

國立交通大學

資訊工程系

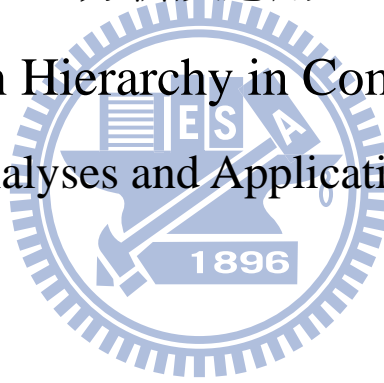
博士論文

複雜網路中具縮影性質之階層：

分析及應用

The Abstraction Hierarchy in Complex Networks:

Analyses and Applications



博士生：鄭家胤

指導教授：孫春在 教授

中華民國九十八年七月

國立交通大學

資訊工程系

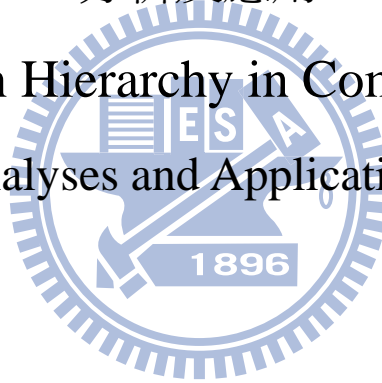
博士論文

複雜網路中具縮影性質之階層：

分析及應用

The Abstraction Hierarchy in Complex Networks:

Analyses and Applications



博士生：鄭家胤

指導教授：孫春在 教授

中華民國九十八年七月

複雜網路中具縮影性質之階層：
分析及應用

The Abstraction Hierarchy in Complex Networks:
Analyses and Applications

研究生：鄭家胤

Student：Chia-Ying Cheng

指導教授：孫春在

Advisor：Chuen-Tsai Sun



A Dissertation

Submitted to

Department of Computer Science

College of Computer Science

National Chiao Tung University

for the Degree of

Doctor of Philosophy

in

Computer Science

July 2009

Hsinchu, Taiwan, Republic of China

中華民國九十八年七月

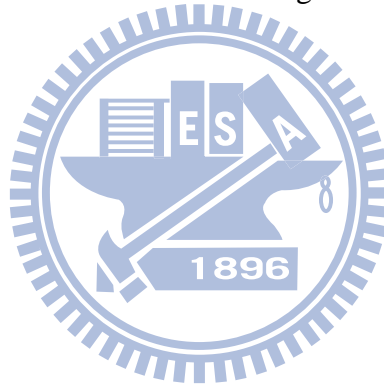
Abstract (in Chinese)

以複雜網路的型態來呈現複雜系統中的互動關係是一種方便且行之有年的研究方法，包括在生物學、生態學、社會學等等的領域上，除了讓研究者有不同於以往該領域傳統議題的新觀點外，許多新的方法也因此被提出來以解決在各種不同複雜系統上的問題。其中，最重要也最具挑戰性的一個問題就是，如何將複雜網路做分群(在社會學領域被稱之為共同體(*community*)或是群組(*group*))，在生物學上被稱為基塊(*motif*)或是模組(*module*))。如何(1)找出模組，(2)階層性的組織，及(3)這兩者對應到真實世界的關係，一直是研究者的焦點所在。儘管已經有一些成功的研究，但是至今仍沒有一個標準的衡量方法可以來找出模組或是階層性組織。以階層式組織來說，大多數的研究專注於其模組在不同階層上垂直面向的關係之探討-其可用來表示"包含(*inclusion*)"，"因果(*causality*)"和"調控(*regulation*)"關係;但往往忽略了其在同一階層上水平面向的關係之研究-其可用來提供給研究者在某一階層上的網路的縮影(*abstraction*)或是骨架(*backbone*)。在本論文研究中，我提出了一雙向式尋找模組及建構階層組織的方法，其同時考慮了各個模組間垂直和水平的關係來建構出該複雜系統的金字塔階層(*pyramid hierarchies*)，此方法除了被人工網路驗證外，也被應用在生物及社會網路上，其結果顯示該方法在擷取複雜系統之資訊上卓越的效能。

Abstract (in English)

The use of nodes and links to assemble networks is convenient for representing interactions in complex systems. This benefits researchers in biology, ecology, sociology and other biological and social sciences. In addition to supporting alternative views of complex domains, network research is also supporting new methods for solving problems in a range of domains. One particularly important and challenging problem is partitioning networks into clusters (called communities or groups in social science research and motifs or modules in biology). Research in these areas has focused on identifying modules and hierarchical organizations that correspond to real-world meanings (e.g., biological functions or economic and political constraints). Despite a number of successful examples, no uniform measure of modularity or standard hierarchical structure exists. Most current descriptions of hierarchical organizations are limited to vertical relationships between modules at different hierarchical levels, thus overlooking horizontal relationships that express associations among modules at the same level. Vertical relationships can be used to represent inclusion hierarchies and to describe causality/regulation. Horizontal relationships complement these by providing abstractions of original networks of interest at various levels in a hierarchy (Fig. 1).

In this dissertation I describe a proposal for a two-way simultaneous module-finding and hierarchy-building strategy. I take both vertical and horizontal relationships between modules into consideration when building pyramid hierarchies in which each layer represents an abstraction of lower-level networks. This dissertation also contains descriptions of tests for this proposed approach, using networks consisting of anywhere from tens to hundreds of nodes and links, and in domains that include artificial random networks, social networks, and biological networks. The results demonstrate its performance for information mining from complex systems.



Acknowledgements

I would like to express my deep gratitude to my advisor, Professor Chuen-Tsai Sun, for teaching me how to define a problem, how to determine its essential components, and how to identify relationships among those components. I am especially grateful for the way he modeled the right attitude for students to take when addressing new topics. That is a skill I will use both in my research and in everyday life.

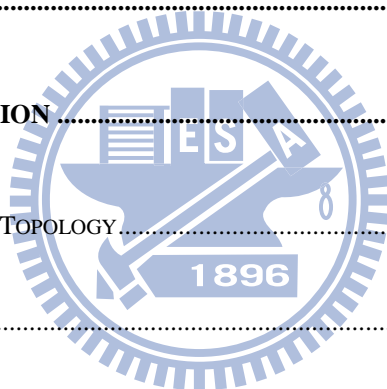
I also thank my co-advisor, Professor Yuh-Jyh Hu, for his strong support during the last two years of my Ph.D. research and writing. As a mentor he discussed in great detail new ideas and ways to implement them. As a friend he shared his life and work values with me, which helped me stay centered during the entire process.

For all others who were patient with me as I completed this project, I extend my thanks for your support and patience.

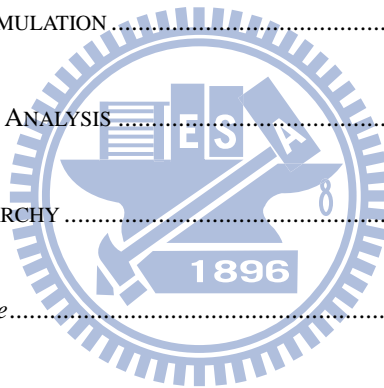
Finally, I give thanks to my Lord Jesus Christ: "And we know that all things work together for good to those who love God, to those who are called according to His purpose" (Romans 8:28).

Contents

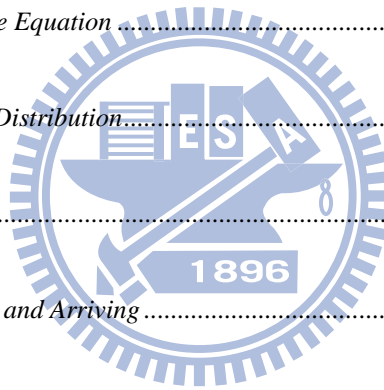
ABSTRACT (IN CHINESE)	I
ABSTRACT (IN ENGLISH)	II
ACKNOWLEDGEMENTS	IV
CONTENTS	V
LIST OF TABLES	IX
LIST OF FIGURES	XI
CHAPTER 1 INTRODUCTION	1
1.1 COMPLEX NETWORK TOPOLOGY	3
1.1.1 <i>Randomness</i>	4
1.1.2 <i>Small-world property</i>	5
1.1.3 <i>Scale-free distributions</i>	7
1.2 COMPLEX NETWORK STRUCTURE	8
1.2.1 <i>Motifs</i>	8
1.2.2 <i>Communities</i>	11
1.2.3 <i>Hierarchical modularity</i>	11
1.3 NETWORK DYNAMICS	13



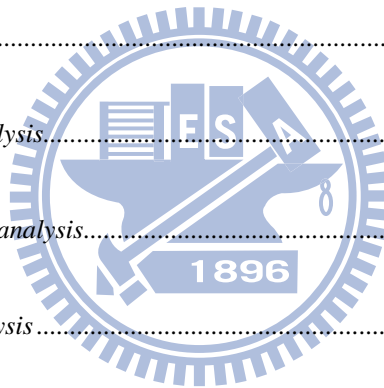
1.3.1	<i>Cellular automata</i>	13
1.3.2	<i>Preferential attachment</i>	14
CHAPTER 2	STATIC NETWORKS AND DYNAMIC PROCESS CHARACTERIZATION	
	AND ANALYSIS	16
2.1	NETWORK MOTIF DETECTION.....	18
2.1.1	<i>General: Bridge and Brick Network Motif-Detecting Algorithm</i>	21
2.1.2	<i>Specific: Bridge and Brick Network Motif-Detecting Algorithm</i>	26
2.2	SOCIAL NETWORK SIMULATION.....	27
2.3	EPIDEMIC DYNAMICS ANALYSIS.....	29
2.4	ABSTRACTION HIERARCHY.....	33
2.4.1	<i>Proximity Measure</i>	34
2.4.2	<i>Network Abstraction</i>	43
CHAPTER 3	NETWORK MOTIF EXPERIMENTS	45
3.1	GENERAL: BRIDGE AND BRICK NETWORK MOTIF-DETECTING ALGORITHMS	45
3.1.1	<i>Validation</i>	49
3.1.2	<i>Experiments</i>	52
3.1.3	<i>Conclusion</i>	60



3.2	SPECIFIC: BRIDGE AND BRICK NETWORK MOTIF-DETECTING ALGORITHMS.....	63
3.2.1	<i>Validation</i>	63
3.2.2	<i>Experiments</i>	67
3.2.3	<i>Conclusion</i>	76
CHAPTER 4 SOCIAL NETWORK SIMULATION EXPERIMENTS		77
4.1	FRIENDSHIP EVOLUTION AND THE THREE-RULE MODEL	77
4.1.1	<i>Friendship Selection Methods</i>	79
4.1.2	<i>Friendship Update Equation</i>	80
4.1.3	<i>Fitting a Normal Distribution</i>	81
4.2	EXPERIMENT	83
4.2.1	<i>Effects of Leaving and Arriving</i>	85
4.2.2	<i>Effects of Breakup Threshold</i>	88
4.2.3	<i>Effects of Resources</i>	89
4.2.4	<i>Effects of Initial Friendship</i>	90
4.2.5	<i>Distribution of Co-Directors</i>	92
4.2.6	<i>Sampling</i>	94
4.3	CONCLUSION.....	95



CHAPTER 5	EPIDEMIC DYNAMICS EXPERIMENTS	97
5.1	EPIDEMIC DYNAMICS IN COMPLEX NETWORKS	98
5.2	EXPERIMENTS	102
5.3	CONCLUSION.....	111
CHAPTER 6	ABSTRACTION HIERARCHY EXPERIMENTS	112
6.1	BACKGROUND.....	113
6.2	VALIDATION	118
6.3	EXPERIMENTS	121
6.3.1	<i>Club network analysis</i>	122
6.3.2	<i>Football network analysis</i>	125
6.3.3	<i>PPI network analysis</i>	127
6.3.4	<i>Metabolic network analysis</i>	134
CHAPTER 7	CONCLUSION	140
CHAPTER 8	REFERENCE.....	142



List of Tables

TABLE 1. AN UPDATE RULE FOR A ONE-DIMENSIONAL, TWO-STATE CELLULAR AUTOMATON.	14
TABLE 2. BRIDGE AND BRICK SUBGRAPH FREQUENCIES IN FOUR COMPLEX NETWORK CATEGORIES (FOR VALIDATION PURPOSES).	51
TABLE 3. BRICK AND BRIDGE MOTIFS IN FOURTEEN REAL WORLD NETWORKS, INCLUDING EDGE AND NODE DEFINITIONS, NETWORK SIZES, AND REFERENCES.	60
TABLE 4. DESCRIPTIONS OF FIVE GENE REGULATION NETWORKS: EDGE AND NODE DEFINITIONS, NETWORK SIZES, AND REFERENCES.	70
TABLE 5. BRICK AND BRIDGE MOTIFS IN FIVE GENE REGULATION NETWORKS.	72
TABLE 6. TERMS AND ABBREVIATIONS FOR INITIALIZED PARAMETERS.	84
TABLE 7. TERMS AND ABBREVIATIONS FOR STATISTICS.	85
TABLE 8. EFFECTIVE DIRECTIONS OF THE PARAMETERS ON $\langle k \rangle$, C , L	92
TABLE 9. CORRELATIONS BETWEEN $\langle k \rangle$, C , L FROM EXPERIMENTS.	92
TABLE 10. SUMMARY OF BIOLOGICAL SIGNIFICANCE OF MODULES BASED ON GO BIOLOGICAL PROCESS ANNOTATIONS.	131
TABLE 11. SUMMARY OF WITHIN-MODULE CONSISTENCY OF METABOLIC PATHWAY CLASSIFICATION BASED ON KEGG.	137



List of Figures

FIG. 1. A PYRAMID OF THE COMPLEX NETWORK WITH VERTICAL AND HORIZONTAL RELATIONSHIPS.	3
FIG. 2. THE COMPARISON BETWEEN THE RANDOM NETWORK AND THE SCALE-FREE NETWORK.....	8
FIG. 3. 13 POSSIBLE OF TRIAD MOTIFS DEFINED BY ALON.	9
FIG. 4. COMMUNITIES CAN BE DEFINED AS GROUPS OF NODES SUCH THAT THERE IS A HIGHER DENSITY OF EDGES WITHIN GROUPS THAN BETWEEN THEM.	11
FIG. 5. THE HIERARCHICAL NETWORK AND ITS DEGREE DISTRIBUTION.	13
FIG. 6. NETWORK MOTIFS EXAMPLE.	20
FIG. 7. LINK-WEIGHTED VALUE CALCULATING EXAMPLE. THE LINK-WEIGHTED VALUE WEIGHT (A, B) OF EDGE (A, B) IS 0 WHILE WEIGHT (B, C).....	21
FIG. 8. THE SMALL-WORLD MODEL. BLACK SIGNIFIES STRONG LINKS AND RED WEAK LINKS.	27
FIG. 9. THREE-RULE MODEL FLOW DIAGRAM.	29
FIG. 10 FLOWCHART FOR A SIS EPIDEMIOLOGICAL SIMULATION MODEL.....	33
FIG 11. FOUR SIMPLE NETWORKS TO ILLUSTRATE PROXIMITY MEASURES.....	39
FIG 12. A SIMPLE UNDIRECTED WEIGHTED NETWORK.	41
FIG. 13. PERCENTAGES OF BRIDGE AND BRICK MOTIFS IN SMALL-WORLD NETWORKS ACCORDING TO DIFFERENT REWIRING RATIOS.....	52
FIG. 14. BRIDGE MOTIF RATIO PROFILES FOR THREE ELECTRICAL CIRCUITS (s208, s420 AND s838).....	54

FIG. 15. BRIDGE MOTIF RATIO PROFILES FOR TWO SOCIAL NETWORKS.	56
FIG. 16. BRICK MOTIF RATIO PROFILES FOR TWO SOCIAL NETWORKS.	56
FIG. 17. BRICK-BRIDGE MOTIF RATIO PROFILES FOR TWO REGULATION NETWORKS (ONE BACTERIA AND ONE EUKARYOTE).	57
FIG. 18. BRIDGE MOTIF RATIO PROFILES FOR SEVEN FOOD WEBS.....	58
FIG. 19. RELATIONSHIPS BETWEEN CLUSTERING COEFFICIENTS AND DIFFERENT REMOVAL RATIOS FOR THREE <i>E. COLI</i> LINK TYPES. RED, RANDOM; GREEN, STRONG; BLUE, WEAK.	65
FIG. 20. RELATIONSHIPS BETWEEN CLUSTERING COEFFICIENTS AND DIFFERENT REMOVAL RATIOS FOR THREE <i>S. CEREVISIAE</i> (YEAST) LINK TYPES. RED, RANDOM; GREEN, STRONG; BLUE, WEAK.....	66
FIG. 21. COMPARISON OF ORIGINAL (BLUE CURVE) AND ALTERED (RED CURVE) BRICK MOTIF RATIO PROFILES FOR <i>E. COLI</i> AFTER RANDOMLY REMOVING 40% OF ITS LINKS. ALTERED RESULTS REPRESENT AVERAGE VALUES FOR 30 RUNS.	66
FIG. 22. COMPARISON BETWEEN ORIGINAL BRICK MOTIF RATIO PROFILES AND ALTERED BRICK MOTIF RATIO PROFILES FOR <i>S. CEREVISIAE</i> (YEAST) AFTER RANDOMLY REMOVING 40% OF ITS LINKS.	67
FIG. 23. DISTRIBUTION OF LINK WEIGHTS IN <i>E. COLI</i> . AVERAGE MEAN AND STANDARD DEVIATION OF LINK WEIGHTS FOR RANDOMIZED NETWORKS WERE CALCULATED AS 0.90 ± 0.04	67
FIG. 24. COMPARISONS OF TRIAD SIGNIFICANCE PROFILES (TSPs) FOR OUR BRIDGE AND BRICK MOTIFS AND MILO ET AL.'S [7], [28] <i>E. COLI</i> MOTIFS.....	74

FIG. 25. COMPARISONS OF TRIAD SIGNIFICANCE PROFILES (TSPs) FOR OUR BRIDGE AND BRICK MOTIFS AND MILO ET AL.'S [7], [28] <i>S. CEREVISIAE</i> (YEAST) MOTIFS.	75
FIG. 26. BRICK MOTIF RATIO PROFILES FOR TWO GENE REGULATION NETWORKS: <i>E. COLI</i> AND <i>S.</i> <i>CEREVISIAE</i> (YEAST).	75
FIG. 27. BRIDGE MOTIF RATIO PROFILES FOR THREE GENE REGULATION NETWORKS: <i>C. ELEGANS</i> , <i>SEA</i> <i>URCHIN</i> , AND <i>DROSOPHILA</i>	75
FIG. 28. BETA14 PDF CURVES AT DIFFERENT AVERAGES OF 0.1, 0.5, AND 0.9.	83
FIG. 29. COMPARISON OF BETA AND NORMAL DISTRIBUTIONS.	83
FIG. 30. EXAMPLE OF A STATISTICALLY STATIONARY STATE USING THE PROPOSED MODEL.	85
FIG. 31. $\langle k \rangle$, C AND L VARYING IN BREAKUP THRESHOLD θ WITH DIFFERENT LEAVING AND ARRIVING PROBABILITY P	87
FIG. 32. $\langle k \rangle$, C AND L VARYING IN FRIEND-REMEMBERING Q VALUE WITH DIFFERENT DISTRIBUTIONS OF FRIEND-MAKING RESOURCES.	88
FIG. 33. $\langle k \rangle$, C AND L VARYING IN FRIEND-REMEMBERING Q VALUE WITH DIFFERENT DISTRIBUTIONS OF INITIAL FRIENDSHIP F_0	89
FIG. 34. TWO-RULE MODEL DEGREE DISTRIBUTION $P(k)$	90
FIG. 35. $\langle k \rangle$, C AND L VARYING IN LEAVING AND ARRIVING PROBABILITY P	92
FIG. 36. MODEL ACQUAINTANCE NETWORK SAMPLES.	94

FIG. 37 PREVALENCE P IN STEADY STATE AS A FUNCTION OF EFFECTIVE SPREADING RATE λ	101
FIG. 38. RELATIONSHIP BETWEEN EFFECTIVE SPREADING RATE AND STEADY DENSITY OF THE SIS EPIDEMIOLOGICAL MODEL ON THREE TYPES OF COMPLEX NETWORK PLATFORMS.	104
FIG. 39. HOW THE AMOUNT OF AN INDIVIDUAL'S ECONOMIC RESOURCES AFFECT STEADY DENSITY CURVES.	106
FIG. 40. RELATIONSHIP BETWEEN RATIO OF TRANSMISSION COSTS TO AN INDIVIDUAL'S ECONOMIC RESOURCES AND CRITICAL THRESHOLD.....	106
FIG. 41. HOW DIFFERENT DISTRIBUTION TYPES OF INDIVIDUAL ECONOMIC RESOURCES (DELTA, UNIFORM, NORMAL, POWER-LAW) AFFECT STEADY DENSITY CURVES AND CRITICAL THRESHOLDS OF INFECTIOUS DISEASE DIFFUSION IN A SCALE-FREE NETWORK.	108
FIG. 42. A UNIFORM ($N = 5, R = 2$) AND NORMAL DISTRIBUTION (STANDARD DEVIATION = 2) OF INDIVIDUAL ECONOMIC RESOURCES WITH AVERAGE VALUE $\langle R \rangle$ OF 16.	109
FIG. 43. INDIVIDUAL ECONOMIC RESOURCES IN A POWER-LAW DISTRIBUTION.	109
FIG. 44. HOW DIFFERENT TYPES OF INDIVIDUAL ECONOMIC RESOURCE DISTRIBUTIONS (DELTA, UNIFORM, NORMAL, AND POWER-LAW) AFFECT STEADY DENSITY CURVES AND CRITICAL THRESHOLDS OF INFECTIOUS DISEASE DIFFUSION IN A SCALE-FREE NETWORK.	110
FIG. 45. A UNIFORM ($N = 5, R = 3$) AND NORMAL DISTRIBUTION (STANDARD DEVIATION = 3) OF INDIVIDUAL ECONOMIC RESOURCES WITH AVERAGE VALUE $\langle R \rangle$ OF 16.	110

FIG. 46. INDIVIDUAL ECONOMIC RESOURCES IN A POWER-LAW DISTRIBUTION.	111
FIG. 47. EXAMPLE OF A DENDROGRAM FROM CONVENTIONAL HIERARCHICAL CLUSTERING.	118
FIG. 48. VALIDATION OF TWO-WAY MODULE-FINDING-HIERARCHY-BUILDING STRATEGY.	120
FIG. 49. ABSTRACT NETWORK CORRESPONDING TO HIERARCHICAL LEVEL THREE AND TWO.	121
FIG. 50. CLUSTERING RESULTS OF ZACHARY’S KARATE CLUB NETWORK.	124
FIG. 51. THE ANALYSIS OF THE FOOTBALL NETWORK.	127
FIG. 52 THE P-VALUE OF THE CORRESPONDING NODES AT DIFFERENT LEVELS.	132
FIG. 53. THE MST OF PPI AT LEVEL FOUR AND FIVE, AND HAVE BLUE AND RED COLORS, RESPECTIVELY.	133
FIG. 54. THE MAPPING BETWEEN THE MODULES WE FOUND AND THE REAL GO ID.	134
FIG. 55. THE PYRAMID OF ABSTRACTION DISCLOSED FROM A METABOLIC NETWORK.	138
FIG. 56. EXAMPLE OF THE VERTICAL RELATIONSHIPS IN AN ABSTRACTION PYRAMID DISCLOSED FROM A METABOLIC NETWORK.	138
FIG. 57. EXAMPLE OF THE HORIZONTAL RELATIONSHIP AT THE THIRD LEVEL OF AN ABSTRACTION PYRAMID.	139

Chapter 1 Introduction

Complex networks consist of sets of items called vertices or nodes and connections between them called edges. There are many examples of systems in the form of networks (also called “graphs” in mathematics): the World Wide Web, the Internet, social networks (acquaintances or other connections), distribution networks (e.g., blood vessels, postal delivery routes), organizations and business relations, neural networks, metabolic networks, food webs, and research paper citations, among many others.

The network concept is proving to be a very useful tool for studying complex systems [1-3]. While no general theory of complexity exists [2, 4, 5], there is a growing collection of related theories, paradigms, and tools, many of them associated with physics and mathematics [5, 6]. They support explanations of complex phenomena such as collective behavior observed in ferromagnetic phase transitions, herding behavior, disease epidemics, and opinion formation—all examples of local interactions that create global order [5, 7]. We also know that even very simple systems such as discrete logistic growth models (logistic maps) can display rich and complex dynamics. To a certain extent, self-organized criticality explains how some

systems manage to operate near criticality in the absence of fine-tuning [5, 8]. Fractal geometry helps explain how and why certain forms and structures in nature arise—for instance, vascular systems. As a new tool for studying complex phenomena, network theory uses a mix of statistical mechanics, graph theory, and dynamical systems theory [9-11].

The majority of network-related problems can be placed in one of two categories: for static networks, relationships between network structure and function, and for dynamic networks, global rules tied to network evolution [12-15]. In the following section I will introduce fundamental findings associated with these two categories.



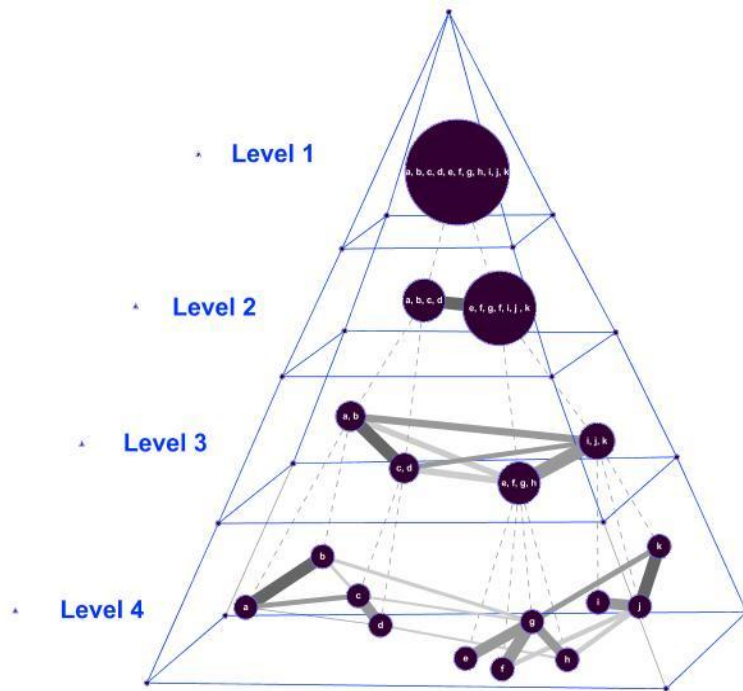
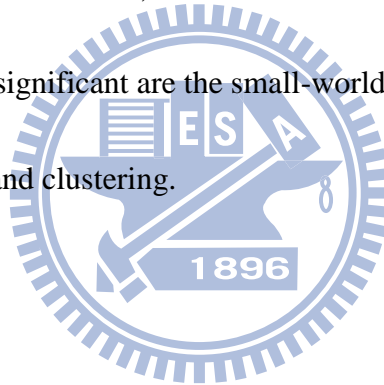


Fig. 1. A pyramid of the complex network with vertical and horizontal relationships.

1.1 Complex Network Topology

Many systems in nature and technology consist of large numbers of highly interconnected dynamical units [2, 16]. Examples include coupled biological and chemical systems, neural networks, social interaction, and the Internet. An initial approach to capturing the global properties of such systems is to model them as graphs whose nodes represent dynamical units (e.g., neurons in the brain or individuals in a social system) and whose links represent interactions between units. Of course, this is a very strong approximation that requires translating interactions

between dynamical units (generally dependent on temporal, spatial, and many other details) into simple binary numbers designating the existence or lack of links between two corresponding nodes. Such approximations provide simple yet informative representations of whole systems. The development of powerful and reliable data analysis tools represent better mechanisms for exploring the topological properties of multiple networked systems, thus supporting topological analyses of interactions in a diverse range of systems (e.g., communication, social, and biological). These efforts reveal that despite inherent differences, most real networks have the same topological properties [1, 5]. The most significant are the small-world effect, degree scale-free distributions, correlations, and clustering.



1.1.1 Randomness

The first non-regular network model [17, 18] was introduced by Paul Erdős and Alfred Rényi in the late 1950s [19]. In this dissertation I will variously refer to this as the random model, the Erdős-Rényi model, or the ER model. The ER model of a random network starts with N nodes and connections between pairs of nodes at a p probability, resulting in graphs with approximately $N(N-1)/2$ randomly placed links (Fig. 2, part Aa). Node degrees follow a Poisson distribution (Fig. 2, part Ab),

indicating that most nodes have approximately the same number of links (close to the average degree $\langle k \rangle$). The tail (high k region) of the $P(k)$ degree distribution decreases exponentially, indicating the rarity of nodes that significantly deviate from the average.

