. K., Marcotte, E. M.: How complete are current yeast and human protein-interaction networks? *Genome Biology* 7, 120 (2006).

- [71] Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D.: Protein interactions: two methods for assessment of the reliability of high throughput observations. *Molecular and Cellular Proteomics* 1, 349-356 (2002).
- [72] Grigoriev, A.: On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Research* 31, 4157-4161 (2003).
- [73] Stumpf *et al.*: Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences* 105, 6959-6964 (2008).
- [74] Rottger, R., Ruckert, U., Taubert, J., Baumbach, J.: How little do we actually know? - On the size of gene regulatory networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9, 1293-1300 (2012).
- [75] Wilson, D., Charoensawan, V., Kummerfeld, S. K., Teichmann, S. A.: DBD-taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Research* 36 (Database issue), D88-D92 (2008).