1.1.2 Small-world property

This property was first investigated in the 1960s in a social context, as part of a series of experiments designed by Milgram [20, 21] to estimate the number of steps in acquaintance chains. In his first experiment, Milgram asked randomly selected people in Nebraska to send letters that would eventually arrive at the home of an individual living in Boston, identified only by his name, occupation, and city of residence. The step-by-step letters could only be sent to individuals that the current sender knew by first name, and who were presumably closer to the final recipient. Milgram kept track of the paths followed by the letters and of the demographic characteristics of their handlers. At the time of these experiments, the commonly held belief was that it would take hundreds of steps for letters to reach their final destination, but Milgram found that the number of links needed to reach the targeted individual was six. Dodds et al. [22] have recently replicated Milgram's experiment using e-mail, completing

enough connecting chains so as to allow for a thorough statistical characterization.

The small-world property has been observed in a variety of real networks (including biological and technological [2, 4, 23]), and is now an accepted mathematical property in some network models (e.g., random graphs).

In 1998, Watts and Strogatz [21] proposed a new model for explaining small path lengths and large clustering coefficients that are independent of network size—two properties shared by many real networks. According to their model, the first step is to construct a network with a one-dimensional ring lattice of N nodes (or d -dimensional regular lattice) in which each node is wired to its neighbors up to k th nearest neighbor. Such regular lattices have high average path lengths. Decreasing those lengths requires the rewiring of each link with a p probability to another randomly picked node—a process that establishes long-range connections. A small-world network displays characteristics of a regular lattice for very small p values and an ER network for very large p values, meaning that small-world networks lie somewhere between order and randomness. Average path length in a small-world network is expressed as

$$(N, p) \propto \frac{N}{K} f(pKN) \quad (1.1)$$

$$f(u) = \frac{4}{\sqrt{u^2 + 4u}} \tanh^{-1} \frac{u}{\sqrt{u^2 + 4u}} \quad \text{for } u \gg 1 \text{ or } u \ll 1. \quad (1.2)$$

This function is a constant for $u \ll 1$, and behaves as $\ln(u)/u$ for $u \gg 1$. Accordingly,

the clustering coefficient for small-world networks is $C_{SW} \propto (1 - p)^3$.

Small-world networks share some properties with a number of real networks.

However, their degree distribution has a pronounced peak at $\langle k \rangle = K$ and

exponentially decaying wings for large k , thereby distinguishing them from the power

law degree distributions of networks such as the WWW, the Internet, and many social

networks.

1.1.3 Scale-free distributions

Many scale-free networks are characterized by a power-law degree distribution [24] in

which the probability that a node has k links follows $P(k) \sim k^{-\gamma}$, where γ is the degree

exponent. The probability that a node is highly connected is statistically more

significant than in a random graph (see Fig. 2, part Ba), with network properties often

determined by a relatively small number of highly connected nodes known as hubs. In

the Barabási–Albert scale-free model network model [24], a node with M links is

added to the network at each time point and connects to an already existing node I

with probability $\frac{k_I}{\sum_J k_J}$, where k is the degree of node I and J the index

denoting the sum over network nodes. The network generated by this growth process

has a power-law degree distribution characterized by the degree exponent $\gamma = 3$, a

distribution represented by a straight line on a log–log plot (see Fig. 2, part Bb). The network created using the Barabási–Albert model [24, 25] does not have an inherent modularity, meaning that $C(k)$ is independent of k . Scale-free networks with degree exponents $2 < \gamma < 3$ (a range that is observed in most biological and non-biological networks) are ultra-small, with average path lengths that follow $l \sim \log \log N$. This is significantly shorter than $\log N$, which is characteristic of random small-world networks [21].

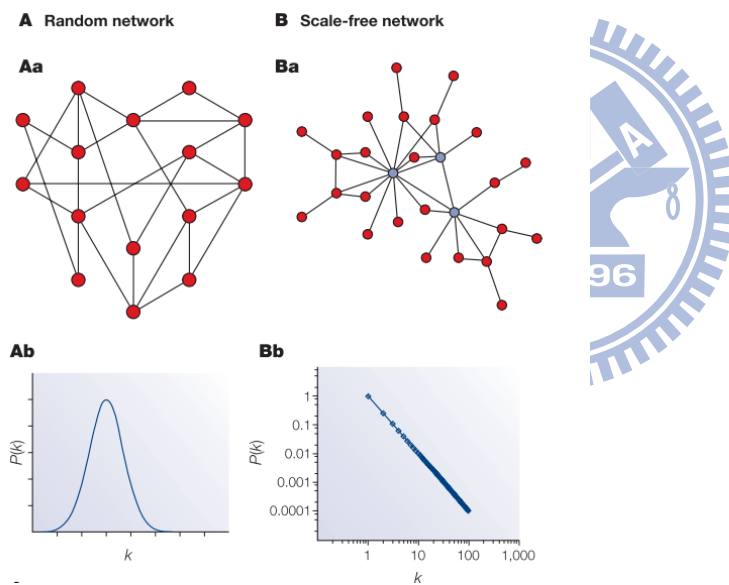


Fig. 2. The comparison between the random network and the scale-free network.

1.2 Complex Network Structure

1.2.1 Motifs

A motif (M) is a pattern of interconnections occurring in either directed or undirected

graphs (G) at a number that is significantly higher than in randomized versions (i.e. in graphs with the same number of nodes, links and degree distribution as the original one, but where the links are randomly distributed) [16, 26]. As a pattern of interconnections, M is usually expressed as a connected (undirected or directed) n -node graph that is a subgraph of G . All the possible three-node connected directed graphs are illustrated in Fig. 3.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13
Motif													

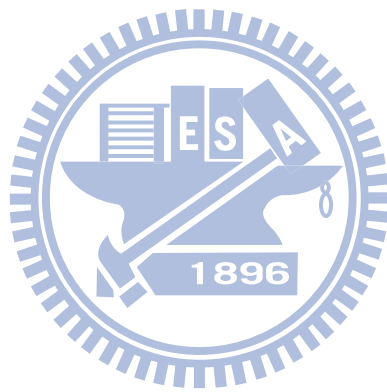
Fig. 3. 13 possible of triad motifs defined by Alon.

The concept of motifs was originally introduced by Alon et al.[16] who studied small n -node motifs in biological and other networks. Significant motif research in a G graph consists of matching algorithms – that is, counting the total number of occurrences of each n -node subgraph M in the original graph and in the randomized graphs. The statistical significance of M is then described in terms of Z-score, defined as

$$Z_M = \frac{n_M - \langle n_M^{rand} \rangle}{\sigma_{n_M}^{rand}} \quad (1.3)$$

Where n_M is the number of times subgraph M appears in G , and $\langle n_M^{rand} \rangle$ and $\sigma_{n_M}^{rand}$ are the mean and standard deviation for the number of appearances in the randomized

network ensemble.



1.2.2 Communities

Community and the first network formalizations of the concept were proposed by social scientists. Given a graph $G(N, L)$, a community (or cluster, or cohesive subgroup) can be expressed as subgraph $G(N', L')$, whose nodes are tightly connected, *i.e.* cohesive. Since the structural cohesion of the nodes of G is quantified in several different ways, there are different formal definitions of community structures(Fig. 4)[27-29].

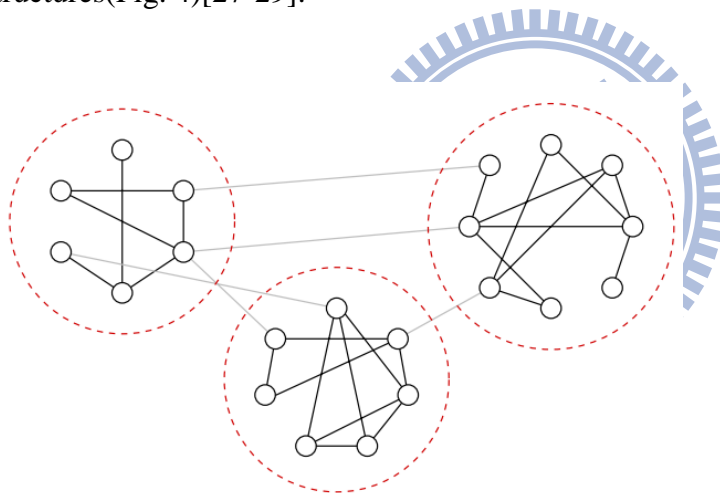


Fig. 4. Communities can be defined as groups of nodes such that there is a higher density of edges within groups than between them.

1.2.3 Hierarchical modularity

To account for the coexistence of modularity, local clustering and scale-free topology

in many real systems, one must assume that clusters combine in an iterative manner to generate hierarchical networks (Fig. 5, part A). The starting point for such construction is a small cluster of four densely linked nodes (e.g., the four central nodes in Fig. 5, part A). Next, three module replicas are generated and three external nodes of the replicated clusters are connected to the central node of the old cluster, thereby producing a large 16-node module. After generating three replicas of this 16-node module, the 16 peripheral nodes are also connected to the central node of the old module, producing a new 64-node module. The hierarchical network model seamlessly integrates a scale-free topology with an inherent modular structure by generating a network that has a power-law degree distribution with degree exponent $\gamma = 1 + n_4/n_3 = 2.26$ (see Fig.5, part B) and a large, system-size independent average clustering coefficient $\langle C \rangle \sim 0.6$. A hierarchical architecture implies that sparsely connected nodes are part of highly clustered areas, with communication between the different highly clustered neighborhoods being maintained by a few hubs (see Fig. 5, part A)[30].

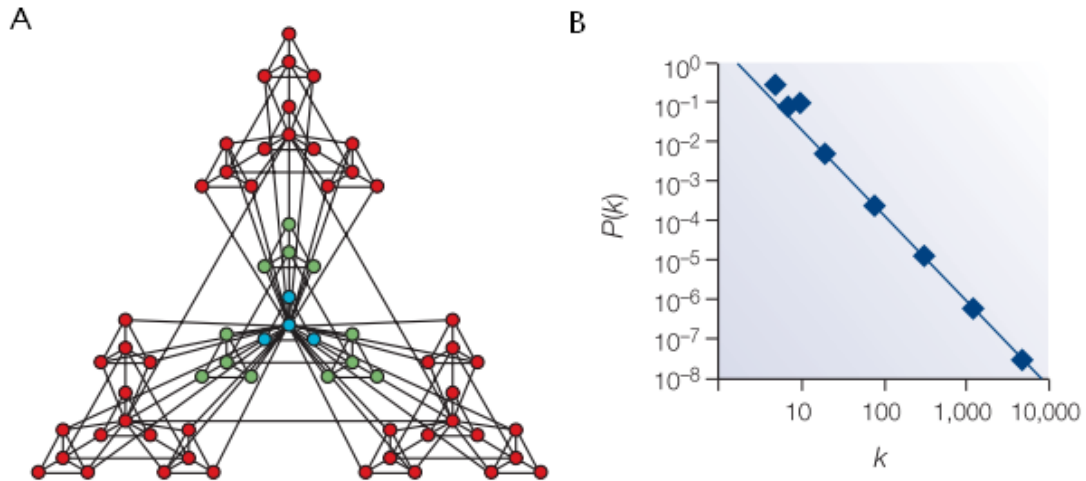


Fig. 5. The hierarchical network and its degree distribution.

1.3 Network Dynamics

Since actual complex networks are not necessarily static, simulating and/or studying the dynamics of the complex networks is a difficult task. The following methods can be used to address this problem in different domains.

1.3.1 Cellular automata

Cellular Automata (CA) [31] are simple examples of discrete dynamical systems. A

cellular automaton consists of a regular cell grid consisting a finite number of states.

Each cell state during time step $t+1$ is determined by states of cells in time step t . In the

example shown in Table 1, a one-dimensional, two-state cellular automaton (i.e., a

bit-string CA) can be defined using the update rule defined in Table 1. In this example,

the state of a cell is determined by its own state and the states of its nearest neighbors.

This type of automaton can easily be schematized as a directed network in which each cell takes inputs from two neighboring cells.

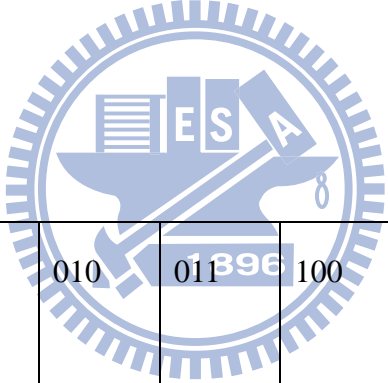
The time evolution of the cell states occur in discrete time steps with synchronous

update. Cellular automata are used to model several phenomena (e.g. pattern

formation) and to study various complexity theory concepts. Many types of cellular

automaton have been proposed, including random Boolean network and random

threshold networks.



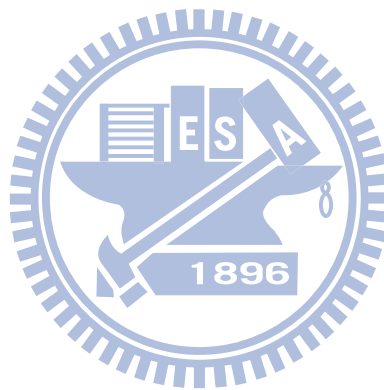
Three-cell block (t)	000	001	010	011	100	101	110	111
Center-cell (t+1)	0	1	0	0	1	1	1	0

Table 1. An update rule for a one-dimensional, two-state cellular automaton.

1.3.2 Preferential attachment

Preferential attachment [24] describes the preference of new nodes to link with more

connected nodes. Hubs are generated via ‘a rich-gets-richer’ mechanism consisting of growth and preferential attachment: the more connected a node is, the more likely new nodes will link to it, meaning that highly connected nodes acquire new links faster than their less connected peers. This mechanism ensures simultaneous the scale-free and hub properties.



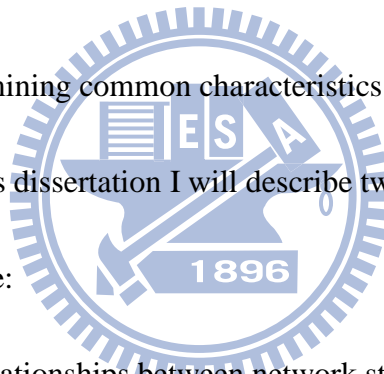
Chapter 2 Static Networks and Dynamic Process Characterization and analysis

No network in nature or technology is totally random—in other words, other non-random mechanisms shape their evolution. The universality of various topological characteristics, from degree distributions [25] to degree correlations [5, 22], motifs [16], and communities [2, 32], can be used as a springboard for studying diverse phenomena and making predictions. Network theory has therefore fundamentally reshaped our understanding of complexity. Even though researchers still lack a universally accepted definition of complexity[32], the role of networks in this area is obvious: all complex systems, from cells to the Internet and from social to economic, consist of an extra-ordinarily large number of components that interact via complex networks. We have long been aware of these networks, but only recently have we acquired the data and tools to probe their topologies, thus giving us a clear understanding of the strong impact of underlying connectivity on a system's behavior. As a result, no single approach to complex systems can succeed unless it exploits network topology.

The requirements of any new theory of complexity require an understanding of the

behavior of systems that we perceive as being complex. We must be capable of predicting how the Internet will respond to attacks and traffic jams, and how cells react to environment changes. Progress in this direction demands an understanding the dynamics of processes, a task made more difficult by the large number of dynamical phenomena-almost as many as complex systems. Examples include the biological study of reaction kinetics using metabolic networks, monitoring the flow of information on computer networks; and exploring the spread of viruses and ideas via social networks.

A major challenge is determining common characteristics among these diverse dynamical processes. In this dissertation I will describe two approaches for responding to this challenge:



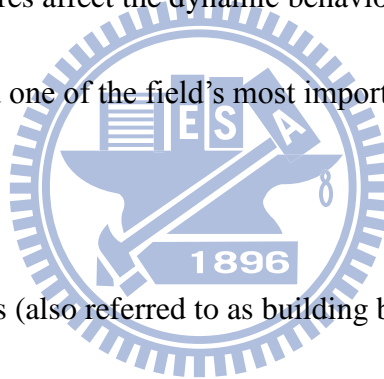
For static networks, find relationships between network structure and function. I will describe such relationships for two types of motifs in Chapter 3, and for network hierarchy in Chapter 6

For dynamic networks, find global rules during network evolution. I will discuss friendship evolution using three-rule model in Chapter 4 and epidemic dynamics with limited resources in Chapter 5.

2.1 Network Motif Detection

Commonalities have been found by complex network researchers in fields ranging from biology to social and computer sciences. Three global features in complex networks have been identified and investigated: highly clustered connections [1, 5], small-world properties [1, 21], and the scale-free phenomenon [5, 7]. Approaches based on quantitative and qualitative analyses of the topological properties of complex networks are serving as the basis for studying how the global features of network topological structures affect the dynamic behavior of networks [16, 33, 34].

This is currently considered one of the field's most important and challenging research topics [35, 36].



Some local structural motifs (also referred to as building blocks) reveal unique and statistically significant patterns when compared with random [16], biological [1], and food web [16, 26] motifs; all are perceived as containing important information.

However, the simple motifs of complex networks that are statistically significant but functionally unimportant are inadequate for investigating network functions and dynamic behaviors [16, 26]. In this dissertation, I will describe an algorithm that simultaneously (a) detects global features and local structures in complex networks,

and (b) identifies functionally and statistically significant network building blocks from complex networks [37].

When considering the global features and local structural motifs of biological networks, it is worth noting that link properties (weights) exert strong impacts on network functions and dynamic behaviors [38-42]. Examples include the role of weak links associated with the six degrees of separation (i.e., small-world) effect of interpersonal networks [40, 41], and the strength of predator-prey interactions that determine the stability of ecological communities [38]. Network researchers have reported that weighted values representing interaction strength can be assigned to all links (edges) in a real network [39, 43, 44]. I therefore considered network motif link strength in terms of bridge motifs (consisting of weak links only or a minimum of one weak link) and brick motifs (consisting of strong links only) (Fig. 6). Network motifs can be separated into two categories: bridge and brick. Using the three-point feed-forward motif as an example, it can be divided into two categories: a three-point feed-forward brick motif (left box) composed of three strong (red) links, and a three-point feed-forward bridge motif (right box) composed of at least one weak (blue) link and a maximum of two strong (red) links as Fig. 6 shows. Bridge motifs connect clusters and reduce the average degree of separation, while brick motifs exhibit the

phenomenon of local clustering in biological networks.

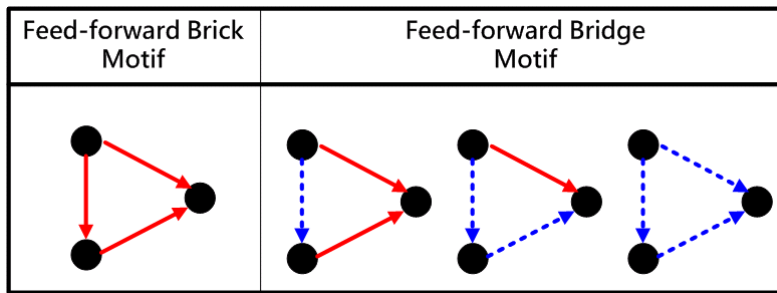
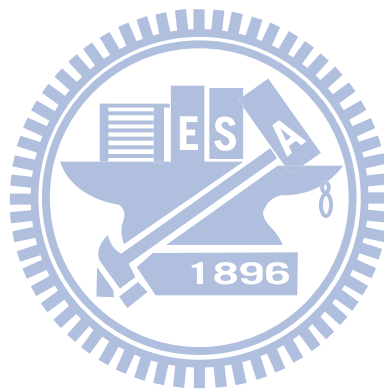


Fig. 6. Network motifs example.



2.1.1 General: Bridge and Brick Network Motif-Detecting

Algorithm

As shown in Figure 7, a link-weighted value that is dependent on the number of all possible paths between two linked nodes equals the summation of the reciprocal values of all possible path lengths except for the link itself. This is expressed as

$$weight(a,b) = \sum_i \frac{1}{length(path_i(a,b))} \quad (2.1)$$

where $path_i(a, b)$ indicates the i th path from node a to node b ; $path_i(a, b) \neq edge(a, b)$;

and $length(path_i(a, b)) \leq$ average network diameter. The length of one path represents

its total number of nodes.

$$\text{Average network diameter} = \frac{\sum_{a,b \in N \wedge a \neq b} \text{ShortestPath}(a,b)}{|N| \times (|N| - 1)} \quad (2.2)$$

$$\text{ShortestPath}(a, b) = \text{Min}(length(path_i(a, b)))$$

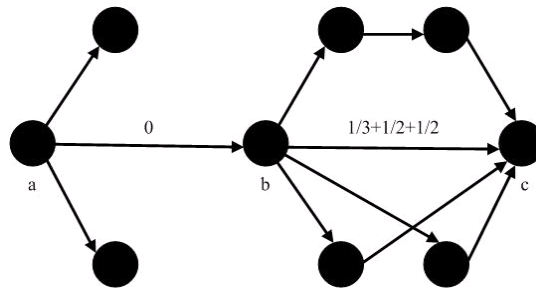


Fig. 7. Link-weighted value calculating example. The link-weighted value weight

(a, b) of edge (a, b) is 0 while weight (b, c)

This definition implies clustering, with any increase in the number of possible paths resulting in an increase in the clustering degree between two linked nodes.

Furthermore, the concepts and algorithms discussed in this dissertation are generalizable to non-directed networks. To ensure that the proposed method can be applied to any complex network, the link-weighted values calculated by the network motif detection method are derived from the number of all possible paths between two linked nodes within all network topological and local connection structures (no preset link quantity). This definition is similar to that of *betweenness* [43, 45]—effects resulting from the removal of network links. Accordingly, the proposed link-weighted value calculation method is assumed to represent the importance of each link in a real network [46, 47].

Also considered were the interactive strengths of individual links in a quantitative real network. To validate the proposal for weighted links, they were compared with quantitative links. However, interactive quantitative links are defined by category-specific functions such as proteins, genes, species, and so on. It is difficult to specify the overall impacts of these links on protein-protein interaction networks [48] and food webs. For example, the number of links between tigers and wild oxen does

not reflect the significance of their connection within an overall food web.

Furthermore, each complex network type has its own measure for interactive strength.

A switching algorithm (i.e., $A \rightarrow B$, $C \rightarrow D$ becomes $A \rightarrow D$, $C \rightarrow B$ if $A \rightarrow D$ and $C \rightarrow B$

do not exist) was used to create random networks according to any given degree

sequence [16, 26]. Results from previous studies indicate that these random networks

have the same number of nodes and edges, as well as node in-degrees (incoming

edges) and out-degrees (outgoing edges) that are identical to those of real networks.

Furthermore, randomized networks preserve the same number of appearances of all

($n-1$) node subgraphs as in real (original) networks [16]. The threshold that

determines the strength of an edge (link) is the mean weighted value of all edges in a

random network ensemble. Accordingly, 1,000 random networks were generated to

serve as a control. Edges were labeled “weak” when their weighted values in these or

real networks were smaller than the threshold minus a double standard deviation ($p =$

0.01); all other edges were labeled “strong.” Researchers can define criteria for strong

and weak links according to their own needs. Finally, all possible motifs were located,

and their distributions in real and random networks were compared.

Milo et al.’s method [16] for identifying bridge and brick motifs in complex networks

was expanded to include the following steps:

1. Calculate the weighted value of each link in a network of interest and an ensemble of random networks to calculate the significance of n -node subgraphs. The goal is to maintain the same number of appearances for all $(n - 1)$ node subgraphs as in the original network.
2. Label all weighted links in the network of interest and random network ensemble as “strong” or “weak” according to a benchmark of two standard deviations from the mean weighted value of all links in the ensemble. Links with weighted values below the benchmark are considered weak.
3. Identify all n -node bridge/brick subgraph types in the network of interest and random network ensemble.
4. Mark all n -node bridge/brick subgraph types by calculating their numbers in the network of interest and random network ensemble. Each n -node bridge/brick subgraph type is selected as a representative motif only if its frequency in the network of interest far exceeds its frequency in the ensemble.

These steps can assist research efforts to understand the functions and roles of identified motifs in a real network and to analyze the dynamic behaviors of complex networks. Regarding method robustness, the proposed approach emphasizes the global and local topological properties of each real network rather than the specific

functions of different network types.

Motif frequency can be used to measure levels of similarity between two networks of interest. In addition, it is possible to calculate the Z-scores for all bridge/brick motifs and significance profiles (SPs) in a network by expanding Milo et al.'s [26, 49, 50] methods. As shown in the following formula, $Z_{Score}(Bridge_i)$ represents the statistical significance of the i^{th} kind of bridge motif in a network:

$$Z_{Score}(Bridge_i) = \frac{N_{real}(Bridge_i) - \langle N_{random}(Bridge_i) \rangle}{STD(N_{random}(Bridge_i))} \quad (2.3)$$

where $N_{real}(Bridge_i)$ represents the time of appearance of the i^{th} type of bridge motif in a network, and $\langle N_{random}(Bridge_i) \rangle$ and $STD(N_{random}(Bridge_i))$ respectively represent the mean and standard deviation of the time of appearance of the i^{th} type of bridge motif in a randomized network ensemble. In the next equation, $SP(Bridge_i)$ is the vector of $Z_{Score}(Bridge_i)$ normalized to a length of 1. This normalization emphasizes the relative significance of the i^{th} type of bridge motif rather than the absolute significance. $Z_{Score}(Brick_i)$ and $SP(Brick_i)$ can be derived in the same manner:

$$SP(Bridge_i) = \frac{Z_{Score}(Bridge_i)}{(\sum Z_{Score}(Bridge_i)^2)^{1/2}} \quad (2.4)$$

$$Z_{Score}(Brick_i) = \frac{N_{real}(Brick_i) - \langle N_{random}(Brick_i) \rangle}{STD(N_{random}(Brick_i))} \quad (2.5)$$

$$SP(Brick_i) = \frac{Z_{Score}(Brick_i)}{\left(\sum Z_{Score}(Brick_i)^2\right)^{1/2}} \quad (2.6)$$

2.1.2 Specific: Bridge and Brick Network Motif-Detecting

Algorithm

The previous method we proposed has solved some problems in different domains successfully. However, since the concept of “neighborhood” is very useful for identifying the motifs or modules in biology. Another version of bridge and brick network motifs is proposed for this specific domain. To ensure that the concepts and methods described in this paper can be applied to any complex biological network, the link-weighted value $Link(v, w)$ for any edge between nodes v and w is expressed as its hypergeometric coefficient $C_{v,w}$ [51]. This value, which is frequently used to measure cluster enrichment and co-occurrence significance, is expressed as:

$$Link(v, w) = C_{v,w} = -\log \sum_{i=|N(v) \cap N(w)|}^{\min(|N(v)|, |N(w)|)} \frac{\binom{|N(v)|}{i} \times \binom{T - |N(v)|}{|N(w)| - i}}{\binom{T}{|N(w)|}} \quad (2.7)$$

where $|N(x)|$ is the neighborhood size of node x and T the total number of nodes in the biological network of interest. The summation in the hypergeometric coefficient $C_{v,w}$

can be represented as the probability of obtaining a number of mutual neighbors between nodes v and w at or above the observed number when the neighborhoods are independent. The hypergeometric coefficient $C_{v,w}$ is consequently defined as the negative log of this summation. Given the neighborhood sizes of the v and w nodes and the T total number of nodes in the biological network, the higher the value of $C_{v,w}$, the higher the number of overlapping neighbors between v and w —an indication that $Link(v, w)$ has a higher clustering coefficient. Otherwise, it does not belong to any specific cluster (Fig. 8). Different link definition differs between the general algorithm and the specific algorithm for detecting bridge and brick motifs, other parts are all the same.

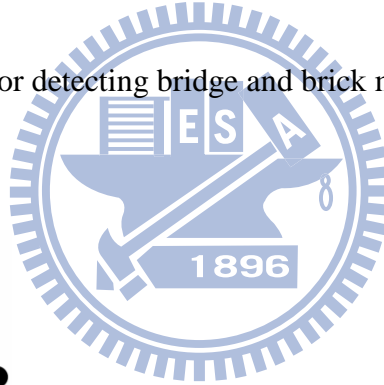
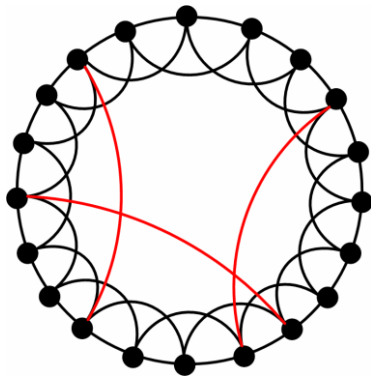


Fig. 8. The small-world model. Black signifies strong links and red weak links.

2.2 Social Network Simulation

Daividsen et al. [52] have proposed a two-rule acquaintance network evolution model.

The first rule addresses how people make new friends (via introductions or meetings-by-chance), and the second rule addresses how friendships are broken when one party dies. The model entails a fixed number of N nodes and undirected links between pairs of nodes representing individuals who know each other [53]. To reflect friendship weakening and strengthening, I added a “friend remembering” rule (Fig. 9). The model repeats the following three rules until the acquaintance network in question reaches a statistically stationary state:

Rule 1 (friend making). Randomly chosen individuals introduce two friends to each other. If this is their first meeting, a new link is formed between them. Randomly chosen persons with less than two friends introduce themselves to one other person at random. Note that the term “introduce” is used to describe meetings by chance as well as meetings via a common friend.

Rule 2 (leaving and arriving). At a p probability, a randomly chosen individual and all associated links are removed from a network and replaced by another person.

Accordingly, acquaintances can be viewed as circles of friends whose members can leave for reasons other than death and enter the circle for reasons other than birth.

Rule 3 (friend remembering). A certain number of friendships are updated, with the number depending on an update proportion b . This proportion and updating will be

explained in detail in the following two sections.

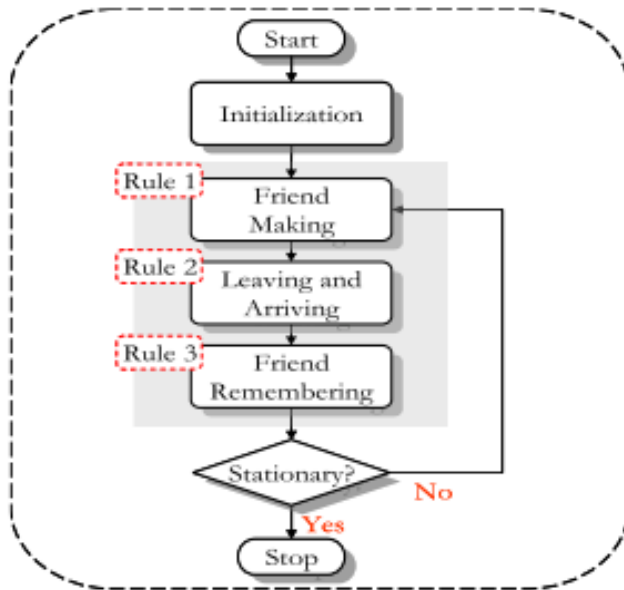


Fig. 9. Three-rule model flow diagram.

2.3 Epidemic Dynamics Analysis

The state transfer concept of SIS models adopted by Pastor-Satorras [54, 55] was applied as the core simulation model architecture. Parameters were incorporated to simulate behavioral and transformative results arising from agent interactions [56-60]. Each agent (node) in a complex network owns a set of properties and behavioral rules that are used to demonstrate the features and statuses of persons in social networks or computers connected to the Internet. A link between two nodes means that the connected agents have a close relationship or share a specific

interaction/communication channel. An infectious disease or computer virus can be transmitted via this link. At each discrete time step, the epidemiological state of each node is determined by its behavioral rules, original epidemiological state, neighbors' epidemiological states, infection rate, and recovery rate. As stated above, $\rho(t)$ is defined as the density of infected nodes present during time step t . When time step t becomes infinitely large, ρ can be represented as a steady density of infected nodes. A computational flowchart for the proposed simulation model is shown in Figure 10. A complex network $G(N, M)$ with N nodes and M links was constructed using the algorithm described previously prior to setting relevant parameters and attributes for the nodes involved in the simulation; discrete time t was set at 0. During simulations, nodes take turns interacting with neighboring agents for specified time intervals. Note that individual agent interactions do not result in immediate influences, and that simultaneous state changes only occur when all agents in a complex network complete their interactions. Accordingly, interaction sequences do not influence interaction processes or results.

At the beginning of each discrete time step, the usable economic resources of each agent v_i are reset to $R(v_i)$, meaning that all agents renew and/or receive supplemental resources. For example, the energy levels of most individuals are revived after a night

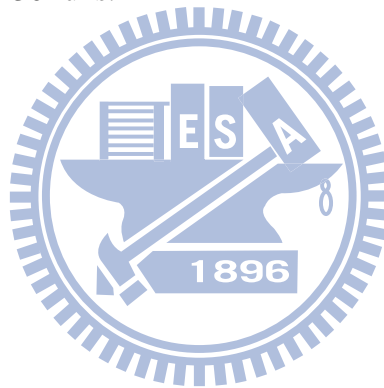
of sleep. In our later experiments, the statistical distribution of individual economic resources could be delta ($r_{Constant}$), uniform, normal, or power-law, as long as the mean value $\langle r \rangle$ satisfied:

$$\langle r \rangle = \frac{\sum_{i=1}^N R(v_i)}{N} = r_{Constant} \quad (2.8)$$

At each discrete time step, each v_j agent interacts with one agent selected from all of its Neighbors(v_i). After the interaction process is completed, agents v_i and v_j must have transmission costs $c(v_i)$ and $c(v_j)$ ($0 \leq c(v_i) \leq R(v_i)$ and $0 \leq c(v_j) \leq R(v_j)$) deducted from their respective economic resources, regardless of the interaction result. If $R(v_i) < c(v_i)$ after the interaction, agent v_i cannot interact with other neighbors because all of its resources have been used up. Otherwise, it repeats the interaction process by choosing another neighboring agent until its resources are exhausted.

Assume that infected and contagious agent v_i is adjacent to susceptible and infection-prone agent v_j . When the two agents come into contact, a combination of infection rate $RateInfect$, agent v_j 's resistance level, and a random number r determines whether or not v_j is infected by v_i . If the random number r is lower than the infection rate $RateInfect$, agent v_j 's epidemiological state becomes I (Infected). Simultaneously, infected agents are cured and become susceptible at a $RateReset$

recovery rate. Without a lack of generality, recovery rate `RateReset` can be assigned as 1, meaning if agent v_j is infected by other agents at discrete time step $t - 1$, it will recover and once again become susceptible at discrete time step t , since it only takes affect according to the definition of the infection disease propagation time scale. At the beginning of an infectious disease simulation, only ten individuals were given I status; all others were given S (Susceptible). During each time step, agents randomly interacted with several neighbors. All epidemic experiments discussed in this paper represent average values for 30 runs.



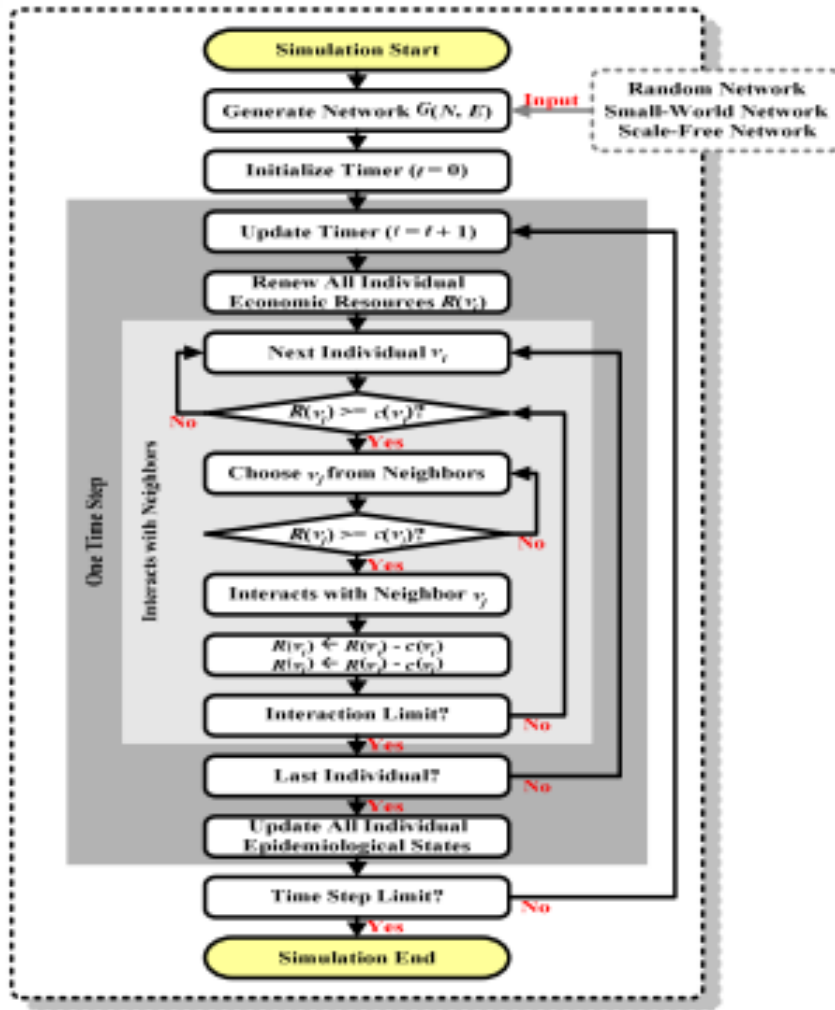


Fig. 10 Flowchart for a SIS epidemiological simulation model.

2.4 Abstraction Hierarchy

The proposed two-way method is considered novel because it emerges from top-down and bottom-up clustering algorithm synergy [61, 62]. Not only does it identify modules in a top-down fashion and construct a hierarchy implied in a complex network from the bottom up, it also produces network abstraction to different degrees

at different levels in the hierarchy. The method consists of three steps: (i) computing between-node proximity, (ii) extracting the backbone (represented by a spanning tree) from the network and using it to partition the network [28, 63], and (iii) generating an abstract network. Iteratively applying the same steps to a newly generated abstract network supports the discovery of an abstraction hierarchy within a complex network [2, 64].

2.4.1 Proximity Measure

Proximity between two nodes can be defined in many ways; since it affects resulting module formation, selecting an appropriate proximity function is very important.

Commonly used measures include Euclidean distance, correlation coefficient and cosine similarity [65, 66]. Module analysis is problem-dependent as stated earlier, in

this dissertation I investigate clustering based on network topology. Conventional

proximity measures are not applicable to clustering problems if network topology

represents the only available information—that is, Euclidean distance cannot be

calculated without node coordinates. Other proximity measures (e.g. edge

betweenness [45] and topological overlap [65, 67, 68]) have recently been proposed

and examined in studies of social, metabolism, protein-protein, and gene networks.

While some successful applications have been reported, they have at least two limitations: (a) edge betweenness of node pairs reflects the global characteristics in a network, but they suffer from high computational costs [64] and are affected by the network incompleteness and noise [32, 64]; and (b) since topological overlap is a local measure, it poses a challenge to identifying any module beyond a locally dense connectivity pattern [69].

Most proximity measures in current use do not take link direction or weight into account. Therefore, any directed weighted network is processed as an undirected unweighted one. To expand its applicability, I propose using a new proximity function for dealing with directions and weights. For the sake of simplicity, I will describe a directed weighted network of n nodes by an $n \times n$ adjacency matrix A , in which each element A_{ij} is the weight of the link from node i to j . A zero-valued weight ($A_{ij}=0$) indicates the absence of a link between those nodes. The proximity function $prox(i,j)$ from node i to j , $i \neq j$ is defined as:

$$prox(i, j) = A_{i, j} + \sum_{k}^{A_{i, k} \neq 0, A_{k, j} \neq 0} \left\{ \frac{A_{i, k}}{W_i^{out} - A_{i, j}} \times \frac{A_{k, j}}{W_k^{out}} \times \min(A_{i, k}, A_{k, j}) \right\},$$

$$W_i^{out} = \sum_m^n A_{i, m} \tag{2.9}$$

where W_i^{out} is the total of the weight of all outgoing node i links. The proximity

function considers not only the effects of common neighbors (i.e. node k), but also link direction and weight. It treats the direct link from node i to j differently than indirect paths between the same nodes through a common neighbor k . The weight of the direct link contributes to $\text{prox}(i, j)$, as indicated by the first term in the above equation. To calculate i -to- j proximity based on an indirect path from i to j by way of k , I divided the path into two sub-paths, from i to k and from k to j . Assuming on an indirect path one node does not always affect all its neighbors, but instead acts probabilistically. Thus, the probability that one node affects another (e.g. i affects k) is defined as the ratio of the link weight between them to the sum of the weights of all outgoing links from node i , not including the direct link from i to j . The probability of an indirect path from i to j by way of k is therefore the product of the probability of paths from i to k and the path from k to j . The proximity between i and j contributed by the indirect path $i-k-j$ is assigned to the probability times the minimum of A_{ik} and A_{kj} . In cases where there is more than one common neighbor of i and j , the sum of the proximity of each indirect path is used.

The examples shown in Fig. 11 illustrate our proximity function and compare it with a related measure, topological overlap [67, 68, 70]; both take common neighbors into consideration. The topological overlap measure T_{ij} between node i and node j ($i \neq j$) is

defined as follows:

$$T_{ij} = \frac{l_{ij} + a_{ij}}{\min(d_i, d_j) + 1 - a_{ij}}, \quad (2.10)$$

where l_{ij} is the number of common neighbors shared by node i and j , d_i is the degree of node i , and $a_{ij}=1$ if a direct link exists between i and j (otherwise, $a_{ij}=0$). The $1 - a_{ij}$ quantity in the denominator prevents the denominator from becoming zero in case where $\min(d_i, d_j)=0$. The inclusion of a_{ij} in the numerator is to make T_{ij} explicitly dependent on the direct link between i and j .

Given the network in Fig. 11A, $T_{ac}=0.5$ and $\text{prox}(a,c)=1$. To evaluate the effects of direct links, one direct link was added between nodes a and c (Fig. 11B). If we compare the network to a gene regulation model, a can be interpreted as a regulator, b as an intermediate gene, and c as a target. Since gene a can regulate gene c either directly, or through the intermediate gene b , the proximity between a and c in Fig. 11B should be higher than that in Fig. 11A. Increase in proximity were found in both measures—that is, $T_{ac}=1$ and $\text{prox}(a,c)=2$ vs. $T_{ac}=0.5$ and $\text{prox}(a,c)=1$. The network shown in Fig. 11C is different from the one in Fig. 11A in that node a and b both have more neighbors. Considering the network as a model of gene regulation, it means that gene a and gene b in Fig. 11C have more possible targets. Consequently, the influence

of gene a on gene c may be diminished. For node a and node c in Fig. 11C, $\text{prox}(a,c)$ decreases to $1/12$, which corresponds reasonably to the gene network model. In contrast, according to the topological overlap measure, the proximity between nodes a and c is 0.5 , which is the same as that in Fig. 11A. Topological overlap measure fails to distinguish between Fig. 11A and 11C. To complete the illustration, we added one direct link between a and c , and created the network shown in Fig. 11D. An increase in proximity results from either measure.

Even though the proposed proximity function is a local measure(similar to topological overlap), it shows better discrimination in network topology and requires less in computational costs than global measures such as edge betweenness. Incorporating the proximity function into a two-way module-finding and hierarchy-building strategy, it is possible to gather global characteristics and to detect the hierarchical structure of a network. We validated our approach using hierarchically nested random networks as in (11); a detailed description of the results will be given in later sections.

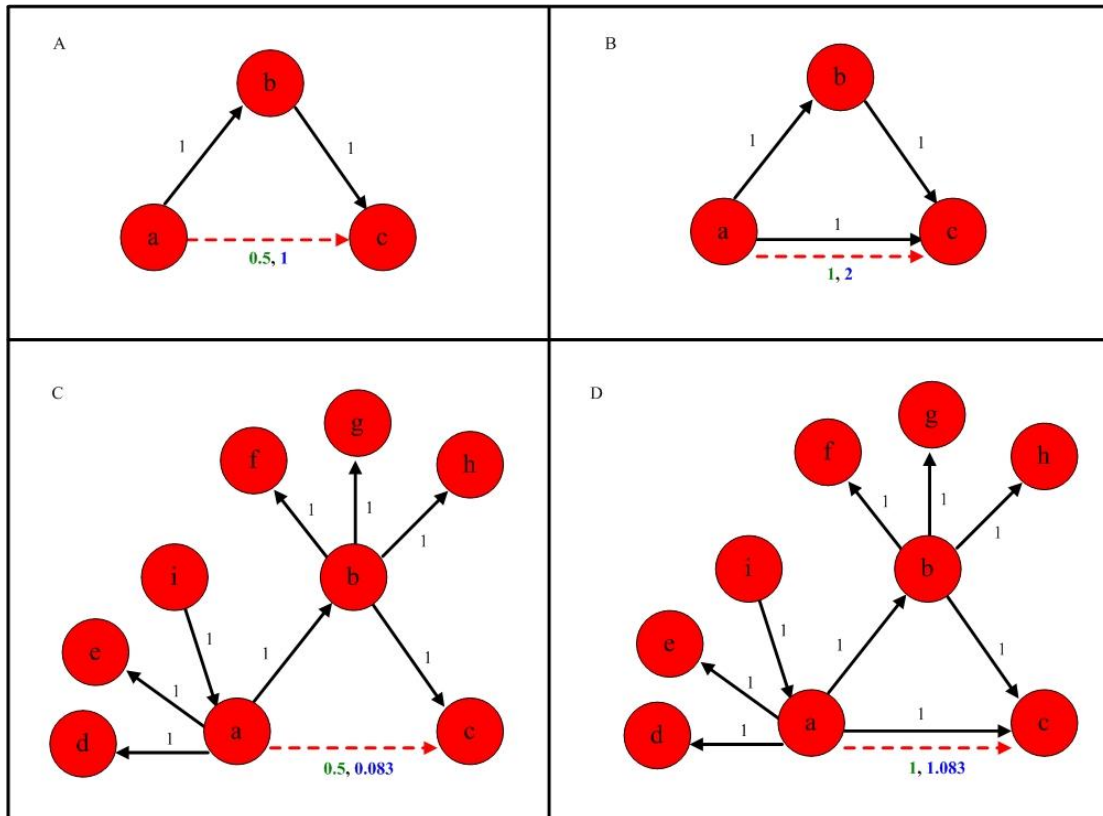


Fig 11. Four simple networks to illustrate proximity measures.

Extracting Network Backbone and Partitioning Network

An optimal solution for network partitioning (based on a criterion function) emerges after enumerating all possibilities, but it is computationally prohibitive for large networks. In response to this problem, a graph-theoretic approach to partitioning was adopted [71]. After computing the proximity between all node pairs, it is possible to build a maximum spanning tree that includes all network nodes, which are all connected with the maximum link proximity sums. Since links with less significant

proximities are discarded, the maximum spanning tree acts as the network backbone.

To reduce computational costs, partitioning is performed based on the maximum spanning tree instead of the original network.

Two subtrees can be obtained by removing one link from a tree, with each subtree representing one module. One tree can be partitioned into many subtrees (i.e.

modules/clusters) by repeating the same process on each subtree. Given the maximum spanning tree, resulting modules can be examined by removing one link from a

(sub)tree. A link is selected for removal if the $M=\{M_1,M_2,M_3,\dots,M_n\}$ set of modules

meets the following criteria after removal:

$$\forall M_a, M_b \in M, a \neq b \quad S_{intra}^{M_a} > S_{inter}^{M_a, M_b} \quad \text{and} \quad S_{intra}^{M_b} > S_{inter}^{M_a, M_b} \quad (2.11)$$

Where $S_{intra}^{M_k} = \sum_{i,j \in C_k} A_{i,j}$ is the sum of the proximity of each intralink within M_k , and

$S_{inter}^{M_a, M_b} = \sum_{i \in C_a, j \in C_b} A_{i,j}$ is the sum of the proximity of each interlink between M_a and M_b .

These criteria for modules are similar to those described in [64] and [72], except that

link weight (i.e., proximity) is considered instead of node degree. The simple example

network shown in Fig. 12 demonstrates the advantage of using the link weight criteria.

Without taking the weight into account, intuitively the most appropriate partition of

the network is to cut the link between node C and node F , and obtain two modules.

According to some previous module definitions [67, 72] that consider the degrees of

nodes only, the simple network will be partitioned in the way above. However, in practice, if the weight represents the significance of connectedness, the network should be considered as a whole. Our criteria for modules take weights into account; therefore, the network cannot be divided based on our criteria. In the case of an unweighted network, by treating each link as one with a constant weight, e.g. one, this simple network will be partitioned into two modules as expected according to our criteria. Without losing generality, this simple network demonstrates that our criteria for modules are more realistic, and can subsume the previous definitions of modules [72]. Note that the proximity sum is calculated for the network rather than the tree, thus preventing information loss. In the proposed model, the tree is only used for evaluating which nodes may form clusters, thus reducing the search space of the original network.

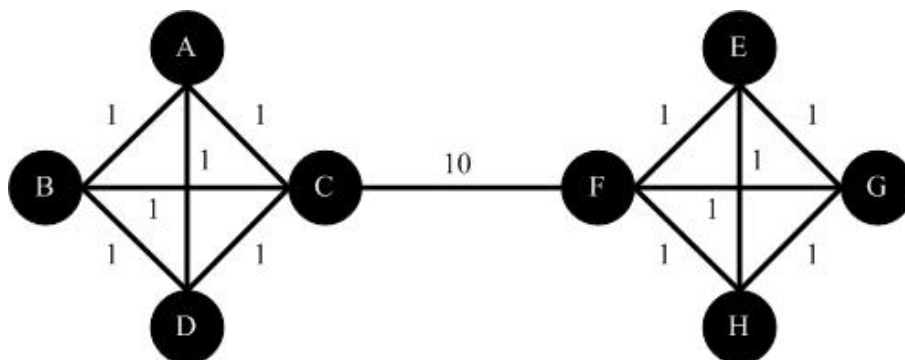


Fig 12. A simple undirected weighted network.

Pseudocode for the partitioning procedure is shown below, starting with a single module represented by a maximum spanning tree. The input includes the network Net in question; M_1 is its maximum spanning tree. The output consists of partitioning result in clusters.

Procedure Network_Partition (Net, M_1)

$M = \{M_1\}$ //M keeps the modules for further analysis

Repeat

{

Select a cluster $M_i \in M$, and remove M_i from M.

Put M_i into D. //D stores the final clusters

Put all the links of M_i in L_i .

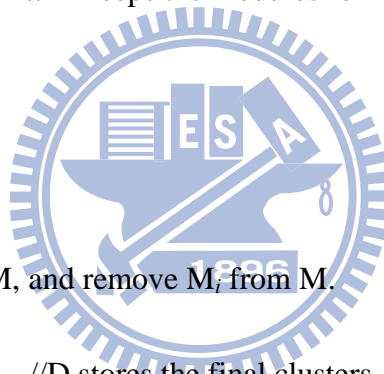
While (L_i is not empty)

{

Set the link in L_i with min proximity as l_{min} .

Remove l_{min} from L_i .

Generate two modules (i.e. subtrees) M_a and M_b by removing l_{min} from M_i .



Add M_a and M_b to M .

If M does not satisfy criteria [4]

{

Remove M_a and M_b from M .

Restore l_{\min} to M_i . //put the link l_{\min} back to the tree M_i

}

Else

{

Remove M_i from D . // because M_i is legally split into M_a and M_b

Break; //break out of While loop

}

} until M is empty.

Output D .

2.4.2 Network Abstraction

After the partition of the network, each module is treated as a supernode [7, 73] and a

network of the supernodes is viewed as an abstraction of the original network. An

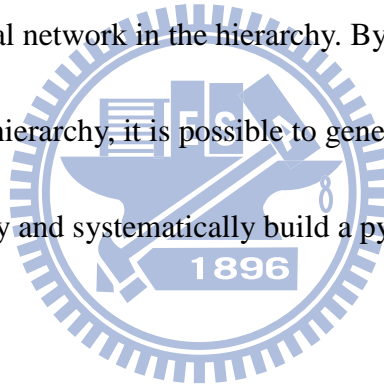
abstract network reveals the general framework of the original network without any

loss of principal characteristics. The proximity between a pair of supernodes (e.g.

module M_a and M_b) is defined as

$$prox_{super}(M_a, M_b) = \frac{1}{|M_a| \cdot |M_b|} \sum_{\forall m \in M_a, n \in M_b} prox(m, n) \quad (2.12)$$

where $|M_a|$ is the number of nodes in module M_a . Proximity between all possible supernode pairs are computed and normalized to a z-score. Links that have z-scores below a pre-set threshold are considered insignificant and therefore discarded. The resulting supernode network (an abstraction of the original network) is placed one level higher than the original network in the hierarchy. By repeating the same process with other networks in the hierarchy, it is possible to generate additional abstract networks and to consistently and systematically build a pyramid of abstraction from the bottom up(Fig. 1).



Chapter 3 Network motif Experiments

Researchers are making progress toward defining organizing principles that govern the formation and evolution of complex biological networks. Considered a major challenge in computational system biology, predicting network behaviors [74] and functions requires the identification of functionally and statistically important motifs. To understand their structural organizing principles and evolutionary mechanisms, bridge motifs can be defined as consisting of weak links only or at least one weak link and multiple strong links; brick motifs can be defined as consisting of strong links only. Next, an algorithm is proposed for performing two simultaneous tasks: detecting global statistical features and local connection structures in biological networks, and locating functionally and statistically significant network motifs.

3.1 General: Bridge and Brick Network

Motif-Detecting Algorithms

Commonalties have emerged from studies of complex networks in fields ranging from biology to social and computer sciences. Three global features in complex networks have been identified and investigated, including highly clustered connections [21, 39,

75], small-world properties [21, 75-78], and the scale-free phenomenon [1, 24, 39, 79].

Approaches based on quantitative and qualitative analyses of the topological properties of complex networks are being utilized to study how the global features of network topological structures affect the dynamic behavior of networks [1, 39, 80-84] 39-40]. This is currently considered one of the field's most important and challenging research topics [1, 39].

Some local structural motifs (building blocks) reveal unique and statistically significant patterns when compared with random [16, 80, 85-90], biological [16, 87], and food web [16, 38] motifs; all are thought to contain important information.

However, simple motifs of complex networks that are statistically significant but functionally unimportant are clearly inadequate for investigating network functions and dynamic behaviors [16, 82, 88, 90-93]. An algorithm is therefore proposed to perform two tasks: simultaneously detect global features and local structures in complex networks, and identify functionally and statistically significant network building blocks from complex networks.

When considering the global features and local structural motifs of biological networks, it is worth noting that link properties (weights) exert strong impacts on network functions and dynamic behaviors [38-40]. Examples include the role of weak

links in the six degrees of separation (i.e., small-world) effect of interpersonal networks [40, 41] and the strength of predator-prey interactions that determine the stability of ecological communities [38]. Network researchers have reported that a weighted value representing interaction strength can be assigned to each link (edge) in a real network [43, 44]. I therefore took into consideration network motif link strength in terms of two categories: bridge motifs (consisting of weak links only or a minimum of one weak link) and brick motifs (consisting of strong links only) (Figs. 3 and 6). Bridge motifs connect clusters and reduce average degree of separation, while brick motifs exhibit local clustering in biological networks.

Validation experiments were performed to confirm weighted link and network motif definitions. Due to the links' non-directional characteristic, only two kinds of motifs were identified for the three-node scenarios: ID = 8 and ID = 13 (Fig. 3). Four well-known types of theoretical complex networks with specific topological properties were examined to validate the proposed algorithm: regular, scale-free, random, and Watts and Strogatz's small-world (Table 2) [21]. Due to their small-world properties, more bridge than brick motifs were found in scale-free and random networks. Regular networks with a Moore neighborhood structure only contain brick motifs, due to that structure's high clustering property (minus any

shortcuts). Watts and Strogatz's small-world networks are formed by rewiring 1% of the links of regular networks containing only a few bridge motifs; when more than 5% are rewired, bridge motifs outnumber brick motifs (Fig. 13). It is therefore suggested that bridge motifs indicate the presence of small-world properties and brick motifs the presence of local clustering properties as follows:

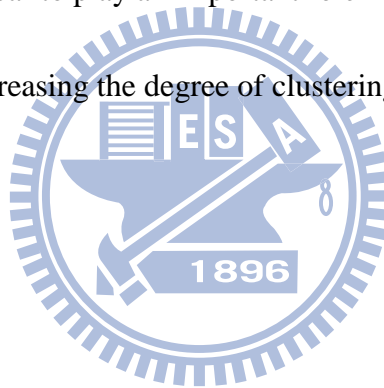
1. Regular. The Moore neighborhood concept was applied to a two-dimensional lattice, with each node linked to its eight adjacent cells [94]. For this type of network only brick motifs were found. To maintain the same in- and out-degree distributions in random and regular networks, individual nodes in random networks can link with any other cell except their eight adjacent cells. As clustering in a random network decreases, the threshold of the weighted value of its links also decreases. Therefore, all links in regular networks are strong (exclusively brick motifs).

2. Scale-free. Here the degree of distribution (i.e., the number of edges per node) obeys a long-tailed power-law distribution, in which the majority of nodes have only a few links, but a small number of nodes have many links. Scale-free networks were found to be composed of many bridge motifs and a small number of brick motifs consisting of nodes with high degrees of separation.

3. Random. As predicted, a dominant motif did not emerge from a comparison of

1,000 random networks. Accordingly, random networks served as a successful null hypothesis for the proposed algorithm.

4. Small-world. Links were rewired in two-dimensional regular networks with Moore neighborhood structures; 0.01, 0.05, 0.1 and 0.5 percent of all links were rewired. In the 0.01 trial, some of the brick motifs became bridge motifs. As the rewiring percentage increased, the number of bridge motifs increased and number of brick motifs decreased. At a rewiring ratio of 1, small-world networks change into random networks. Brick motifs appear to play an important role in reducing the degree of separation, as well as in increasing the degree of clustering in scale-free networks.



3.1.1 Validation

We performed validation experiments to confirm the definitions of weighted links and network motifs. Due to the links' non-directional characteristic, only two kinds of motifs were identified for the three-node scenarios: ID = 8 and ID = 13 (Fig. 3). We looked at four well-known types of theoretical complex networks with specific topological properties to validate our algorithm: regular, scale-free, random, and Watts and Strogatz's small-world (Table 2) [21]. Due to their small-world properties, we

found more bridge than brick motifs in scale-free and random networks. Regular networks with a Moore neighborhood structure only contain brick motifs due to the structure's high clustering property (minus any shortcuts). Watts and Strogatz's small-world networks are formed by rewiring 1% of the links of regular networks containing only a few bridge motifs; when more than 5% of the links are rewired, bridge motifs outnumber brick motifs (Fig. 13). We therefore suggest that bridge motifs indicate the presence of small-world properties and brick motifs the presence of local clustering properties as follows:

Regular. We applied the Moore neighborhood concept to a two-dimensional lattice, with each node linked to its eight adjacent cells [94]. For this type of network we found brick motifs only. To maintain the same in- and out-degree distributions in random and regular networks, individual nodes in random networks can link with any other cell except its eight adjacent cells. As clustering in a random network decreases, the threshold of the weighted value of its links also decreases. Therefore, all links in regular networks turn out to be strong (exclusively brick motifs).

Scale-free. Here the degree of distribution (i.e., the number of edges per node) obeys a long-tailed power-law distribution in which the majority of nodes have only a few links, but a small number of nodes have many links. We found that scale-free

networks are composed of many bridge motifs and very few brick motifs consisting of nodes with high degrees of separation.

Random. As predicted, we failed to find a dominant motif during our comparison of 1,000 random networks. Accordingly, random networks served as a successful null hypothesis for our algorithm.

Small-world. We rewired links in two-dimensional regular networks with Moore neighborhood structure using rewiring percentages of 0.01, 0.05, 0.1 and 0.5 of all links. In the 0.01 trial we found that some of the brick motifs became bridge motifs.

As the rewiring percentage increased, the number of bridge motifs increased and number of brick motifs decreased. At a rewiring ratio of 1, small-world networks change into random networks. Brick motifs appear to play an important role in reducing the degree of separation and increasing the degree of clustering in scale-free networks.

Table 2. Bridge and brick subgraph frequencies in four complex network categories (for validation purposes).

Category	Nodes	Edges	Subgraph Type	ID	N_{Real}	$N_{\text{Random}} \pm \text{STD}$	Z_{Score}
Regular	900	7200	Bridge	8	0	24983.2 \pm 39.0	-640.61
			Brick	8	14400	40.8 \pm 17.4	824.81
			Bridge	13	0	0.0 \pm 0.0	0.00
			Brick	13	3600	58.6 \pm 8.2	430.13
Scale-Free	900	1800	Bridge	8	4355	4099.7 \pm 53.7	4.75
			Brick	8	45	258.9 \pm 47.1	-4.54
			Bridge	13	2	7.0 \pm 2.6	-1.95
			Brick	13	0	8.8 \pm 3.5	-2.54

Random	900	1800	Bridge	8	1229	1226.1±27.9	0.11
			Brick	8	537	536.9±27.8	0.01
			Bridge	13	0	0.4±0.7	-0.64
			Brick	13	0	0.6±0.7	-0.81
WS Small-World #1 (rewiring % = 0.01)	900	7200	Bridge	8	2399	25029.7±38.0	-595.92
			Brick	8	12573	58.3±18.6	674.60
			Bridge	13	320	7.6±2.8	113.47
			Brick	13	3111	51.5±7.1	430.65
WS Small-World #2 (rewiring % = 0.05)	900	7200	Bridge	8	9434	24713.2±73.5	-207.80
			Brick	8	8100	656.3±64.5	115.49
			Bridge	13	991	25.1± 5.3	182.65
			Brick	13	1681	35.0±6.1	268.59
WS Small-World #3 (rewiring % = 0.10)	900	7200	Bridge	8	13386	24047.0±111.4	-95.69
			Brick	8	6089	1519.0±99.9	45.73
			Bridge	13	1063	30.9±4.9	209.23
			Brick	13	1029	30.7±5.6	179.17
WS Small-World #4 (rewiring % = 0.50)	900	7200	Bridge	8	22649	22935.4±148.5	-1.93
			Brick	8	3973	4244.9±153.7	-1.77
			Bridge	13	213	56.2±8.9	17.64
			Brick	13	47	17.7±4.1	7.11

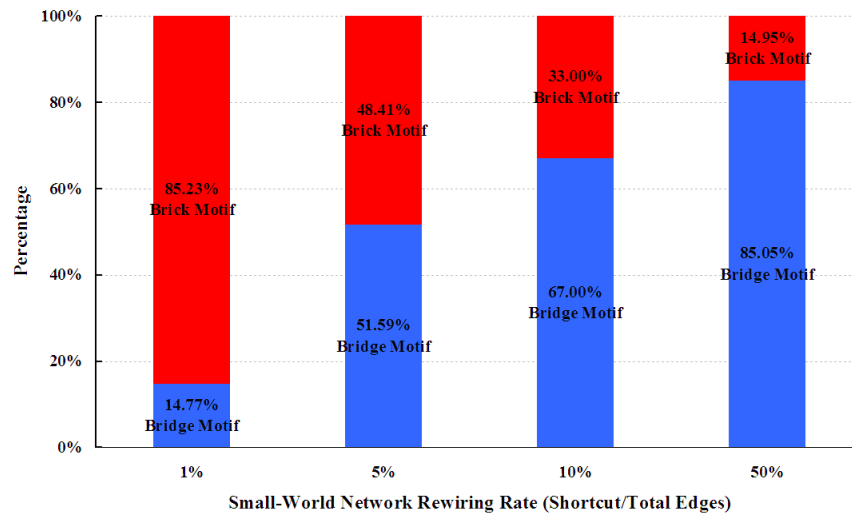


Fig. 13. Percentages of bridge and brick motifs in small-world networks according to different rewiring ratios.

3.1.2 Experiments

The proposed method was applied to several biochemistry (transcriptional gene regulation) and ecology (food web) networks to identify bridge and brick network

motifs. Networks and sources are listed in Table 3. All data and programs (including source code) are available online at <ftp://www.csie.cgu.edu.tw/bbm/>.

Several engineering (electronic circuit) and social networks (Table 3) were used to demonstrate that the proposed motif detection method is both general-purpose and robust. It was also compared with Milo et al.'s [26] original method for complex network analysis. In electronic circuits consisting of digital fractional multipliers (data from an ISCA89 benchmark) [26], nodes represent logic gates and flip-flops and edges represent directed electronic transmission paths. Experimental results indicate that the s208, s420, and s838 electronic circuit networks contain significant numbers of bridge motifs. Here the low degree of clustering is considered trivial because designers often try to simplify connection structures and numbers of electronic components [77]. The identified feedback bridge motif (consisting of weak-tie links only) fulfills this requirement as described by Kundu et al. [95] (ID = 9) (Figs. 14, Table 3). As its name implies, the feedback bridge motif indicates the existence of a feedback structure without redundancy in the three above-named electronic circuits—again proven by Kundu et al. [95], who also reported that redundant circuits seldom appear in simple electronic circuits such as s208, s420, and s838. However, they also note that redundant wires and components are frequently added to more

complex electronic circuits (e.g., s15850, s35932, s38417 and s38584) to prevent accidental system failures. The over-simplification of electronic circuits can result in large numbers of errors [77] or complete system breakdowns when one component fails. Accordingly, it is necessary to add an appropriate level of redundancy as a means of bypassing failed components or substituting for the original path [77, 95]. Strong-tie links represent alternative paths and weak-tie links represent simplified electronic circuits. Combined, simplification and duplication help prevent unexpected system breakdowns.

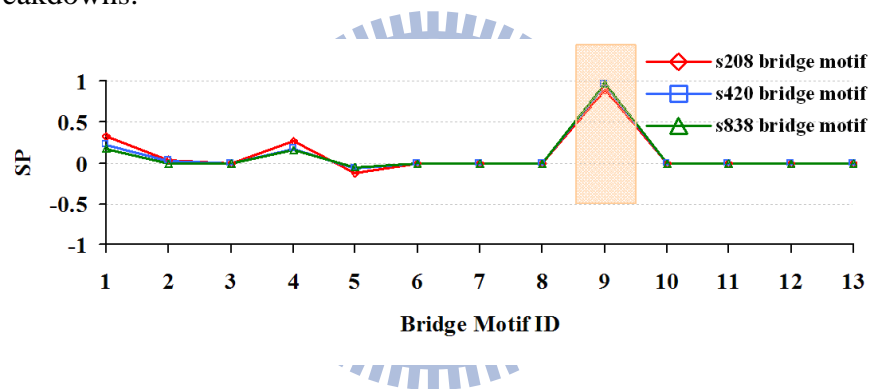


Fig. 14. Bridge motif ratio profiles for three electrical circuits (s208, s420 and s838).

In the two social networks that were analyzed, nodes represent individuals in a group and edges represent positive sentiments directed from one group member to another based on responses to questionnaire items. Similar characteristics were found between

two networks, one consisting of prison inmates ($N = 67$ nodes, $E = 110$ edges) and the other of college students in a leadership course ($N = 32$, $E = 96$). The inmates responded to the question, “Who are your closest friends in your cellblock?” The students were asked to name three classmates they would invite to serve on a committee (correlation coefficient $c = 0.92$ to 0.96 [96, 97]). According to Milo et al.’s [16] methods, both social networks belong to the same superfamily. Strong similarities between the two networks were also identified according to the triad significance profile (TSP) of bridge motifs ($c = 0.92$) (Fig. 15, Table 3), but not according to the TSP of brick motifs ($c = 0.6$) (Fig. 16, Table 3). Also found was a significantly higher number of bridge motifs (i.e., more “nodding acquaintances”) in the prisoner network. The significantly larger number of brick motifs in the leadership class network indicates that small, strong groups are easily formed. The bridge and brick motifs can be used to further analyze network topological structures, functions, and differences.

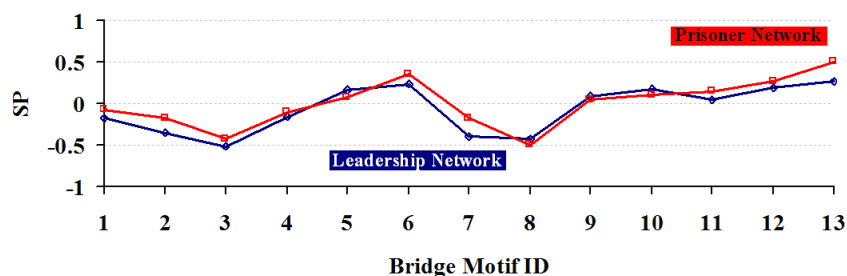


Fig. 15. Bridge motif ratio profiles for two social networks.

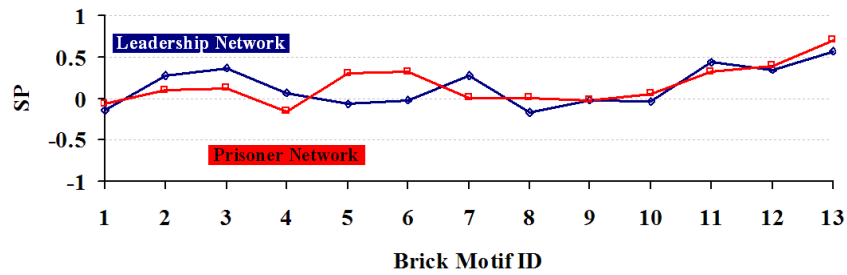


Fig. 16. Brick motif ratio profiles for two social networks.

In gene regulation networks for one bacteria (*Escherichia coli*) and one eukaryote (the yeast *Saccharomyces cerevisiae*) [26], each node represents a gene or operon that encodes a transcription factor (TF); edges denote the TFs themselves. Many TFs are encoded within operons, therefore directed links represent direct transcriptional modulation from a TF to an operon or from a TF-contained operon to another operon [26]. More bridge than brick subgraphs were found in both networks (they are not called motifs until they reach statistical significance). Furthermore, the two transcription networks had the same feed-forward bridge motif (ID = 5), indicating that the transcription networks have, at minimum, non-replaceable interactions without intermediate interactions with other genes (Fig. 17 and Table 3). This suggests that the weak-tie link that provides a unique path for controlling the signal

exerts a significant impact on the signal processing function of transcription networks [26, 80]. When analyzing the relationship between coherent (incoherent) FFLs and brick (bridge) FFLs, I identified *E. coli*'s 34 coherent and 8 incoherent FFLs (Table 3) [34, 98, 99]. Accordingly, differences in coherent (incoherent) FFL frequencies cannot be explained simply in terms of the relative abundances of bridge and brick motifs in a network.

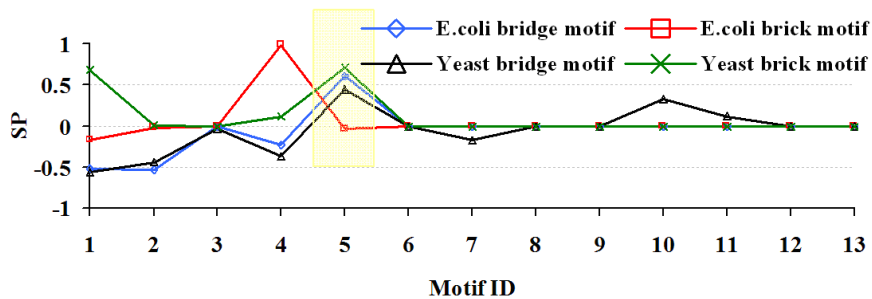


Fig. 17. Brick-bridge motif ratio profiles for two regulation networks (one bacteria and one eukaryote).

In the seven analyzed food webs [100], nodes represent groups of species and edges connect predator and prey nodes. Two studies have shown that strong interactions (similar to the definition of weak-tie links used here) between two consecutive levels of a trophic chain have a significant effect on food web stability and dynamics [38, 101]. A strong interaction indicates a strong predator preference for one prey species and a low potential for intermediate species—a phenomenon that supports the

proposal that weak-tie links exert certain impacts on food webs. Also in the seven food webs, the numbers of bridge motifs were significantly higher than the numbers of brick motifs, especially feedback (ID = 5) and three-point chains (ID = 2) (Fig. 18, Table 3). This confirms Jordi's [38] claim that these two motifs exert significant impacts on ecosystem food webs. The reason why ecosystems containing these two types of bridge motifs easily become unbalanced is likely because they have many weak links—in other words, it is difficult to find substitute nodes or links for the purpose of preserving ecosystem stability.

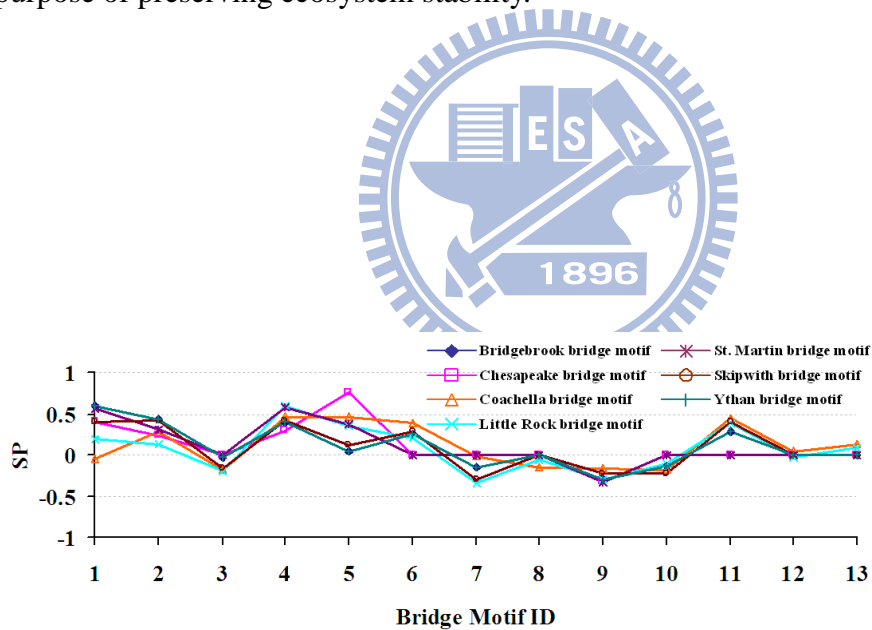


Fig. 18. Bridge motif ratio profiles for seven food webs.



Table 3. Brick and bridge motifs in fourteen real world networks, including edge and node definitions, network sizes, and references.

Category	Common Feature	Directed Network	Nodes	Links	Motif Type	ID	N_{Real}	$N_{Random} \pm STD$	Z_{Score}
Gene Regulation (transcription) [16, 102]	Directed graph in which nodes represent genes and edges are directed from one gene to another, regulated by the transcription factor.	E.coli	424	519	Bridge	5	42	7.5±3.1	11.14
		Yeast	688	1079	Bridge	5	67	13.8±3.8	14.04
Social [96, 97]	Directed graph in which nodes represent people and edges indicate friendships between two persons.	Leader	32	96	Brick Brick	7 11	38 5	22.1±9.5 1.5±1.3	1.67 2.59
		Prisoner	67	182	Bridge Brick	6 12	11 5	2.0±1.4 0.5±0.7	6.42 6.26
Food Webs [100]	Seven different ecosystems. Directed graph in which nodes represent groups of species and edges connect predator and prey nodes.	LittleRock	92	984	Bridge	11	93	41.3±6.2	8.33
		Ythan	83	391	Bridge	2	1182	850.1±86.0	3.86
		St. Martin	42	205	Bridge	5	244	180.4±20.0	3.18
		Chesapeake	31	67	Bridge	5	21	11.2±4.0	2.42
		Coachella	Bridge	2	275	192.5±14.8	5.57		
			Bridge	4	252	110.3±15.1	9.38		
			Bridge	6	110	68.1±5.3	7.84		
		Skipwith	Bridge	13	10	6.2±1.4	2.83		
Bridge	2		181	140.1±11.3	3.63				
B.Brook	Bridge	4	234	115.2±33.4	3.56				
	Bridge	1	266	123.5±31.2	4.57				
Electrical Circuits [77]	ISCAS89 benchmark set of sequential logic electronic circuits.	s208	122	189	Bridge	9	10	0.90±1.0	9.23
		s420	252	399	Bridge	9	20	0.9±0.9	20.13
	Directed graph in which nodes represent logic gates and flip-flops.	s838	512	819	Bridge	9	40	0.9±1.3	30.2

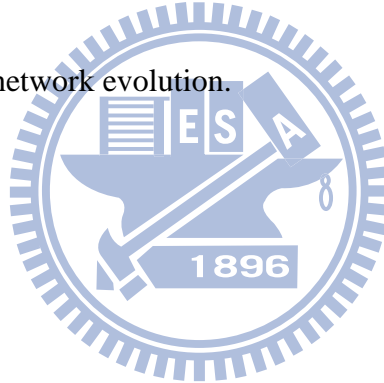
3.1.3 Conclusion

According to the definitions of weighted links and network motifs used in this study and the results of the validation experiments using theoretical complex networks, the presence of bridge and brick motifs in a network is closely associated with network topological structures (especially local connections), but not with network size (i.e., number of nodes). In summary, three experimental predictions were tested to verify

the importance and function of bridge and brick network motifs: (a) whether regular networks with Moore neighborhood structures only contain brick motifs due to the structure's strong clustering property; (b) whether the number of bridge motifs increases and the number of brick motifs decreases as rewiring percentages increase in regular networks, with the rewiring process contributing to the formation of networks that exhibit small-world and clustering properties; and (c) whether the combination of more bridge motifs and fewer brick motifs means that a network is less prone to cluster formation.

The proposed method combines two measures, each with its own merits—that is, determining the topological properties of links in real networks and identifying statistically significant motifs in real networks. The combined measures can be used to explore the functions and roles of real network motifs. To locate statistically significant network motifs, Milo et al. [16] propose comparing the real network in question with suitably randomized networks, then selecting patterns (subgraphs) that appear at significantly higher frequencies in the real network. Compared to Milo et al.'s approach, the method described in this chapter simultaneously detects global features and local structures in complex networks and locates functionally and statistically significant network motifs. It is suggested that the combination of these

two methods can (a) assist in locating motifs; (b) help researchers find clusters between bridge motifs and within the brick motifs of complex networks for the purpose of identifying real network functions, behaviors, and similarities; and (c) provide global and local views of the real network in question. Most network motif functions can be identified via network topological structures. Combining a motif structure with its function can help identify complex network properties. Motifs with special topological structures reveal the global features of real networks and significant local structural patterns. This information can help researchers working with design principles and network evolution.



3.2 Specific: Bridge and Brick Network

Motif-Detecting Algorithms

The above-described method successfully addresses some problems in specific domains. However, in biology, the concept of “neighborhood” is especially useful for identifying motifs or modules. Accordingly, in this section, I will propose another version of bridge and brick network motifs in the context biology.

3.2.1 Validation

To validate the respective roles of weak and strong links, equal percentages of each (as well as random links) were removed. For *E. coli* and *S. cerevisiae* [16, 26], the greater the number of strong links removed, the lower the clustering coefficient relative to the randomly removed links. In contrast, the greater the number of weak links removed, the higher the clustering coefficient relative to the randomly removed links (Figs. 19 and 20). Note that the average clustering coefficient increases when weak links are removed—that is, when the clustering coefficient of a weak link’s end node is calculated, its neighbors do not include the same link’s other end node. The average coefficient increases after the weak links are removed because the two end nodes do not share a large number of common neighbors. The average degree of

separation in the network after removing links was not computed, since a network might become broken and disconnected after a link is removed, and the definition of average degree of separation is based on a connected network. Note that the proposed approach is insensitive to data errors: significant network motif sets in the two gene regulation networks do not change a great deal, even when 40% of their edges are removed (Figs. 21 and 22). All altered results (red curves) shown in Figures 21 and 22 represent average values for 30 runs. The sensitivity analysis results confirmed significant similarities between the original and altered networks after randomly removing 40% of their links. According to the triad significance profile (TSP) [26] of brick motifs, the original and altered networks belong to the same superfamily. As shown in Figure 23, link weight distribution is extremely polarized (either 0 or >2), which matches the criterion for distinguishing between strong and weak links (i.e., mean weighted value $\text{LinkAVG} = 0.9$ and standard deviation $\text{LinkSTD} = 0.04$ for all links in 1,000 randomized networks). In most cases random networks have many more weak links than strong links. At least one researcher has suggested that high degree of clustering is a generic feature of biological networks [65].

The link property is a good indicator of cellular function robustness. The simplest strategy for protecting against the failure of a specific component is to provide

alternative ways to perform that component's function. At the molecular level, this backup strategy (or genetic buffering) [103, 104] can be carried out by duplicate genes with identical roles or by different genes that constitute an alternate but functionally overlapping path [83]. Researchers can use brick motifs to explore (a) identical genes that diverge functionally, (b) reasons why the biological networks of unreliable elements still perform reliably [1], and (c) the degeneracy phenomenon [65].

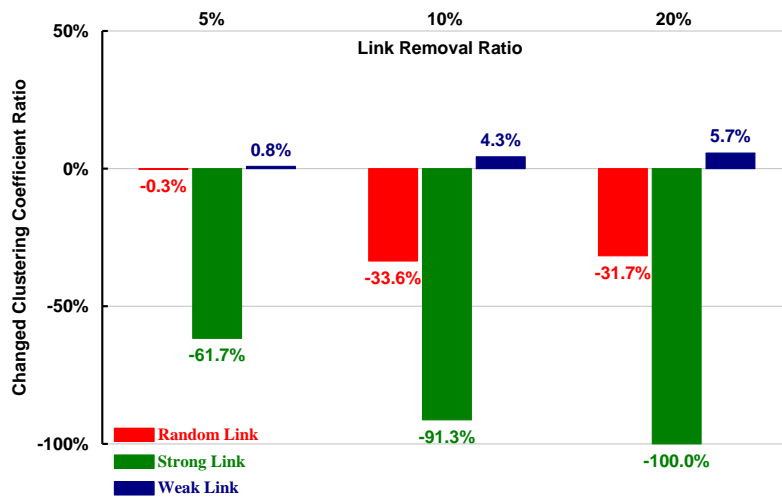


Fig. 19. Relationships between clustering coefficients and different removal ratios for three *E. coli* link types. Red, random; green, strong; blue, weak.

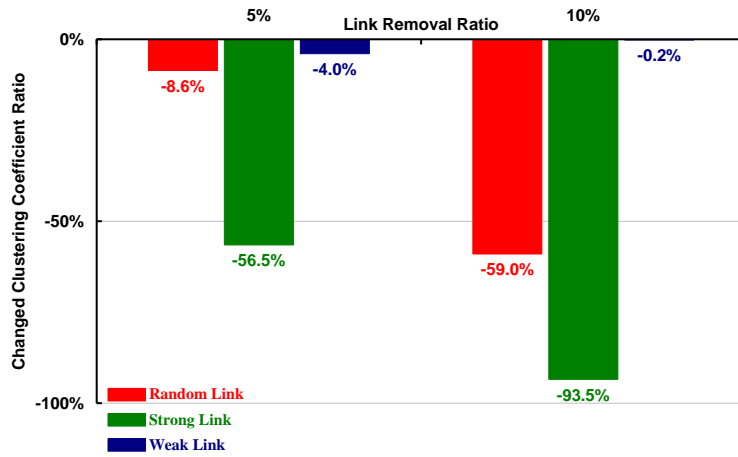


Fig. 20. Relationships between clustering coefficients and different removal ratios for three *S. cerevisiae* (yeast) link types. Red, random; green, strong; blue, weak.

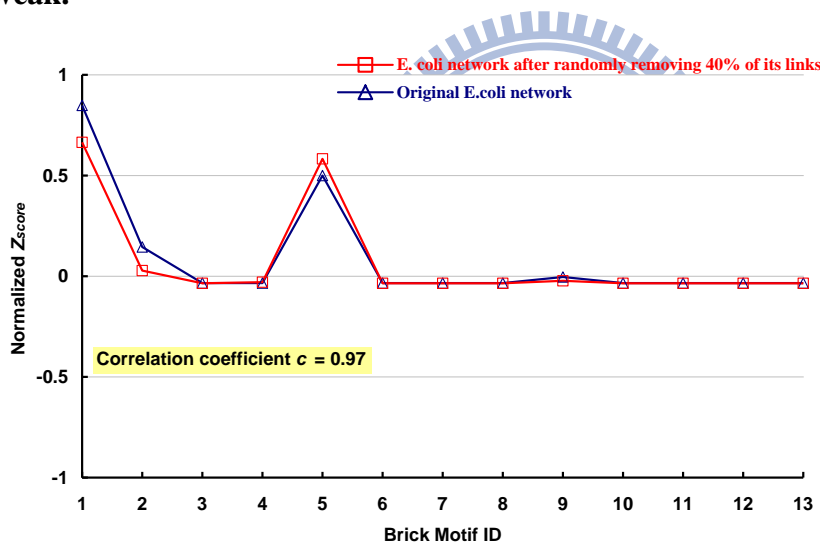


Fig. 21. Comparison of original (blue curve) and altered (red curve) brick motif ratio profiles for *E. coli* after randomly removing 40% of its links. Altered results represent average values for 30 runs.

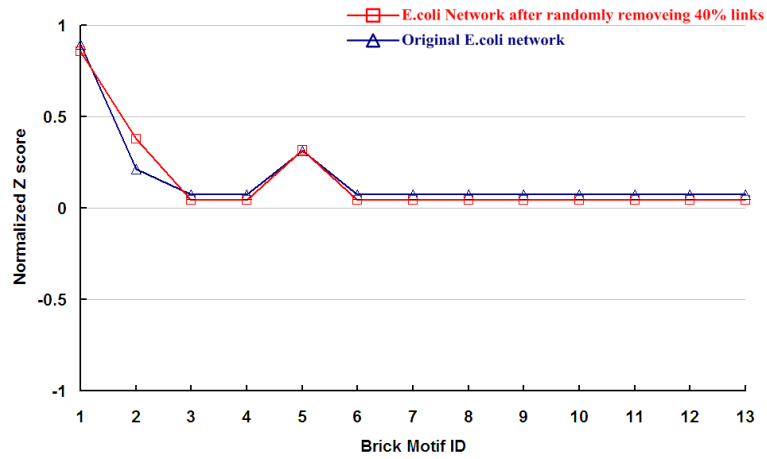


Fig. 22. Comparison between original brick motif ratio profiles and altered brick motif ratio profiles for *S. cerevisiae* (yeast) after randomly removing 40% of its links.

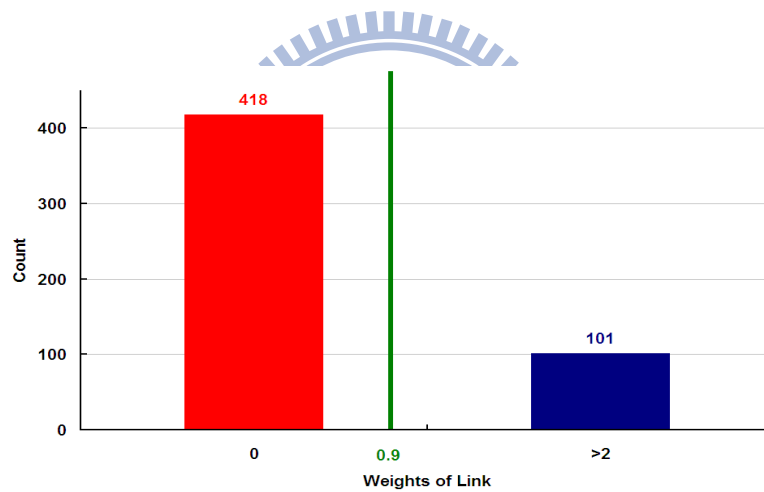


Fig. 23. Distribution of link weights in *E.coli*. Average mean and standard deviation of link weights for randomized networks were calculated as 0.90 ± 0.04 .

3.2.2 Experiments

The proposed method was applied to *E. coli* (bacteria) and *S. cerevisiae* (yeast)

transcriptional gene regulation networks [26]. Network and source data are listed in Tables 4 and 5. In both networks, nodes represent operons (i.e., one or more genes transcribed on the same mRNA [26]) and directed links represent transcriptional regulatory relationships between operons that encode transcription factors (TFs) and operons regulated by TFs. Many v-out and feed-forward loop (FFL) brick motifs were observed in both *E. coli* and *S. cerevisiae* (ID = 1 and 5, respectively) (Table 5). The FFL (a three-gene subgraph) is composed of two input transcription factors, one regulating the other and both jointly regulating a target gene [80]. The observation that FFL bridge motifs do not exist in either network supports previous findings indicating that most motifs do not function in isolation, but overlap with known biological functions [24, 42, 100]. Specifically, one FFL motif cluster overlaps with the flagella motor module and another contains a significant number of elements responsible for regulating the *E. coli* aerobic/anaerobic switch [84]. Since most FFL motifs consist of strong links, it is suggested that many (if not all) FFL motif interactions can be used as parts of other motifs or modules (e.g., for flagella motor, osmoregulated porin gene, oxidative stress response, methionine biosynthesis modules) in a manner that makes the most efficient use of each gene or operon archive [84]. Accordingly, FFL brick motifs are viewed as having an optimal design in

terms of convergent evolution in transcriptional gene regulation networks [105].

The other motif type that is well represented in both networks is the four-gene bi-fan pattern associated with bridge motifs (Table 5). The bi-fan consists of two input transcription factors, one never regulating the other, but both jointly regulating two target genes. In *E. coli*, 208 of the 209 observed bi-fan motifs combined to create dual motif clusters in which most links are shared by at least two adjacent motifs in addition to multiple non-adjacent motifs [84]. No bi-fan brick motifs were found, but 107 bi-fan bridge motifs that did not overlap with other motifs were noted, indicating that they function by themselves. These observations suggest a low co-regulation ratio for two operons in which one regulates the other.

Using the bi-fan bridge motif consisting of *aroL*, *mtr*, *TrpR*, and *TyrR* as an example, the combination of the *TyrR* protein and *TrpR* repressor is responsible for regulating other aromatic amino acid transport genes [106]. The *TyrR* protein plus either phenylalanine or tyrosine is responsible for *mtr* gene activation, while a combination of the *TrpR* repressor plus tryptophan represses the *mtr* gene [107]. Both *TyrR* and *TrpR* regulate the expression of the *aroL* gene-encoding enzyme shikimate kinase II in *E. coli* [84]. Also found were 51 brick motifs (ID = 206) consisting of combinations of FFL and bi-fan motifs. As Dobrin [98] reports, these motifs form a heterologous

motif superstructure. The present results for *S. cerevisiae* are similar to those for *E. coli*. After comparing these with Milo et al.'s [16], it was determined that v-out (ID = 1) and FFL brick motifs (ID = 5) play important roles in both networks (Figs. 20 and 21). Furthermore, the brick motif ratio profiles in the two gene regulation networks are very similar (correlation coefficient $c = 0.96$) (Fig. 26), even though they contain relatively few brick motifs [16].

An effort was made to learn more about the relationship between coherent (incoherent) FFLs [88] and brick (bridge) FFLs. Since each of the three FFL interactions can be either activating or repressing, FFLs have eight possible structural types [80, 84]. The four incoherent FFL types act as sign-sensitive accelerators that shorten the response time of target gene expression following stimuli in one direction (e.g., off to on), but not the other. The four coherent FFL types act as sign-sensitive delays. *E. coli* contains 34 coherent FFLs, 8 incoherent FFLs [84], 29 brick-coherent FFLs, and 6 brick-incoherent FFLs. Accordingly, the difference in coherent (incoherent) FFL frequencies cannot be simply explained by the relative abundances of brick and bridge motifs in a network.

Table 4. Descriptions of five gene regulation networks: edge and node definitions,

network sizes, and references.

Network Type	Common Feature	Directed Network	Nodes	Edges	Description
Gene Regulation (transcription)	Directed graph in which nodes represent genes and edges are directed between genes, regulated by transcription factor.	<i>E. coli</i>	424	519	<i>Escherichia coli</i> [88]
		<i>S. cerevisiae</i> (yeast)	688	1079	<i>Saccharomyces cerevisiae</i> [108]
		<i>Drosophila</i>	110	307	<i>Drosophila melanogaster</i> www.csa.ru/Inst/gorb_dep/inbios/genet/genet.htm
		<i>Sea urchin</i>	43	58	<i>Sea urchin</i> [108]
		<i>C. elegans</i>	280	2170	<i>C. elegans</i> (all synaptic connections used; not restricted to those with ≥ 5 synapses) [26]



Table 5. Brick and bridge motifs in five gene regulation networks.

Network	Nodes	Links	Motif Type	ID	N_{Real}	$N_{Random} \pm STD$	Z_{Score}
<i>E. coli</i>	424	519	Brick	1	402	22.8±19.8	19.17
			Brick	5	35	6.9±2.5	11.30
			Bridge		107	46.8±15.0	4.01
			Bi-fan				
<i>S. cerevisia</i> (yeast)	688	1079	Brick	1	416	17.6±13.9	28.74
			Brick	5	35	5.8±2.9	10.06
			Bridge		1673	276.2±41.2	33.99
			Bi-fan				
<i>Drosophila</i>	110	307	Brick	1	354	123.1±27.8	8.31
			Brick	2	264	108±23.6	6.61
			Brick	5	109	29±7.7	10.45
			Brick	11	14	2.2±1.6	7.55
<i>Sea urchin</i>	43	58	Brick	1	42	24.4±11.2	1.57
			Brick	6	5	1.4±1.4	2.52
<i>C. elegans</i>	280	2170	Bridge	5	740	489.6±32.9	7.60
			Bridge	6	141	53.6±8.7	10.05
			Bridge	11	213	59.1±8.9	17.22
			Bridge	12	75	24.6±4.8	10.41
			Brick	1	2479	950.5±85.1	17.97
			Brick	5	297	46.5±7.4	33.76
			Brick	11	31	2.8±1.8	15.96
			Brick	12	11	0.5±0.7	15.63

We will use the bi-fan bridge motif consisting of *aroL*, *mtr*, *TrpR*, and *TyrR* as an example. The combination of the *TyrR* protein and *TrpR* repressor are responsible for regulating other aromatic amino acid transport genes [57]. The *TyrR* protein plus either phenylalanine or tyrosine is responsible for *mtr* gene activation, while a combination of the *TrpR* repressor plus tryptophan represses the *mtr* gene [58]. Both *TyrR* and *TrpR* regulate the expression of the *aroL* gene-encoding enzyme shikimate kinase II in *E. coli* [42]. We also found 51 brick motifs (ID = 206) consisting of combinations of FFL and bi-fan motifs. As Dobrin [56] reports, these motifs form a

heterologous motif superstructure. Our results for *S. cerevisiae* are similar to those for *E. coli*. After comparing our results with Milo et al.'s [28], we determined that v-out (ID = 1) and FFL brick motifs (ID = 5) play important roles in both networks (Figs. 20 and 21). Furthermore, the brick motif ratio profiles in the two gene regulation networks are very similar (correlation coefficient $c = 0.96$) (Fig. 26), even though they contain relatively few brick motifs [28]. We made an effort to learn more about the relationship between coherent (incoherent) FFLs [12] and brick (bridge) FFLs. Since each of the three FFL interactions can be either activating or repressing, FFLs have eight possible structural types [13], [42]. The four incoherent FFL types act as sign-sensitive accelerators that shorten the response time of target gene expression following stimuli in one direction (e.g., off to on) but not the other. The four coherent FFL types act as sign-sensitive delays. *E. coli* contains 34 coherent FFLs, 8 incoherent FFLs [42], 29 brick-coherent FFLs, and 6 brick-incoherent FFLs. Accordingly, the difference in coherent (incoherent) FFL frequencies cannot be simply explained by the relative abundances of brick and bridge motifs in a network.

Next, the proposed method was applied to transcription networks that guide development in *Drosophila melanogaster* and sea urchin, and synaptic wiring in *Caenorhabditis elegans* (Table 4). As in the two gene regulation networks, brick TSPs

were more significant than bridge TSPs in these three networks. However, it was also determined that four bridge motifs (ID = 5, 6, 11, and 12) in *C. elegans* are very significant (Table 5), indicating the greater presence of isolated motifs. This suggests that these bridge motifs constitute the main difference between the *C. elegans* network and the *Drosophila* and sea urchin networks (Fig. 27). Similarities (differences) in bridge and brick motifs imply similar (different) key circuit elements in each organism.

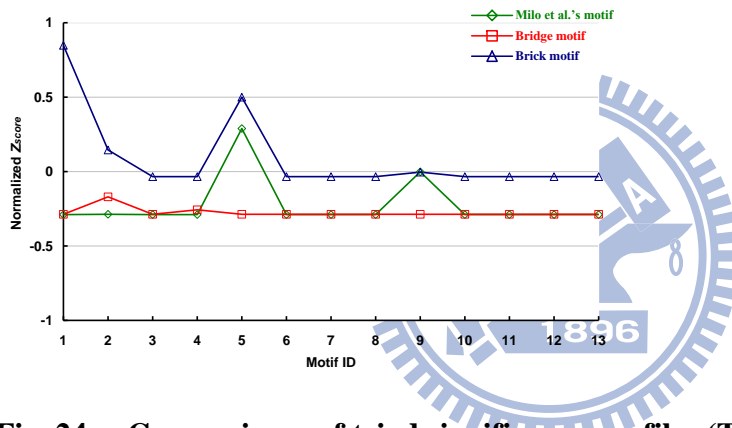


Fig. 24. Comparisons of triad significance profiles (TSPs) for our bridge and brick motifs and Milo et al.'s [7], [28] *E. coli* motifs.

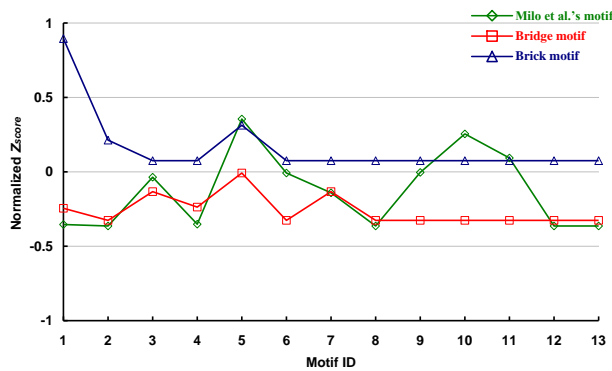


Fig. 25. Comparisons of triad significance profiles (TSPs) for our bridge and brick motifs and Milo et al.'s [7], [28] *S. cerevisiae* (yeast) motifs.

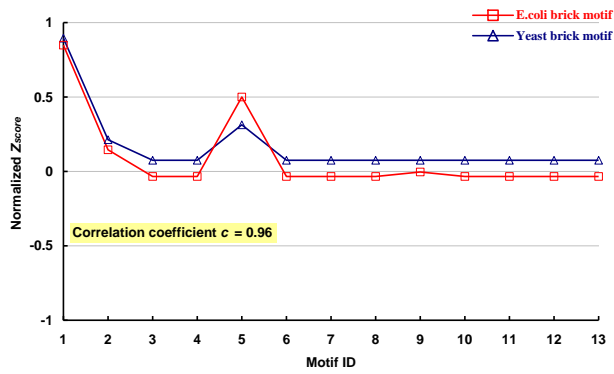


Fig. 26. Brick motif ratio profiles for two gene regulation networks: *E. coli* and *S. cerevisiae* (yeast).

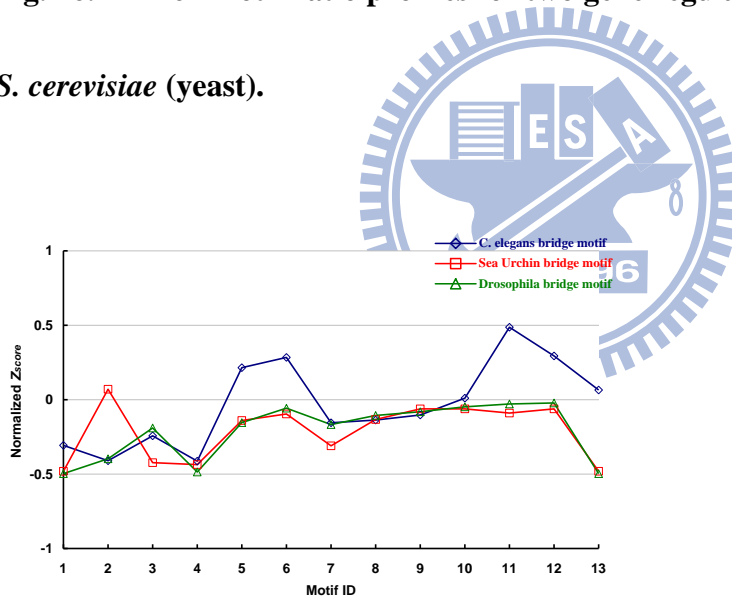
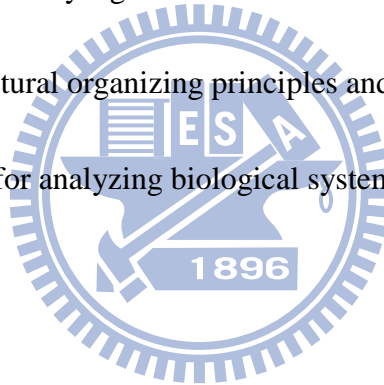


Fig. 27. Bridge motif ratio profiles for three gene regulation networks: *C. elegans*, *sea urchin*, and *Drosophila*.

3.2.3 Conclusion

According to the above definitions of weighted links and network motifs and the results of validation experiments using two gene transcription regulation networks, the presence of bridge and brick motifs in a biological network is closely associated with network topological structures (especially local connections) but not with network size (i.e., number of nodes). Bridge motifs can assist in the identification of isolated motifs, and brick motifs can be used to locate motifs whose functions overlap. This combination of a statistically significant motif and strong or weak-link properties provides insight to the structural organizing principles and functions of networks. It can also serve as a method for analyzing biological system robustness.



Chapter 4 Social Network Simulation

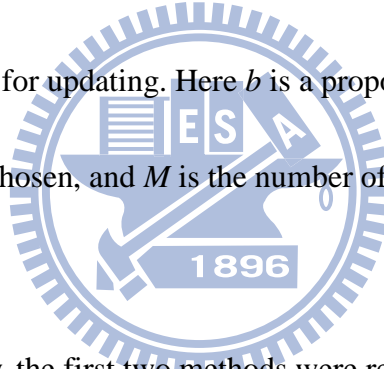
Experiments

To better reflect actual human interactions in social network models, a bottom-up, agent-based modeling and network-oriented simulation approach was used to analyze acquaintance network evolution associated with local interaction rules [23, 109]. In addition to resources and remembering, this approach also considers common friends, meeting by chance, and leaving and arriving. Based on these factors, established friendships can be strengthened, weakened, or broken up [110]. Results from a series of computer simulations indicate that (a) network topology statistics (especially average degree of nodes) are irrelevant to parametric distributions because they rely on average values for initial parameters; (b) resources, remembering, and initial friendships all increase the average number of friends and decrease both degree of clustering and separation; and (c) widely used fieldwork sampling methods cannot capture the actual node degree distributions of social networks.

4.1 Friendship Evolution and The Three-Rule Model

Three selection methods for updating friendships were considered. In the first, *person*

selection, a researcher reviews $b \times N$ persons before picking a specific friend for each one and updating their friendships. Updating does not occur if the chosen person has no friends. In this method, b is a proportion factor for deciding how many persons are chosen, and N represents the number of persons in the network. In the second method, *pair selection*, updating is canceled if paired persons do not know each other—a frequent occurrence, since the network in question is sparse in comparison to a complete graph. In this method, b is a proportion factor for deciding how many pairs are chosen. In the third, *edge selection*, the individual has more direct choice in selecting $b \times M$ friendships for updating. Here b is a proportion factor for deciding how many friendships are chosen, and M is the number of friendships (or edges) at a specific moment.



Without a lack of generality, the first two methods were rejected, since in both cases the number of chosen friendships is in proportion to N (the number of nodes or persons). Since $N \times (N - 1) / 2$ (the upper boundary for the number of friendships) is directly proportional to M (the number of edges or friendships), the edge selection method was adopted for choosing friendships in the experiments.

4.1.1 Friendship Selection Methods

During friend remembering, the model uses the selection method just described to choose a specific number of friendships. If a selected friendship links person u with person v , their friendship is dependent upon three threshold factors: individual remembering, resources, and breakup, expressed as

$$f_{u,v}^{new} = \begin{cases} q \cdot f_{u,v}^{old} + (1 - q) \cdot J \left(D \left(\frac{r_u}{k_u} \right), D \left(\frac{r_v}{k_v} \right) \right), & \text{if } f_{u,v}^{old} \geq \theta \\ 0, & \text{if } f_{u,v}^{old} < \theta \end{cases} \quad (4.1)$$

where $f_{u,v}^{new}$ represents the new friendship between u and v , $f_{u,v}^{old}$ the original friendship, q the old friend remembering, θ the breakup threshold, r_u person u 's friend-making resources, k_u his or her number of friends, and r_v and k_v person v 's resources and friend numbers, respectively. J is a joint function and D a distribution function. For convenience, the friend remembering q , resource r , and breakup threshold θ parameters are normalized between 0 and 1.

Simplification without loss of generality is behind our decision to use $D(x) = x$ as the distribution function and $J(a, b) = (a + b) / 2$ as the joint function. The updated equation is written as

$$f_{u,v}^{new} = \begin{cases} q \cdot f_{u,v}^{old} + (1 - q) \cdot \left(\frac{r_u}{k_u} + \frac{r_v}{k_v} \right) \cdot \frac{1}{2}, & \text{if } f_{u,v}^{new} \geq \theta \\ 0, & \text{if } f_{u,v}^{new} < \theta \end{cases} \quad (4.2)$$

The equation is divided into two parts by the breakup threshold, θ . The first part consists of the terms q (representing the effect of old friendships) and $(1 - q)$ (representing the effect of limited resources). The newly updated friendship may be weakened or strengthened. It may also theoretically equal zero if the new friendship is below the breakup threshold, as shown in the second part of the equation.

4.1.2 Friendship Update Equation

Acting locally, (a) the *friend-making* rule adds links, thereby increasing the average number of friends; (b) the *leaving and arriving* and *friend-remembering* rules both remove links, thereby reducing the average number of friends; (c) increases in the average number of friends $\langle k \rangle$ leads to decreases in the average shortest path length L ; and (d) the directions of the clustering coefficient C and average shortest path length L are reversed.

As opposed to the large number of factors associated with the friend-remembering rule, the leaving and arriving rule has a single parameter: probability p . Factor q

denotes a person's ability to remember friends, thus increasing that person's number of friends. The resource factor r determines an individual's resources for making friends, thereby setting an upper limit. The breakup threshold θ determines the difficulty of cutting off a friendship—a negative influence. The initial friendship factor f_0 is a reflection of how much attention a person is paying when making a new acquaintance—a positive contribution to friend-making. It was expected that parameters q , r and f_0 would exert positive (increasing) influences on $\langle k \rangle$, and that parameters p and θ would exert negative influences on $\langle k \rangle$.

4.1.3 Fitting a Normal Distribution

For sensitivity analyses of skewness and critical parameters affecting distribution, a feasible probability-distribution function (pdf) must be applied. In most situations a normal distribution is considered the best choice, but it did not fit the purposes of this study. Since critical parameters such as initial friendships, old friends remembering, resources, and breakup thresholds have ranges of 0 to 1, a beta distribution was selected—a two-parameter family of continuous probability distributions defined according to the interval [0, 1] with a probability density function of

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (4.3)$$

where B is the beta function and α and β both must be > 0 . I used a beta distribution subset called beta14, which satisfies $\alpha + \beta = 14$, and has $\mu = \alpha / (\alpha + \beta)$ as its average. Figure 28 shows pdf curves for beta14 distributions with averages of 0.1, 0.5, and 0.9; Figure 29 shows pdf curves for comparing beta14 and normal distributions.

Once a simulation reached a statistically stationary level, data for clustering coefficient C , average path length L , average degree of nodes $\langle k \rangle$, average square degree of nodes $\langle k^2 \rangle$, and node degree distribution statistics were collected. Degree distributions in the simulations involved some random rippling, especially for smaller populations. However, since large populations consume dramatically greater amounts of simulation time, Bruce's [111] ensemble average was applied as follows:

$$\bar{p}(k) = \frac{1}{M} \sum_{v=1}^M p_v(k) \tag{4.4}$$

where M is the number of curves to be averaged and $p(k)$ a curve that represents.

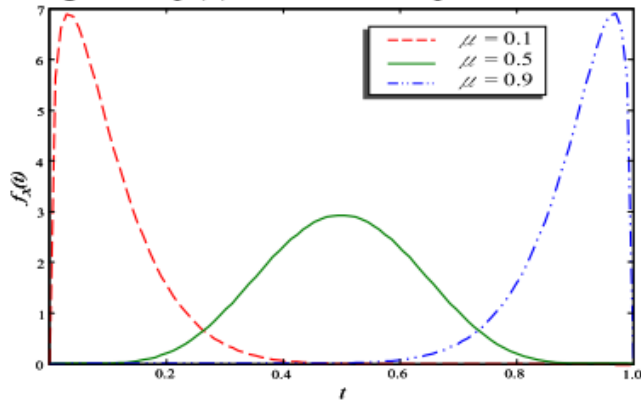


Fig. 28. Beta14 pdf curves at different averages of 0.1, 0.5, and 0.9.

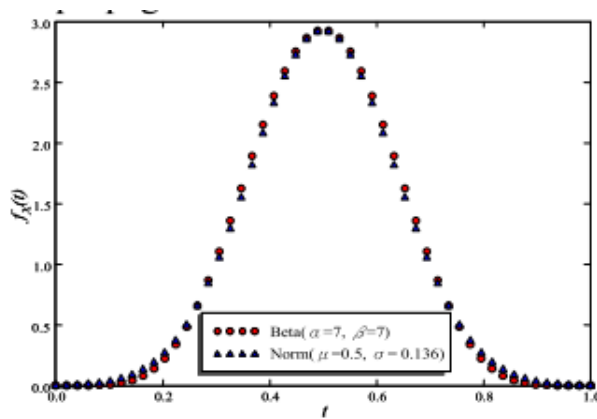


Fig. 29. Comparison of beta and normal distributions.

4.2 Experiment

A simulation using the proposed model starts with parameter initialization and ends once the acquaintance network reaches a statistically stationary state. As shown in Table 6, initialized parameters included number of persons N , leaving and arriving probability p , updated friendship proportion b , old friend remembering q , breakup threshold θ , distribution of friend-making resources r , and distribution of initial

friendship f_0 .

Table 6. Terms and abbreviations for initialized parameters

Abbreviation	Description
N	Number of persons (nodes) in acquaintance network.
M	Number of friendships in acquaintance network.
$f_0, f(t=0)$	Initial friendship distribution.
p	Leaving and arriving probability in rule 2.
b	Proportion of updated friendships in rule 3.
f	Friendships.
q	Old friend remembering.
r	Friend-making resource distribution.
θ, th	Breakup threshold.
M, mu	Mean.
$fixed_value(\mu)$	A distribution that sets its random variable as a fixed value μ .
$beta14(\mu)$	A <i>beta</i> distribution instance in $[0, 1]$ that forces $\alpha + \beta = 14$ with $\mu = \alpha / (\alpha + \beta)$.

Statistically stationary states were determined by observing average degree of nodes $\langle k \rangle$, average square degree of nodes $\langle k_2 \rangle$, clustering coefficient C , and average path length L (Table 7). Each of these four statistics eventually converged to values with slight ripples. Figure 30 presents a statistically stationary state of parameter initialization at $N = 1,000$, $p = 0$, $b = 0.001$, $q = 0.9$, $\theta = 0.1$, r with a fixed value of 0.5 and a $beta14 f_0 (\mu = 0.9)$. Blue solid lines indicate the acquaintance network and green dashed lines in (c) and (d) show an ER random model with the same average degree of nodes as the acquaintance model.

Table 7. Terms and abbreviations for statistics

Abbreviation	Description
$\langle k \rangle$	Average degree of nodes.
$\langle k^2 \rangle$	Average square degree of nodes.
C	Average clustering coefficient.
L	Average shortest path length

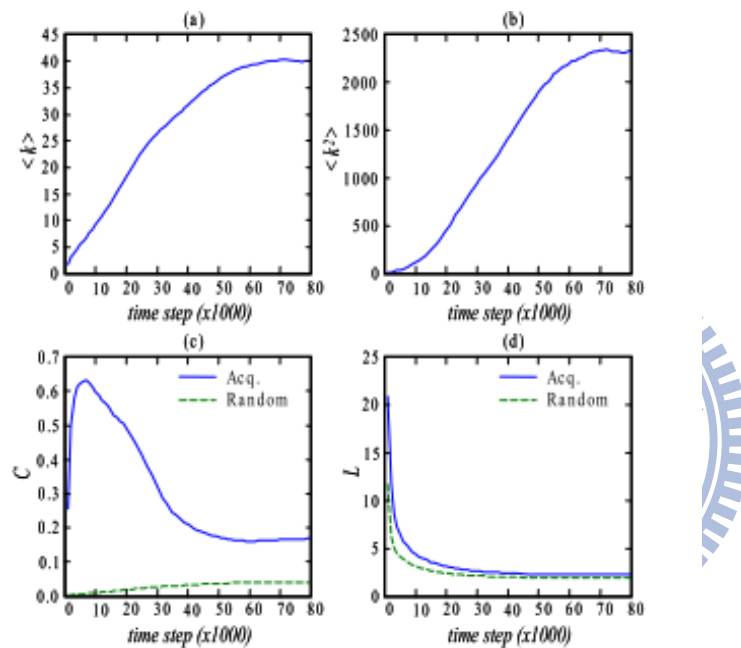


Fig. 30. Example of a statistically stationary state using the proposed model.

4.2.1 Effects of Leaving and Arriving

For comparison purposes I reproduced Davidsen et al.'s [112] simulations using their original parameters of $N = 7,000$ and p at 0.04, 0.01, or 0.0025, and then changed N to

1,000 and tested a broader p range. As noted above, the leaving and arriving probability p is the only rule 2 parameter. In addition to using various degree distribution diagrams, correlations among $\langle k \rangle$, C , and L values were analyzed to determine the effects of changes in p on the acquaintance network.

A $P(k)$ degree distribution from the two-rule model is shown in Figure 28, and all $\langle k \rangle$, C , and L values with parameter initializations for various probability p values are shown in Figure 32. The solid lines in Figure 32 reflect the application of Davidsen et al.'s two-rule model, and the dashed lines reflect the application of the ER model with the same average node degree. Contrasts between the two lines in Figures 31b and c indicate that the acquaintance network has a small world characteristic. According to Figure 32a, the number of friends increases as the lifespan of an individual lengthens. Figure 32 also shows that the clustering coefficient closely follows average degree of nodes but not average path length. A larger p indicates a higher death rate and a lower p a longer life span—in other words, parameter p serves as an aging factor. Relative to other species, humans require more time to make friends, therefore Davidsen et al. only focused on the $p \ll 0.1$ regime. To satisfy the needs of integrity theory, I also explored the $p \gg 0.1$ regime and found that average node degree $\langle k \rangle$ decreased for p values between 0 and 0.5. The decrease slowed once p exceeded 0.1. Note that in the

proposed model, a leaving and arriving probability of 0 means that rule 2 is inactive, and a friendship update proportion of 0 means that rule 3 is inactive. Once rule 3 becomes inactive, the proposed three-rule model becomes the equivalent of Davidsen et al.'s two-rule model. In all of the experiments described in the following sections, N was initialized at 1,000 and b at 0.001.

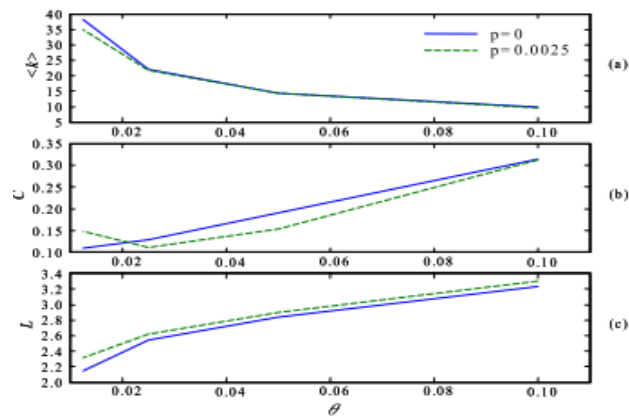


Fig. 31. $\langle k \rangle$, C and L varying in breakup threshold θ with different leaving and arriving probability p .

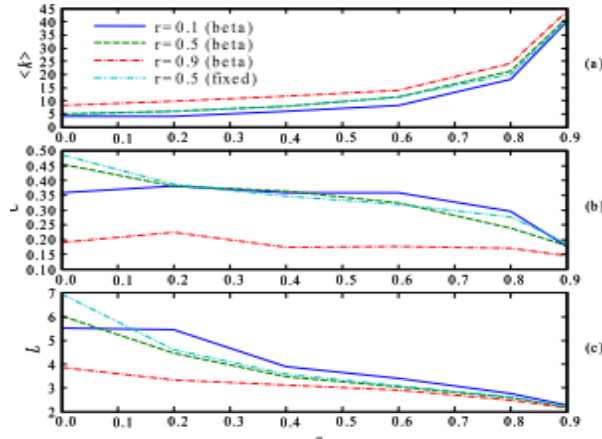


Fig. 32. $\langle k \rangle$, C and L varying in friend-remembering q value with different distributions of friend-making resources.

4.2.2 Effects of Breakup Threshold

To determine the effects of the breakup threshold on the acquaintance network, experiments were performed with parameters initialized at different levels of the friendship-breakup threshold θ . Other initialized parameters were $q = 0.6$ and the constants $r = 0.5$ and $f_0 = 0.5$. The solid lines in Figure 30 represent $\langle k \rangle$, C , and L statistics without rule 2 included ($p = 0$), and the dashed lines represent the same statistics with rule 2 added ($p = 0.0025$). The data indicate that rule 2—which acts as an aging factor on acquaintances in the network—reduced both average degree of nodes $\langle k \rangle$ and clustering coefficient C , and increased average path length L .

According to the data presented in Figure 33, the breakup threshold θ lowered the

average node degree $\langle k \rangle$ and raised both the clustering coefficient C and average path length L . The threshold reflects the ease with which a friendship can be broken. As expected, a higher θ resulted in a smaller number of “average friends” and greater separation between individuals.

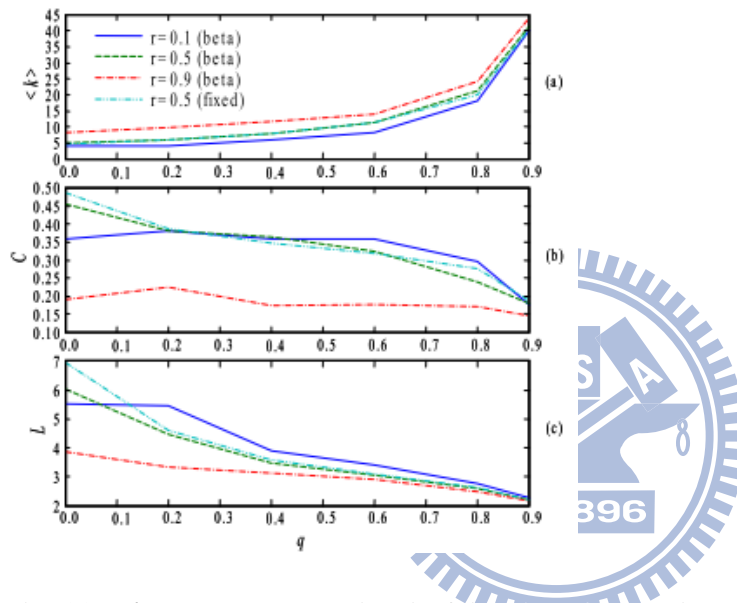


Fig. 33. $\langle k \rangle$, C and L varying in friend-remembering q value with different distributions of initial friendship f_0 .

4.2.3 Effects of Resources

To determine the effects of resources and memory factors on acquaintance networks, I ran a series of experiments using parameters initialized with different friend-making resource r and friend-remembering q values. Initialized parameters also included $p =$

0, $\theta = 0.1$, and a fixed f_0 value of 1. The results indicate that a larger r raised the average node degree $\langle k \rangle$ but lowered both the clustering coefficient C and average path length L (Fig. 31). It remains unclear whether statistical characteristics were influenced by different resource distributions, but they were clearly influenced by different resource averages. In other words, $\langle k \rangle$, C , and L are all affected by different resource averages, but not by different resource distributions. The Figure 34 data also show that an increase in q raised $\langle k \rangle$ and lowered both C and L .

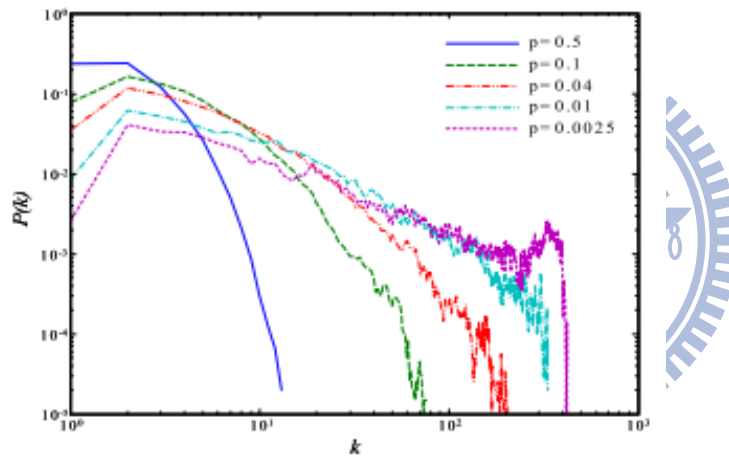


Fig. 34. Two-rule model degree distribution $P(k)$.

4.2.4 Effects of Initial Friendship

Experiments were run using parameters initialized at different initial-friendship f_0 and friend-remembering q values for the purpose of determining the effects of those factors on acquaintance networks. Other initialized parameters were $p = 0$, $\theta = 0.1$,

and a fixed r value of 0.5. The results show that a larger f_0 raised the average degree of nodes $\langle k \rangle$ but lowered both the clustering coefficient C and average path length L (Fig. 35). It was not obvious whether different distributions of initial friendship influenced statistical characteristics, but different averages of initial friendship clearly did. In other words, $\langle k \rangle$, C , and L were affected by different initial friendship averages, but not by different initial friendship distributions. The Figure 35 data also show that the friend remembering q factor raised $\langle k \rangle$ and lowered both C and L . The effects of different parameters on the proposed model were analyzed by relationally cross-classifying all experiments; results are shown in Tables 8 and 9. The plus/minus signs in Table 8 denote positive/negative relations between parameters and statistics. In Table 9 the plus or minus signs denote the strength and direction of correlations. As Table 8 indicates, in addition to the effects of rule 1, positive correlations were found for q , r , and f_0 , with average degree of nodes $\langle k \rangle$ and p and θ having negative correlations with $\langle k \rangle$. Furthermore, each average node degree had a negative relationship with its corresponding average path length. All of the rule 3 parameters affected the clustering coefficient C and average length L in a positive manner, while rules 1 and 2 affected C and L negatively. Note that friendships were initialized in rule 1 and updated in rule 3.

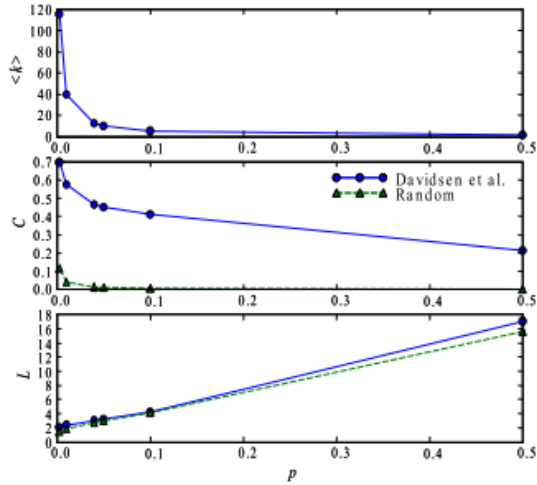


Fig. 35. $\langle k \rangle$, C and L varying in leaving and arriving probability p .

Table 8. Effective directions of the parameters on $\langle k \rangle$, C , L

Statistics	Rule 1	Rule 2		Rule 3		
		p	q	θ	r	f_0
$\langle k \rangle$	+	-	+	-	+	+
C	+	-	-	+	-	-
L	-	+	-	+	-	-

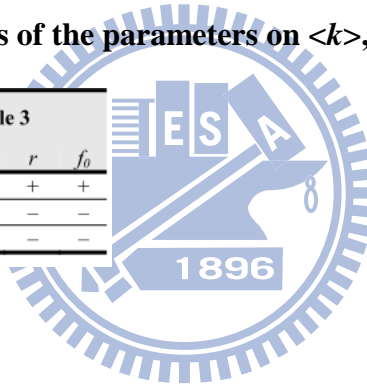


Table 9. Correlations between $\langle k \rangle$, C , L from experiments

Experiments	Variational Parameters	$C-\langle k \rangle$	$L-\langle k \rangle$	$C-L$
4.1	P	+++	—	—
4.2	θ, p	—	—	+++
4.3	q, r	—	—	+++
4.4	q, f_0	—	—	+++

4.2.5 Distribution of Co-Directors

The three main properties of social networks are (a) the small-world phenomenon, (b)

high-clustering, and (c) skewed node degree distribution [39]. The focus here was on the third property. Interlocking board of director networks show a remarkable node degree distribution that is very different from either scale-free or ER random networks [24]. A clear example is the connections among the nearly 8,000 directors on the boards of Fortune 1000 companies in 1999; the corresponding degree distribution shows a strong peak and fast (approximately exponential) tail decay, much faster than a power-law distribution, but slower than a Poisson or normal distribution [113].

A node degree distribution comparison between one of the proposed acquaintance networks at a statistically stationary state (blue solid curve) and co-directors for Davis' [112] boards-of-director data (green dashed curve) is presented in Figure 36. Selected acquaintance network parameters were initialized at $N = 1,000$, $p = 0$, $b = 0.001$, $q = 0.4$, $\theta = 0.1$, $r = 0.5$ (fixed) and $f_\theta = 0.5$ (fixed). Davis' data are for the 8,000 directors described in the previous paragraph. Both curves exhibit similar peaks and long tails that do not decay smoothly.

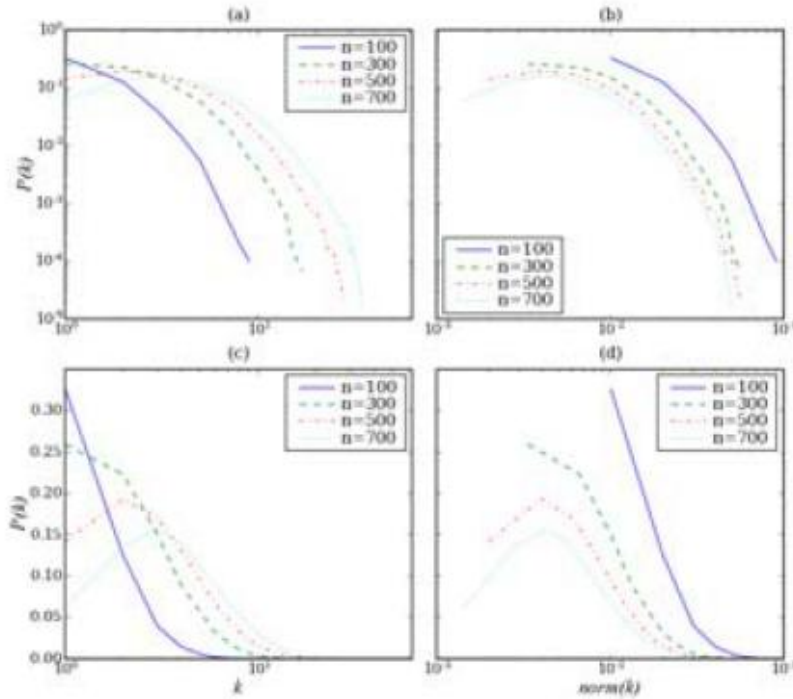
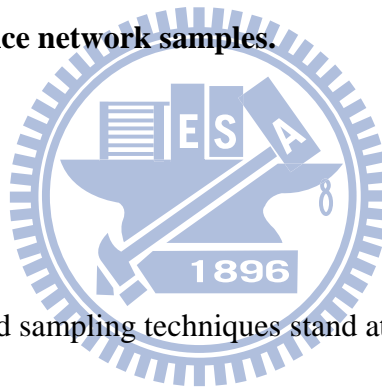


Fig. 36. Model acquaintance network samples.



4.2.6 Sampling

Surveys, questionnaires, and sampling techniques stand at the center of traditional social science research and are considered cheaper and more practical than collecting large amounts of census data. However, the effectiveness of these methods for analyzing social networks has not been examined—the motivation for running an arbitrary simulation of the proposed model after reaching a statistically stationary state, and for collecting a sample of nodes. Initialized parameters were $N = 1,000$, $p = 0$, $b = 0.001$, $q = 0.4$, and $\theta = 0.1$; constants were $r = 0.5$ and $f_0 = 0.5$.

Figure 11 presents degree distribution $P(k)$ data after sampling at 100, 300, 500 and 700 nodes. Figures 11a and 11b are log plots with log scaling on the x and y axes; these were used to determine if distributions were scale-free. Figures 11c and 11d are semi-log plots with log scaling on the x axis only; these were used to determine if distributions were exponential. Degrees in Figures 11b and 11d are post-normalization, as required for different numbers of sampled nodes. Each curve in Figure 11 represents an ensemble average of 100 sampling repetitions. The solid lines in Figure 11 reflect a lower sampling ratio of 0.1, considered common for traditional surveys and sampling techniques. The dotted lines reflect a higher sampling ratio of 0.7, considered common for a census. Turns in the direction of the y-axis were observed for high sampling but not for low. The degree distribution clearly lost its original shape after sampling.

4.3 Conclusion

Experimental simulations are a necessary aspect of social network research, not only due to expenses and other difficulties involved with fieldwork, but also because widely used sampling approaches cannot capture real social network distributions, since distributions for higher sampling rates differ from those for lower sampling rates.

Taking a bottom-up, agent-based modeling and network-oriented simulation approach to modeling reflects the evolution mechanism of real social networks. Building on insights from previous studies, I applied local and interactive rules to acquaintance network evolution. This approach produced new findings that can be used to explore human activity in specific social networks—for example, rumor propagation and disease outbreaks.



Chapter 5 Epidemic Dynamics

Experiments

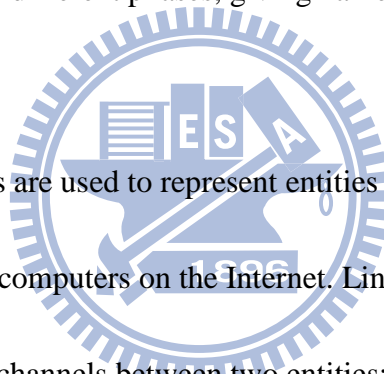
Whether or not a critical threshold exists when epidemic diseases are spread in complex networks is an interesting problem in many disciplines. In 2001, Pastor-Satorras et al. [114] used a computational simulation to show that epidemic diseases which spread through scale-free social networks do not have positive critical thresholds. However, they ignored two key factors that have a large impact on epidemic dynamics: economic resource limitations and transmission costs. Every infection event entails tangible or intangible costs in terms of time, energy, or money to the carrier, recipient, or both. Here we apply an agent-based modeling and network-oriented simulation approach to analyze the influences of resource limitations and transmission costs on epidemic dynamics and critical thresholds in scale-free networks. Our results indicate that when resources and costs are taken into consideration, the epidemic dynamics of scale-free networks are very similar to those of homogeneous networks, including the presence of significant critical thresholds.

5.1 Epidemic Dynamics in Complex Networks

In a standard epidemiological model, all individuals in a population can be roughly classified into a small number of states, including Susceptible (S), meaning that an individual is vulnerable to infection but has not yet been infected; Infected (I), meaning that an individual can infect others; and Removed (R), meaning that an individual has either recovered, died, or otherwise ceases to pose any further threat.

Generally speaking, epidemiologists use combinations of these states to represent orders of transition between different phases, giving names such as SIR and SIS to their models.

In complex networks, nodes are used to represent entities such as organisms in biological environments or computers on the Internet. Links indicate close relationships or interaction channels between two entities; those with direct connections are called *neighbors* [21, 39]. When simulating the transmission dynamics of epidemic diseases in complex networks, epidemiologists usually assume that nodes in complex networks randomly run through an SIS cycle (Susceptible \rightarrow Infected \rightarrow Susceptible). During each time step, all susceptible nodes connected to one or more infected nodes are subject to a probability ν infection rate. Infected nodes recover at a probability ε recovery rate, and once again become susceptible. Based on



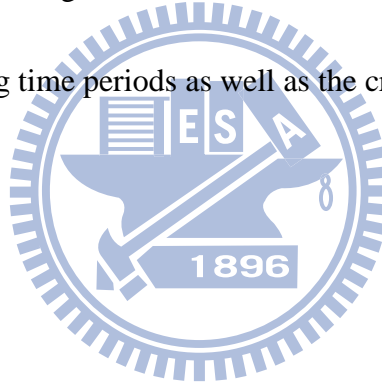
the infection ν and recovery ε rate definitions, effective spreading rate λ is defined as $\lambda = \nu / \delta$. Without a lack of generality, recovery rate ε can be assigned a value of 1, since it only affects individuals during a period of disease propagation.

Pastor-Satorrasni [115] define $\rho(t)$ as the density of infected nodes at time t . When time t becomes infinitely large, ρ can be represented as a steady density of infected nodes. Using these definitions, Pastor-Satorras [115] applied dynamic mean-field theory to the SIS model and proposed homogeneous mixing hypothesis according to the topological features of homogeneous networks for obtaining the stable density of infected nodes ρ during long time periods as well as the critical threshold λ_c , That

hypothesis is expressed as

$$\rho = \begin{cases} 0 & \lambda < \lambda_c \\ \frac{\lambda - \lambda_c}{\lambda} & \lambda \geq \lambda_c \end{cases}$$

$$\lambda_c = \frac{1}{\langle k \rangle}$$



(5.1)

According these Equations, a positive and nonzero critical threshold λ_c exists in a homogeneous network based on the SIS model. If the value of the effective spreading rate exceeds the critical threshold ($\lambda \geq \lambda_c$), the infection spreads and gains persistence. If the effective spreading rate is below the critical threshold ($\lambda < \lambda_c$), the infection dies at an exponential speed. In summary, the primary prediction of an SIS

epidemiological model in a homogeneous network is the presence of a positive critical threshold, proportional to the inverse of the $\langle k \rangle$ average number of neighbors of every node, below which epidemics die and epidemic states are impossible.

Pastor-Satorras relaxed their homogeneity assumption for homogeneous networks and obtained the critical threshold λ_c in a scale-free network as following Equation. The results indicate that in scale-free networks with a connectivity exponent of $2 < \gamma \leq 3$ and for which $\langle k^2 \rangle \rightarrow \infty$ is the limit of a network of infinite size ($N \rightarrow \infty$), the critical threshold λ_c is very close to 0 ($\lambda_c \rightarrow 0$).

$$\lambda_c = \langle k \rangle / \langle k^2 \rangle \quad (5.2)$$

Pastor-Satorras [116] express the total prevalence ρ for the SIS epidemiological model in a BA scale-free network as a function of the effective spreading rate λ , and compare it to the theoretical prediction for a homogeneous network. As shown in Figure 37, dashed and solid lines represent BA scale-free and WS small-world networks, respectively. The total prevalence ρ of a BA scale-free network reaches 0 in a continuous and smooth manner when the effective spreading rate λ decreases; this indicates an absence of any critical threshold ($\lambda_c = 0$) in a BA scale-free network. As long as $\lambda > 0$, epidemic diseases can be stably transmitted in the network and eventually reach a steady state. This explains why scale-free networks are fragile in

epidemiological spreading situations. Since social networks and the Internet both have “the rich get richer” properties, computer viruses, biologically infectious diseases, and cultural trends can be stably transmitted even when initial infection cases occur in small, limited areas.

For finite-size scale-free networks, Pastor-Satorras et al. [117] introduced the concept of maximum connectivity k_c (dependent on N), which has the effect of restoring a connectivity fluctuation boundary and inducing an effective nonzero critical threshold.

According to the definition of maximum connectivity k_c , $\langle k_2 \rangle$ clearly has a finite value in finite-size scale-free networks. However, in this situation the critical threshold (which is not an intrinsic quantity as it is in homogeneous networks) vanishes as network size increases.

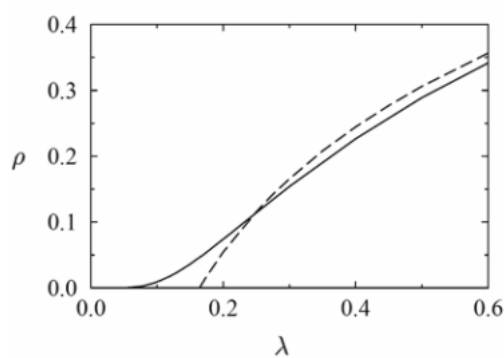


Fig. 37 Prevalence ρ in steady state as a function of effective spreading rate λ .

In addition to the topological characteristics of complex networks, individual differences (e.g., supercarriers and individuals immune to certain infectious diseases) and environmental factors (e.g., mosquito breeding sites) exert considerable influences on transmission dynamics and epidemic disease diffusion. Huang et al. [118] used Watts and Strogatz's small-world networks to investigate the influence of individual differences ("local information") on epidemic simulations. Specifically, they used a sensitivity analysis to show that when an agent-based modeling and network-oriented simulation approach is applied to exploring epidemic transmission dynamics in small-world networks, researchers should focus not only on network topological features, but also on proportions of specific values of individual differences related to infection strength or resistance. Less emphasis should be placed on the details of the topological connection structures of small-world networks and the distribution patterns of individual difference values.

5.2 Experiments

The first simulation experiment focused on the universality and generality of the steady density curve and critical threshold when individual economic resources and transmission costs are taken into consideration. Usable economic resources $R(v_i)$ of

individual v_i at time t was set at 16 units, and transmission cost $c(v_i)$ was set at one unit, thus accounting for 6.25% of the individual's total usable economic resources.

The relationship between effective spreading rate λ and steady density ρ in the SIS epidemiological model was compared using three types of complex network platforms:

small-world, scale-free without transmission costs, and scale-free with limited

individual economic resources and transmission costs. As shown in Figure 38, the

eight simulation experiment suites generated consistent results that did not become

contradictory following changes in node and edge numbers. It is therefore suggested

that the results can be applied to various scale-free networks used to simulate

infectious diseases.

The red curves in Figure 36 indicate that the steady density ρ of the SIS

epidemiological model based on scale-free networks reached 0 in a continuous and

smooth manner when the effective spreading rate λ was decreased, indicating the

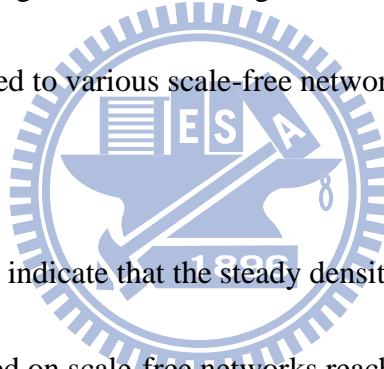
absence of a critical threshold ($\lambda_c = 0$) in scale-free networks without transmission

costs. The blue curves show that infectious diseases do have critical thresholds in

small-world networks (approximately 0.14). If the value of the effective spreading

rate exceeds the critical threshold (i.e., $\lambda \geq \lambda_c$), the infection will spread throughout the

network and eventually reach a steady density $\rho(\lambda)$. If $\lambda < \lambda_c$, the infection dies almost



immediately. The green curves represent the steady densities $\rho(\lambda)$ of infectious diseases in scale-free networks when individual economic resources and transmission costs are taken into consideration. In addition to being very similar to the blue steady density curves in small-world networks, the green curves have critical thresholds (again approximately 0.14). One conclusion drawn from the results of the first simulation experiment is that individual economic resources, transmission costs, and average vertex degree exert significant influences on epidemic dynamics and critical thresholds in scale-free networks. The same conclusion can be applied to the second and third simulation experiments.

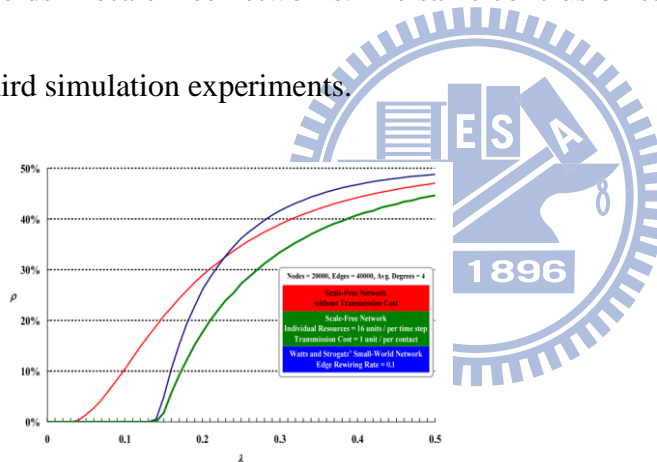


Fig. 38. Relationship between effective spreading rate and steady density of the SIS epidemiological model on three types of complex network platforms.

The second experiment focused on the relationship between the ratio of transmission costs to the total amount of economic resources (hereafter referred to as “the ratio”)

and critical threshold. To evaluate the influence of the ratio on epidemic dynamics and critical thresholds, I used ten economic resource quantities (4, 8, 12, 16, 20, 24, 28, 32, 36 and 40 units) and assigned the transmission cost $c(v_i)$ of each interaction event as a single unit accounting for 25%, 12.5%, 8.33%, 6.25%, 5%, 4.17%, 3.57%, 3.13%, 2.78% and 2.5% of an individual's economic resources, respectively. As shown in Figure 39, the critical threshold significantly increased as the ratio grew. For instance, when the $R(v_i)$ resources of individual v_i at time t were designated as 8 units, the critical threshold was approximately 0.22 (pink curve)—significantly greater than that of a small-world network with the same number of nodes and edges (blue curve) and the same average number of vertex degrees (Fig. 38). The opposite was also true: when the $R(v_i)$ of individual v_i at time t was designated as 40 units, the shape of the steady density curve (Fig. 39, red curve) was very close to that of the scale-free network without transmission costs (black curve), and the critical threshold was reduced to 0.09. As shown in Figure 40, a linear correlation exists between the critical threshold and the ratio. Another interesting observation was that the steady density curve grew at a slower rate as the ratio increased—that is, the ratio and steady density had a negative linear correlation. One conclusion drawn from the second simulation experiment is that when transmission costs increase or economic resources decrease,

the critical thresholds of spreading infectious diseases in scale-free networks grow, while steady density shrinks according to the diffusion rate.

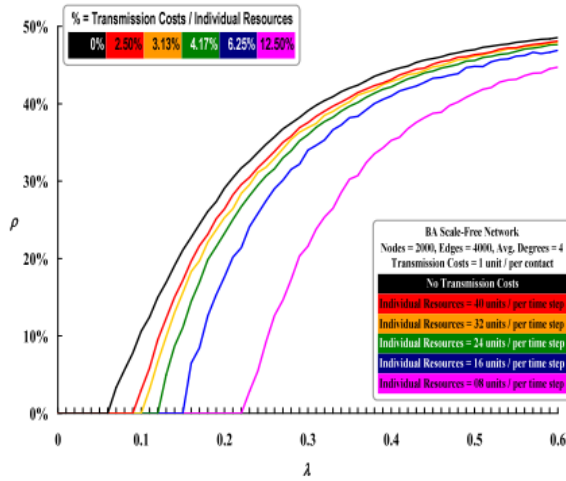


Fig. 39. How the amount of an individual's economic resources affect steady density curves.

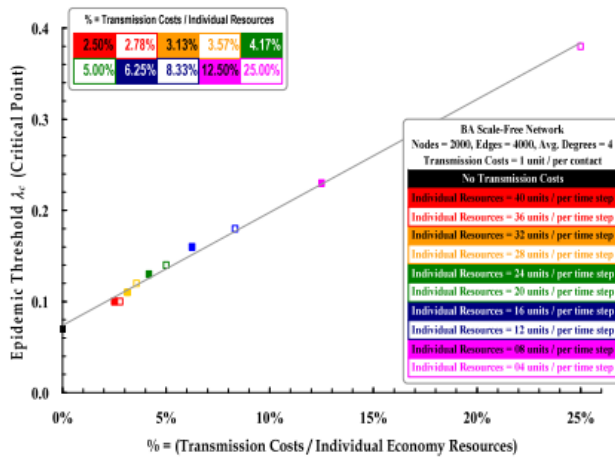


Fig. 40. Relationship between ratio of transmission costs to an individual's economic resources and critical threshold.

The third simulation experiment was designed to determine how different distribution types (delta, uniform, normal, power-law) of economic resources and their statistical distribution parameters (standard deviation in a normal distribution, number of values and range in a uniform distribution) affect the steady density curves and critical thresholds of infectious disease diffusion in scale-free networks marked by limited individual economic resources and transmission costs. The orange, green, and purple steady density curves in Figures 41 and 44 represent the delta (fixed value of 16), uniform, and normal distributions of individual economic resources, respectively; normal distributions are shown in Figures 42 and 45. All had the same critical threshold (≈ 0.14), and their steady density curves almost overlapped with each other when the average values of the economic resources were the same. However, as shown in Figures 43 and 46, if those same economic resources reflected a power-law distribution (i.e., the majority of individuals had extremely limited economic resources, and a small number had the most), and no correlation existed between the amount of an individual's economic resources and vertex degree, the resulting steady density curve (pink) grew more slowly than those of the other three distributions, even though they all had the same critical threshold.

The same results were produced as long as the average values of the individual economic resources were the same (Figs. 41 and 44). The steady density curves and critical thresholds were almost identical across different distribution types, regardless of whether the individuals' economic resources obeyed a uniform distribution with a range of 2 or 3 (Figs. 42 and 45, green bars) or a normal distribution with a standard deviation of 2 or 3 (Figs. 42 and 45, purple curves). From the two significantly different groups of steady density curves in Figures 41 and 44 (orange, green, and purple versus pink), it was concluded that as long as researchers ensure that economic resources do not obey a power-law distribution, they can simply assign each individual's $R(v_i)$ at time t as the average value $\langle r \rangle$ of the statistical distribution derived from the real-world scenario, and thereby facilitate the requirements of their experiments without affecting simulation results.

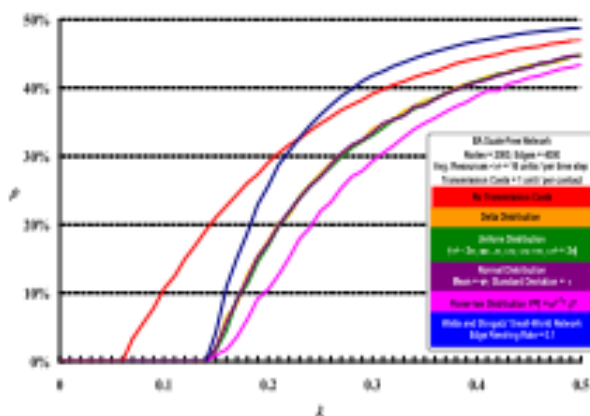


Fig. 41. How different distribution types of individual economic resources (delta,

uniform, normal, power-law) affect steady density curves and critical thresholds of infectious disease diffusion in a scale-free network.

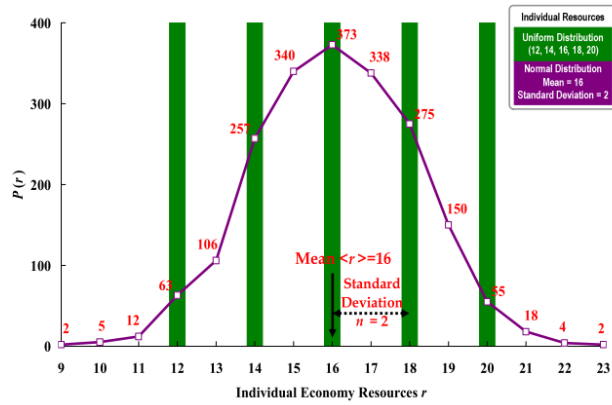


Fig. 42. A uniform ($n = 5, r = 2$) and normal distribution (standard deviation = 2) of individual economic resources with average value $\langle r \rangle$ of 16.

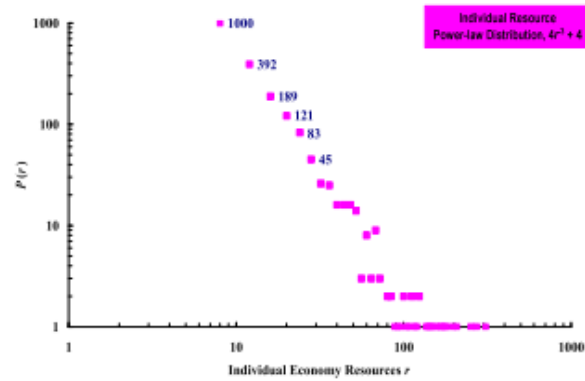


Fig. 43. Individual economic resources in a power-law distribution.

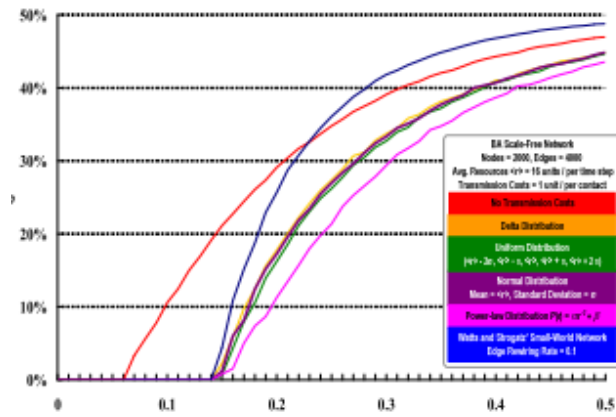


Fig. 44. How different types of individual economic resource distributions (delta, uniform, normal, and power-law) affect steady density curves and critical thresholds of infectious disease diffusion in a scale-free network.

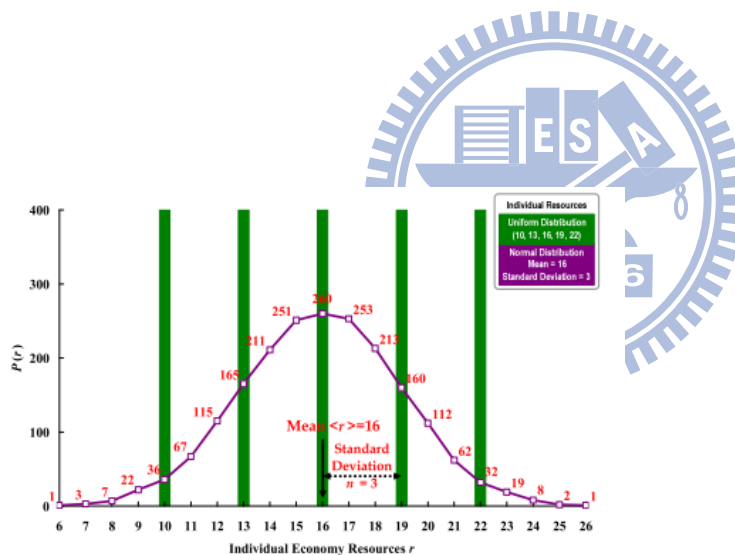


Fig. 45. A uniform ($n = 5, r = 3$) and normal distribution (standard deviation = 3) of individual economic resources with average value $\langle r \rangle$ of 16.

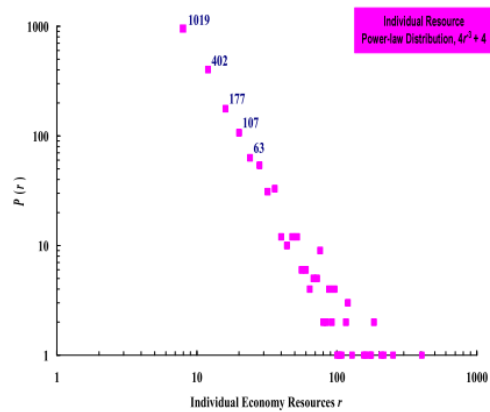


Fig. 46. Individual economic resources in a power-law distribution.

5.3 Conclusion

Agent-oriented modeling and complex networks were used to construct infectious disease simulation models for the purpose of investigating how economic resources and transmission costs influence epidemic dynamics and thresholds in scale-free networks. The results indicate that when economic resources and transmission costs are taken into consideration, a critical threshold does in fact exist when infectious diseases are spread within a scale-free network. These conclusions may help epidemiologists and public health professionals understand core questions of disease epidemics, predict epidemic dynamics and diffusion, and develop effective public health policies and immunization strategies.

Chapter 6 Abstraction Hierarchy

Experiments

Networks consisting of node and link assemblies are viewed as convenient representations of interactions in complex systems. Many complex networks in fields ranging from biology to sociology have been analyzed, and complex network research has cultivated alternative views of and advanced new methodologies for solving problems in complex domains. An important and challenging problem concerns partitioning individual networks into *clusters*—also called *communities* or *groups* in social networks [29, 45, 65] and *motifs* or *modules* in biology [16, 72, 119]. Research in this area has focused on identifying clusters and their hierarchical organizations in a manner that corresponds to such real-world meanings as biological functions [67, 98, 119], economic laws, and political constraints [2, 29]. Despite a number of successful examples, no uniform measure of modularity or standard structure of hierarchy has been universally accepted [4, 32, 67, 69, 98]. The definition of community implied by modularity is not necessarily consistent with its optimization [32]; consequently, a perfect partition enforced by modularity optimization may not correspond to a network's actual community structure [32, 120]. Furthermore, a hierarchical

architecture can represent different relationships between hierarchical levels. They can be as general as inclusion/nesting relationships (e.g., those found in logic gates in D-flips, metabolic pathways [121, 122], or domestic airline flights [69]), or as specific as gene-target regulations observed in *E. coli* genetic regulatory networks (e.g., the manner in which IHF regulates OmpR and OmpR regulates FlhDC [34]).

Most current forms of module hierarchical organization are limited to vertical relationships between modules at different hierarchical levels, thus overlooking horizontal relationships among modules at the same level. Vertical relationships support representations of inclusion hierarchies [67, 69] and causality/regulation [98]. To complement these, horizontal relationships can be used to provide abstractions of original networks of interest at various levels within a hierarchy. Domain experts can therefore focus on interconnections among modules at each hierarchy level [63].

6.1 Background

Advancements in complex network research have supported the identification of many significant network properties from various domains and disciplines—for example, small-world architectures [21], scale-free connectivity [24], and network motifs (modules) [16, 26]. Several researchers believe that hierarchical relations exist

in complex networks, and have demonstrated the potential application of the hierarchy concept to system-level analyses (e.g., cellular processes or domestic flights in the US) [43, 64].

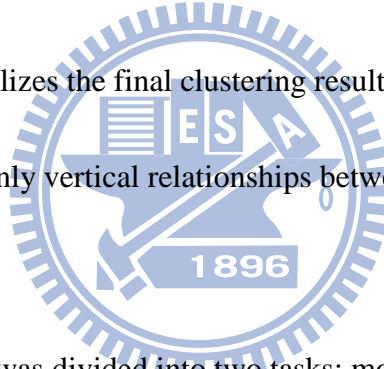
One widely used method for finding hierarchical data organization is hierarchical clustering [61, 123, 124], in which data are grouped into sequences of nested clusters, ranging from singletons to clusters consisting of agglomerations of all individuals in a fashion or vice versa in a divisive style. Both techniques organize data into hierarchical structures that are usually depicted as dendrograms (e.g., Fig. 47).

Compared to divisive clustering, agglomerative clustering has a computational advantage in that it only considers $O(n^2)$ merges of data given n data points, rather than $O(2^n)$ possible divisions of data into two nonempty groups. However, it also has a drawback in its tendency to find only cores of clusters (leaving out peripherals), since core nodes often have shorter distances between each other, and therefore merge earlier during the agglomerative process [1]. As shown in Figure 47, the root of the dendrogram represents the whole dataset, and each leaf singleton is regarded as a data point. The intermediate nodes contain data points proximal to each other, and the height in the dendrogram expresses the distance between nodes. Agglomerative clustering starts with $n=5$ single data points. Based on a distance metric, it calculates

the proximity matrix for the n singleton clusters. It selects two clusters with the minimum distance, and merges them into a new cluster. By repeating the same process, a series of merge operations is applied to force all data points into one cluster eventually. To obtain the final clustering results, we must cut the dendrogram at an appropriate level. We show two possible cut lines. Divisive clustering, on the contrary, starts with one cluster that holds all data points. It iteratively selects a cluster and splits it into two clusters based on a distance metric. The procedure repeats until each cluster contains a single instance. Like agglomerative clustering, we must draw a cut line at an appropriate level in the dendrogram to generate the final clusters. For either approach, the selection of a cut line is crucial to the final clusters. In this example, the red cut line, which leads to two clusters $\{a,b\}$ and $\{c,d,e\}$, is apparently more reasonable than the blue one, which creates a singleton $\{c\}$ incorrectly. In contrast, divisive clustering provides clearer insights into the main data structure because larger clusters are generated earlier; for this reason it is less likely to suffer from accumulations of erroneous decisions [125].

Both agglomerative and divisive clustering techniques produce hierarchical trees for visualizing internal hierarchical data structures, regardless of whether data are actually hierarchically organized. A height threshold in a dendrogram can arguably be

selected according to some metric, so that clusters and their hierarchical relations above the threshold can be regarded as genuine. However, the question remains whether post-clustering analyses that are independent of the clustering process can be effective. A variant of divisive clustering, called *box clustering*, has been proposed [32]. Based on a significance test that compares the original network and a null model, box clustering verifies that the network under analysis has an inclusion/nested hierarchy, and iteratively identifies (in an unsupervised fashion) modules at each level in the hierarchy [66] until no further hierarchical levels can be located via module division. This method visualizes the final clustering result in the form of a box-model clustering tree that shows only vertical relationships between different hierarchical levels.



Complex network analysis was divided into two tasks: module identification and hierarchy construction. In addition to definitions of modules and hierarchies, the degree of difficulty of accomplishing these tasks depends on network type—for instance, whether it is weighted or directed determines problem complexity. Due to computational complexity, most current methods are limited to unweighted or undirected networks. Here I will propose a novel approach that is both general and efficient enough to be applied to networks in various domains, whether weighted or

unweighted and whether directed or undirected.

Unlike conventional top-down or bottom-up hierarchical clustering approaches [126, 127], this approach uses a two-way module-finding and hierarchy-building strategy. It initially partitions a network into disconnected modules (i.e., subgraphs) in a top-down fashion. Each module at the bottom level (containing multiple nodes) is then treated as a new node. Based on the network topology, links and their weights between new nodes can be derived to obtain a new and higher-level network. By repeating the same process it is possible to build a multi-level hierarchy from the bottom-up. Unlike most hierarchy studies that only focus on vertical relationships between modules at different levels, the proposed approach also provides explicit information for horizontal relationships via a network consisting of nodes at each hierarchical level.

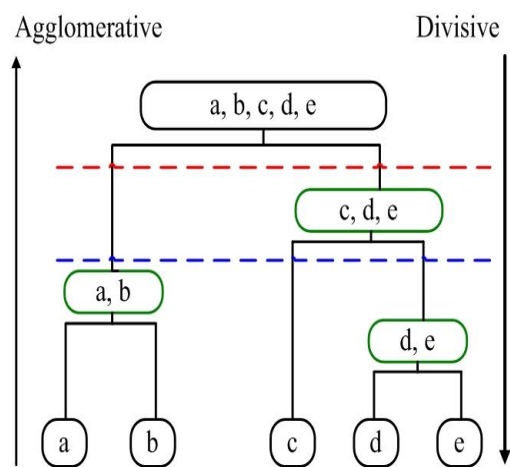


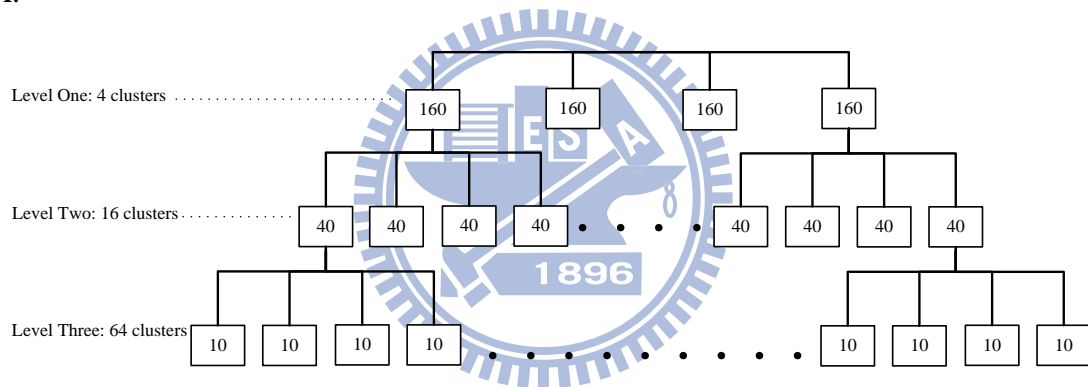
Fig 47. Example of a dendrogram from conventional hierarchical clustering.

6.2 Validation

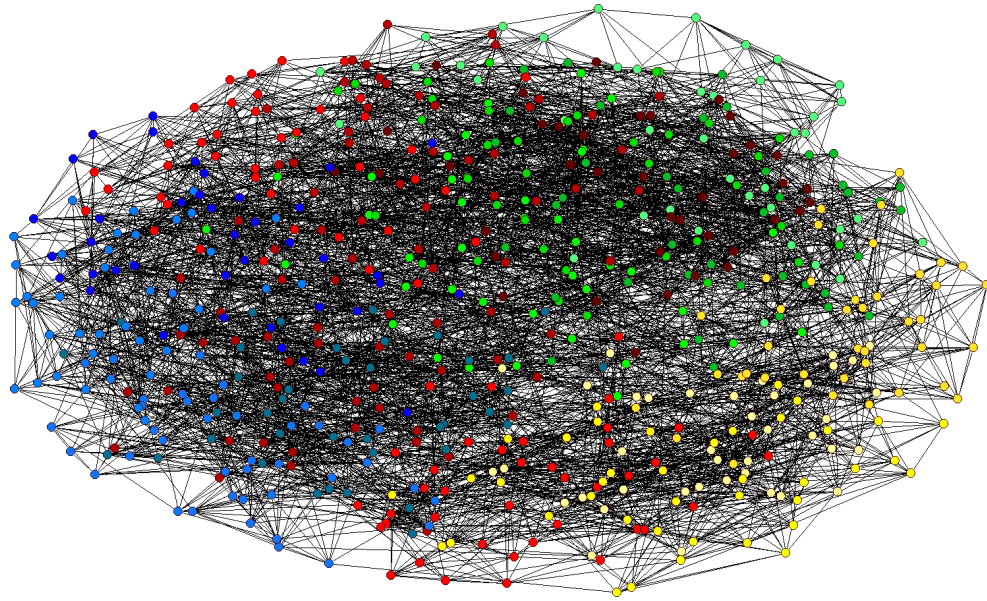
The capability of the proposed approach, named Pyramabs, was verified in order to uncover the inclusion hierarchy (based on an ensemble of random networks as proposed in [69]). Average results from tests using 30 randomly generated 640-node artificial networks are shown in Figure 48. Unlike previously established methods, Pyramabs is capable of describing explicit horizontal relationships between modules by means of abstract networks integrated within hierarchies. Figure 48 presents the abstract network at the second and third hierarchical levels for a random 640-node graph. Link color and thickness indicates the significance of connections between module pairs (red circles). Figure 48(A) shows the abstract network for the original 640-node random network. Modules were easily divided into 16 groups according to observed link color and thickness, with each group consisting of four modules. Clustering corresponded to partitioning at the second level of the hierarchy (see also Fig. 49[A]). As shown in Figure 49(B), four clusters matching the theoretical partitioning at the top hierarchical level were found. According to these results, Pyramabs is capable of detecting the inclusion hierarchy in the 640-node random

network. Furthermore, from the original network it derived weighted connections among the modules at each hierarchical level as an abstract network, thus providing greater insight into the random network.

A.



B.



C.

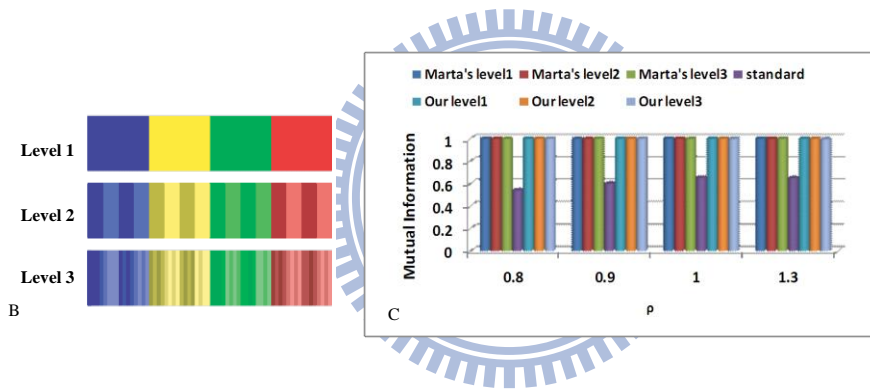
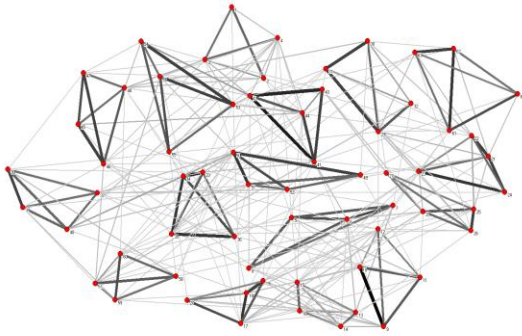


Fig 48. Validation of two-way module-finding-hierarchy-building strategy.

A.



B.

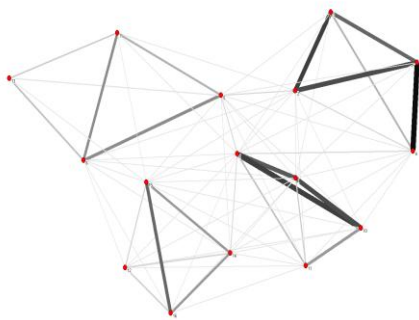
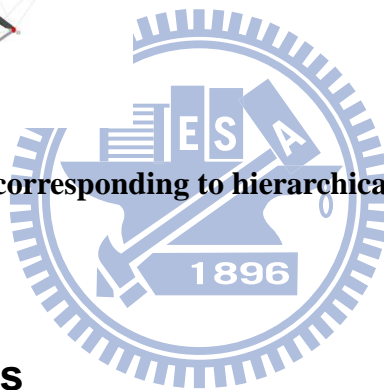


Fig 49. Abstract network corresponding to hierarchical level three and two.



6.3 Experiments

A series of experiments using the proposed approach was conducted using artificial and real-world datasets in various domains. Following the lead of Sales-Pardo et al. [69], the approach was first tested with nested random networks having a hierarchical structure. As stated above, it was possible to use artificial network data to validate the method's ability to identify inclusion hierarchies implied in networks. To evaluate its applicability to real-world problems, I tested it with three real-world datasets with

different characteristics: social networks, protein-protein interactions [12], and metabolic pathways. The results indicate that the proposed method is capable of (a) uncovering inherent hierarchies and significant modules in complex networks, and (b) providing abstractions of complex networks to different degrees.

6.3.1 Club network analysis

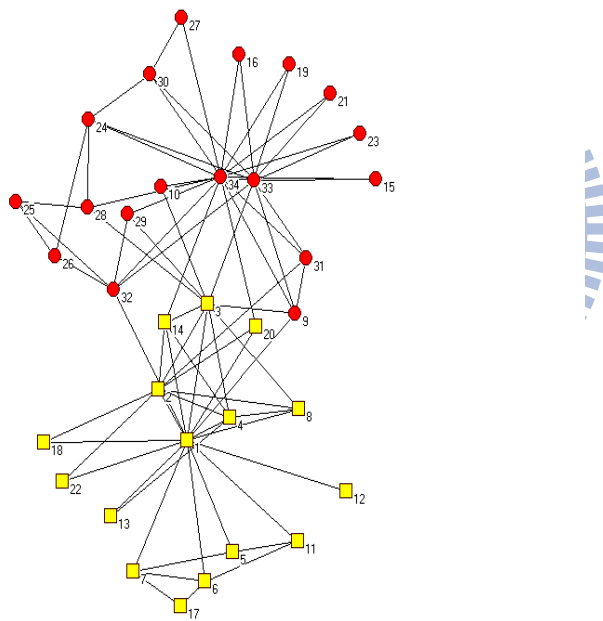
Two social networks with known module structures were selected to test Pyramabs.

The first, Zachary's karate club network, consists of 34 nodes and 78 edges [126]. It represents friendship patterns among members of a karate club at an American university over a two-year period in the early 1970s. The club split into two groups (administrators and instructors), but a previously used method [1, 69] identified a partition that agreed with the actual split with one exception: a misclassified node 3.

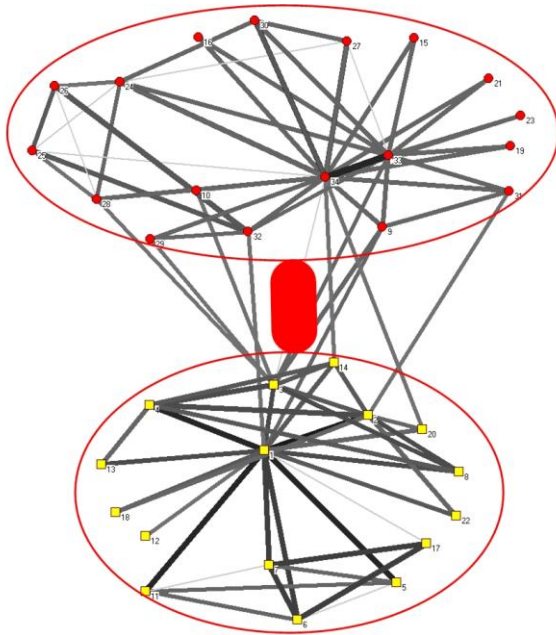
Pyramabs partitioned the club network into two modules that corresponded to the actual factions described in Zachary's study. As shown in Fig. 50, (A) The Zachary's karate club network consisting of 34 nodes and 78 edges. The nodes (i.e. club members) were categorized into two groups, administrators and instructors, colored in red circles and yellow squares respectively. (B) The proximity network of the Zachary's karate club. A thicker and darker link between nodes indicated a closer

proximity. (C) The backbone, represented by a spanning tree, extracted from the proximity network. Pyramabs identified the correct cut between node 3 and node 9. Based on the tree and the cut line, we identified two clusters as shown in (B) with two red circles connected by a thick red link. The abstract network of the two modules corresponded to the correct partition of the karate club.

A.



B.



C.

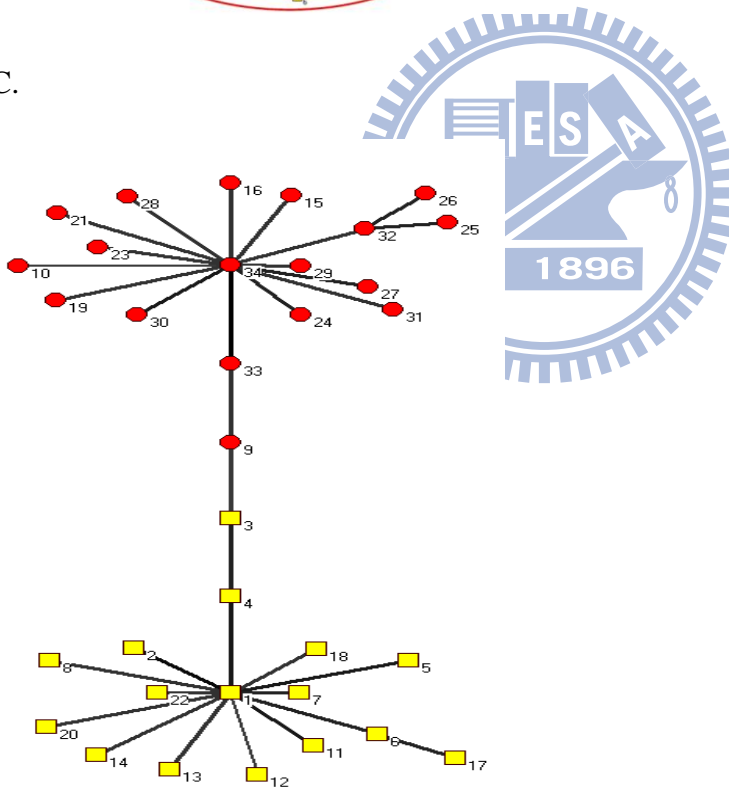
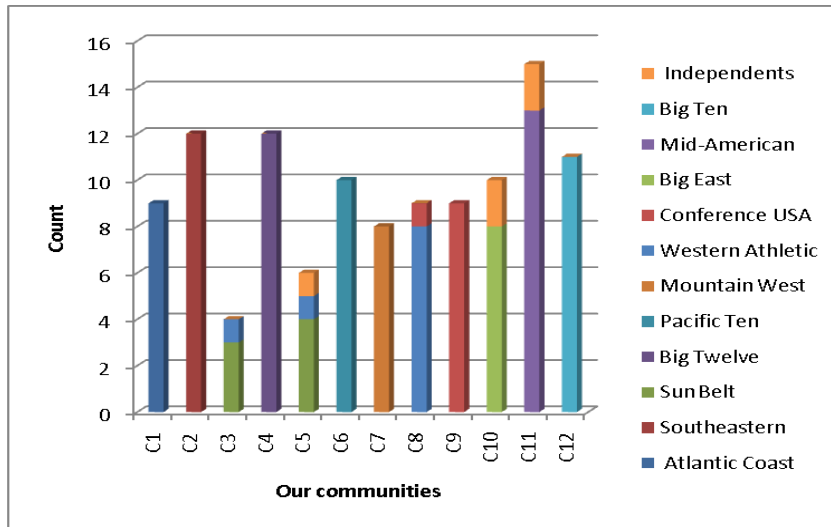


Fig. 50. Clustering results of Zachary's karate club network.

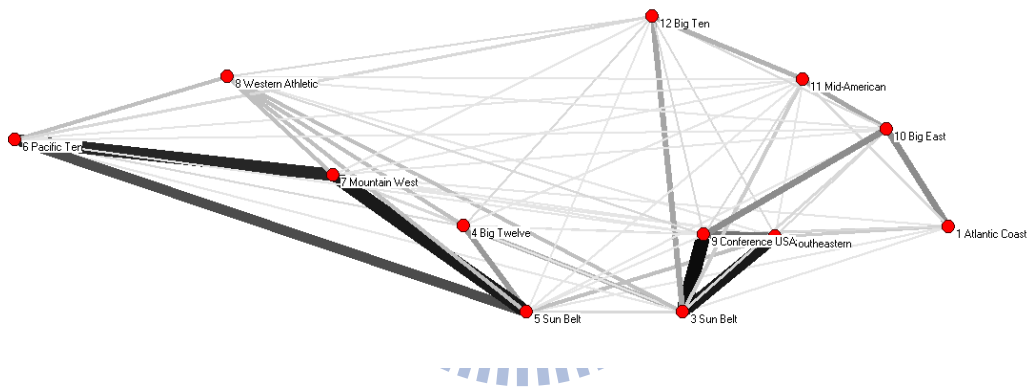
6.3.2 Football network analysis

The next test of Pyramabs used the NCAA college football network of 115 nodes and 613 links [43], representing the schedule of games between American Division I college football teams during the 2000 season. The teams were organized into different conferences, and intra-conference games were scheduled more often than inter-conference games. An abstraction pyramid consisting of two hierarchical levels was detected, and the 12 communities identified at the bottom level mapped well to the 12 conferences (Fig. 51A). Nearly all teams correctly clustered with other members in the same conference, and the scattering of independent teams across several communities reflected their lack of membership in any conference. The results shown in Figure 51A are comparable to those reported in [43]. Figure 51B shows the abstract network of communities detected by the original football game network. Unlike most previous efforts based on a bottom-layer abstract network, Pyramabs also correctly identified two modules (Fig. 51C) that were separated geographically by the Mississippi River (Fig. 51D).

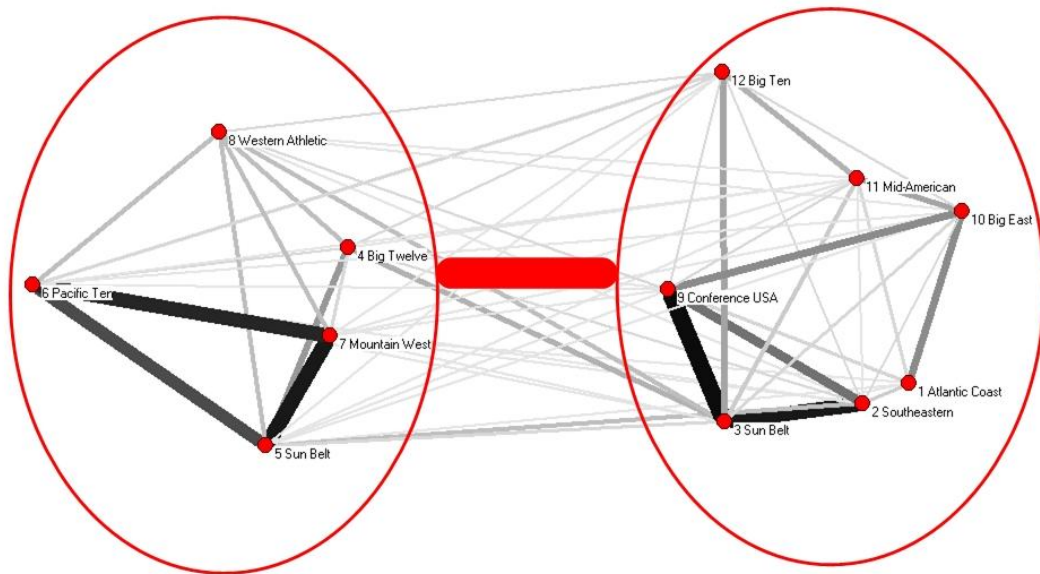
A.



B.



C.



D.

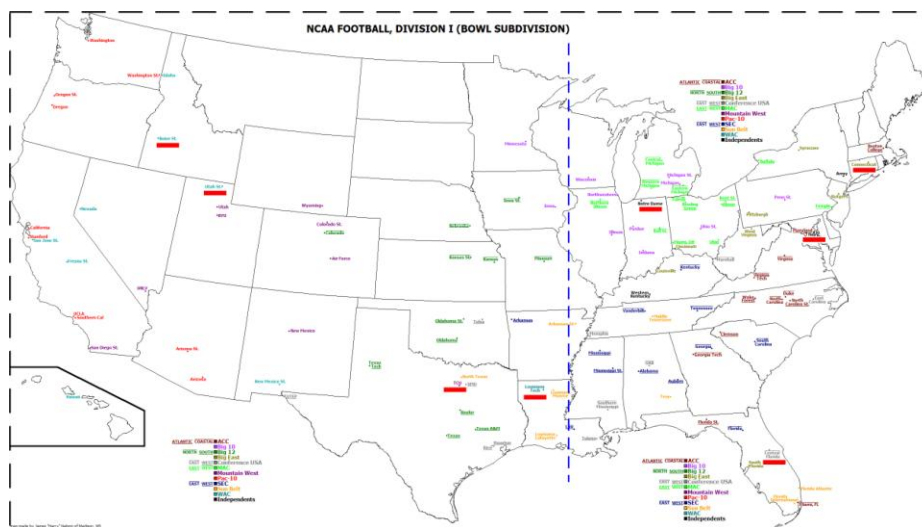


Fig. 51. The analysis of the football network

6.3.3 PPI network analysis

We tested Pyramabs on the yeast core protein interaction network downloaded from the DIP database [127]. We first applied two methods [48], the Expression Profile

Reliability Index and the Paralogous Verification Method, for reliability assessment to filter the high-throughput protein interaction data and to remove self-connecting links as previously reported [64, 72]. Then, the final interaction network was obtained, which consisted of 2440 proteins connected by 6241 links. We ran Pyramabs on this network and discovered a hierarchy of five abstraction levels. The numbers of modules at each level were 207, 72, 16, 3 and 1 from the fifth (bottom) to the first (top) level, respectively.

We evaluated the biological significance of the identified modules based on the Gene Ontology biological process annotations using Gene Ontology Go Term Finder of SGD (<http://www.yeastgenome.org/>). Based on a binomial distribution, the GO term Finder calculates a p -value that reflects the probability of observing the co-occurrence of proteins with a given GO annotation in a certain module by chance. The smaller the p -value, the more consistent is the module with the GO annotations. Our results are shown in Table 10. We also included the results of Luo *et al.* [64] and Raddichi *et al.* [72] for reference, as neither of these was capable of extracting a hierarchy from a complex network. For comparisons of hierarchy detection, we tested the same network using Sales-Pardo *et al.*'s box clustering [69].

From Table 10, the average p -value decreased as the level moved up, which suggested

that the vertical relationships in the hierarchy identified by Pyramabs corresponded to the GO hierarchy in the sense that the modules at lower levels were merged correctly into larger modules at higher levels. They also mapped well to the GO annotation categories. In addition, we conducted a series of Monte Carlo tests to obtain a baseline for the p -value using random modules; the background p -values (shown in parentheses) in Table 10 are for reference.

We also analyzed the horizontal relationship in the abstract network at each level to ascertain if it could characterize biological meanings. The proximity between two (super)nodes in an abstract network, as defined in eq. [2] (see Methods), measures the significance of the relationship between the nodes. Given two pairs of nodes in the abstract network, (a, b) and (c, d) , when the proximity between a and b is P_{ab} and the proximity between c and d is P_{cd} , $P_{ab} \geq P_{cd}$ suggests that a and b have a closer relationship to each other than c and d .

In our analysis of protein-protein interactions, we verified if a and b had a closer biological relationship than c and d when $P_{ab} \geq P_{cd}$ by evaluating the change in p -value before and after merging nodes (i.e. modules). We ran a sign test on the abstract network at each level in the hierarchy. There were a significant number of positive cases for which the ratio of the decrease in p -value after merging a and b was

larger than that after merging c and d when $P_{ab} \geq P_{cd}$ (at significance level 0.01).

These results demonstrated the feasibility of applying the horizontal relationship for characterizing biological meanings.

In Fig. 53, owing to the complexity, we showed the maximum spanning tree of the abstract network at levels 4 and 5 in the hierarchy instead of the abstract networks themselves. To further describe the vertical and the horizontal relationships, we selected two examples in Fig. 53 and elaborated on them in Fig. 54 and 55.



Table 10. Summary of biological significance of modules based on GO biological**process annotations**

	Total Clusters	Average Cluster Size	Average <i>p</i> -value
Pyramabs (Level 2)	3	723.33	1.04E-69 (1.67E-52) ^b
Pyramabs (Level 3)	16	152.44	2.21E-25 (1.46E-12) ^b
Pyramabs (Level 4)	72	33.86	7.32E-18 (5.67E-07) ^b
Pyramabs (Level 5)	207	8.54	4.65E-10 (3.66E-04) ^b
Luo <i>et al.</i> ^a	86	19.20	2.99E-17
Raddichi <i>et al.</i> ^a	155	12.82	3.82E-13
Sales-Pardo <i>et al.</i> (Level 2) ^c	76	26.41	1.04E-16
Sales-Pardo <i>et al.</i> (Level 3) ^c	101	11.44	4.36E-13
Sales-Pardo <i>et al.</i> (Level 4) ^c	88	7.76	3.51E-08
Sales-Pardo <i>et al.</i> (Level 5) ^c	12	5.37	3.51E-05

Table 1. Footnotes

^aBoth Luo *et al.*'s and Raddichi *et al.*'s methods could only identify single-level modules.

^bNumbers in parentheses are baseline *p*-values obtained by Monte Carlo tests. All baseline *p*-values are larger than the observed (e.g. 1.67E-52 vs. 1.04E-69), suggesting that observed *p*-values are not by chance.

^cIn Sales-Pardo *et al.*'s method, a higher-level module will not necessarily be partitioned further into lower-level sub-modules. Thus, the number of modules does

not necessarily increase as the level goes down (e.g. 88 modules at level 4, but only 12 modules at level 5).

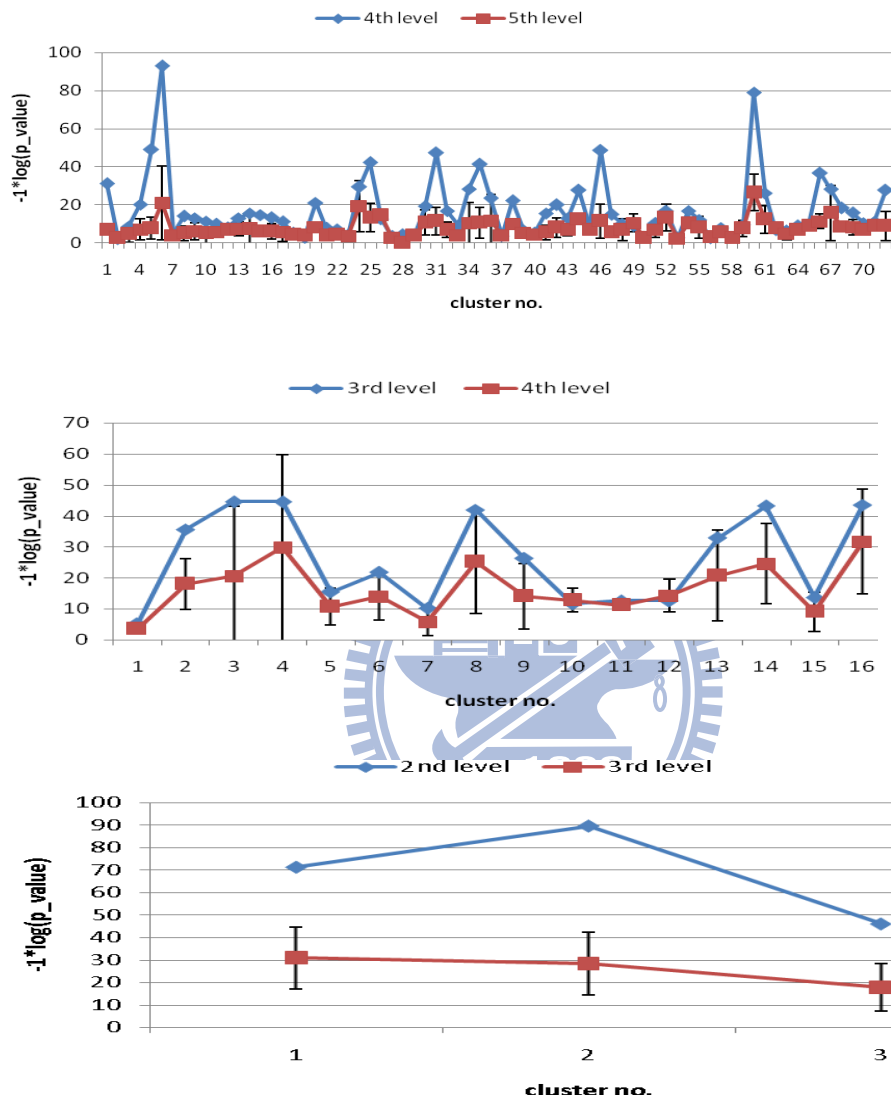


Fig. 52 The p-value of the corresponding nodes at different levels.

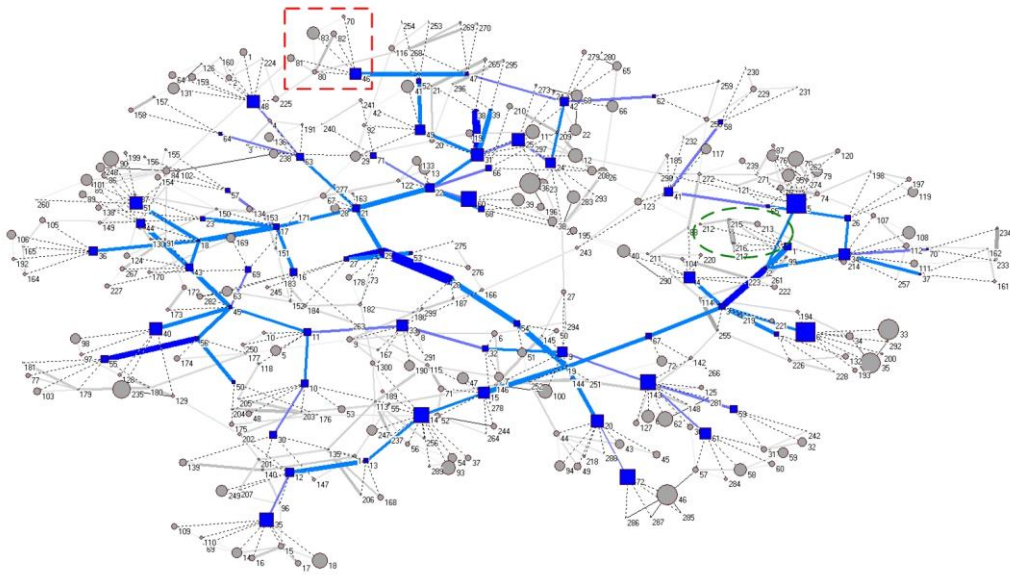


Fig. 53. The MST of PPI at level four and five, and have blue and red colors, respectively.

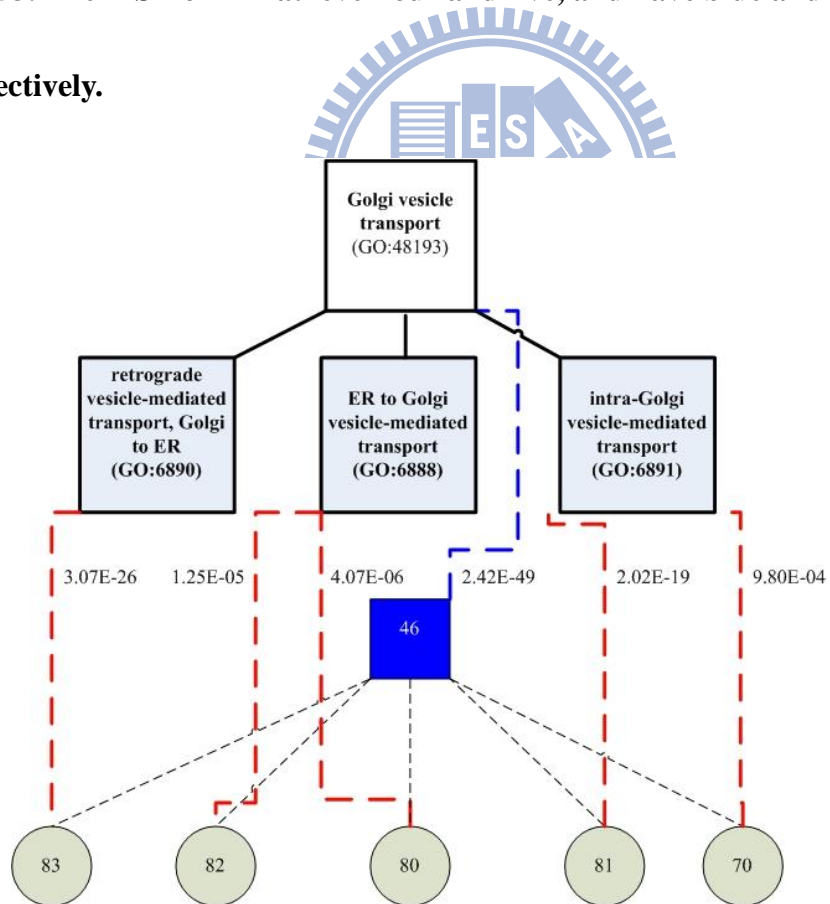


Fig. 54. The mapping between the modules we found and the real GO id.

6.3.4 Metabolic network analysis

Thousands of components in a living cell are dynamically interconnected as a complex network that determines the cell's functional properties [8, 36]. One of the primary examples is cellular metabolism that arises from sophisticated biochemical networks in which numerous metabolites are integrated through biochemical reactions.

To facilitate the identification and characterization of system-level features in biological organizations, we can partition cellular functionality into a collection of modules and organize them in a hierarchy [67].

We tested Pyramabs on the metabolic network of *E. coli* as used previously [69]. This contained 507 nodes and 947 links, where each node represented a metabolic substrate and a link described a reaction. As Pyramabs is flexible enough to deal with undirected or directed networks, the reactions in the metabolic network were treated as undirected and directed, respectively, in different tests for hierarchy discovery.

In KEGG [128], metabolic pathways are classified into 11 categories: Carbohydrate, Energy, Lipid, Nucleotide, Amino acid, Other amino acid, Glycan, PK/NRP, Cofactor/vitamin, Secondary metabolite and Xenobiotics. Each category consists of

several sub-categories (e.g., nucleotide metabolism includes purine metabolism and pyrimidine metabolism). In addition to KEGG PATHWAY classifications, KEGG also provides *pathway modules* that are specifications of sub-networks corresponding to tighter functional units. We measured the within-module consistency of metabolic pathway classification according to KEGG by the p -value based on a hypergeometric distribution. These results are summarized in Table 11(a) and 11(b) after treating the links as undirected and directed, respectively. For a comparative study, we tested the same network using Sales-Pardo *et al.*'s method of box clustering [69], which considered undirected networks only; these results are in Table 11(c).

Figure 55 shows the abstraction pyramid extracted from the metabolic network. To enhance readability, we only show the maximum spanning tree of the abstract network at each level. An example of the vertical relationship between different hierarchical levels is marked by red circles for further analysis (Fig. 56). We have also highlighted by a red rectangle an example of the horizontal relationship at the second level, and compared it against the KEGG PATHWAY (Fig. 57). The vertical relationships disclosed by Pyramabs correctly characterized *inclusion* (or *part of*) relations (e.g. "Pyrimidine metabolism" is included in "Nucleotide metabolism."); the horizontal relationships showed that the modules with a larger proximity in between

belonged to the same pathway category.

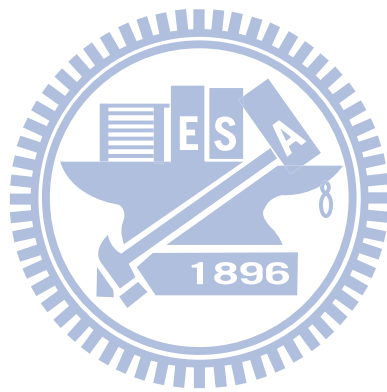


Table 11. Summary of within-module consistency of metabolic pathway

classification based on KEGG.

A.

	Total Clusters	Average Cluster Size	PATHWA Y Category Average <i>p</i> -value	PATHWAY Sub-category Average <i>p</i> -value	Pathway Module Average <i>p</i> -value
Level 2	6	84.5	4.91E-20 (1.52E-02)	3.35E-22 (3.37E-05)	4.88E-09 (2.17E-03)
Level 3	26	19.5	2.78E-10 (5.95E-03)	6.68E-15 (3.29E-03)	4.24E-09 (1.53E-02)
Level 4	104	4.9	2.16E-04 (2.35E-02)	1.13E-07 (6.19E-03)	1.74E-05 (1.52E-02)

B.

	Total Clusters	Average Cluster Size	PATHWA Y Category Average <i>p</i> -value	PATHWAY Sub-category Average <i>p</i> -value	Pathway Module Average <i>p</i> -value
Level 2	5	101.4	3.78E-11 (2.59E-06)	5.27E-16 (5.81E-05)	9.41E-11 (2.07E-03)
Level 3	27	18.8	5.05E-10 (8.37E-03)	7.10E-16 (2.82E-03)	1.57E-09 (1.32E-02)
Level 4	117	4.3	4.98E-04 (2.29E-02)	4.75E-07 (5.08E-03)	4.59E-05 (1.34E-02)

C.

	Total Clusters	Average Cluster Size	PATHWA Y Category Average <i>p</i> -value	PATHWAY Sub-category Average <i>p</i> -value	Pathway Module Average <i>p</i> -value
Level 2	28	18.1	1.82E-08	3.17E-13	1.35E-08
Level 3	111	3.2	1.49E-04	8.80E-08	1.75E-05
Level 4	48	2.8	5.28E-03	6.41E-05	1.64E-03
Level 5	50	2.7	6.51E-03	9.05E-05	1.65E-03

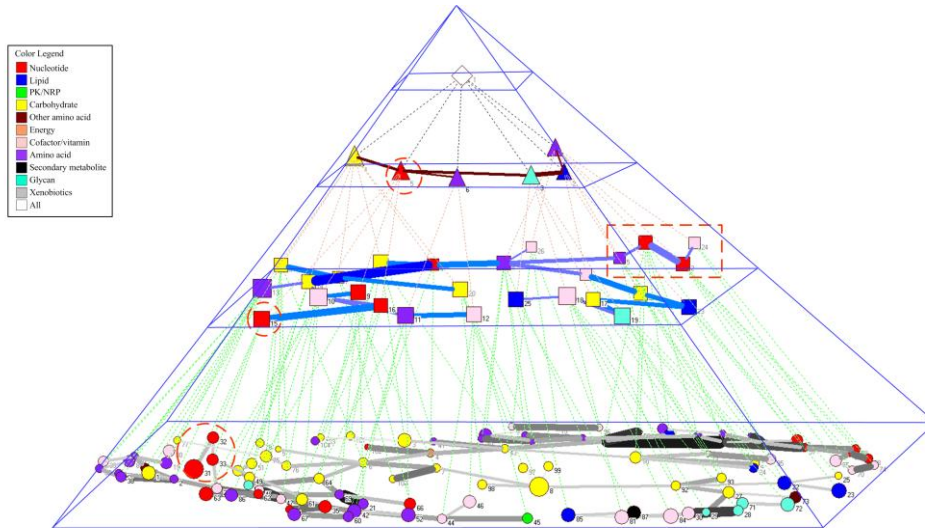


Fig. 55. The pyramid of abstraction disclosed from a metabolic network.

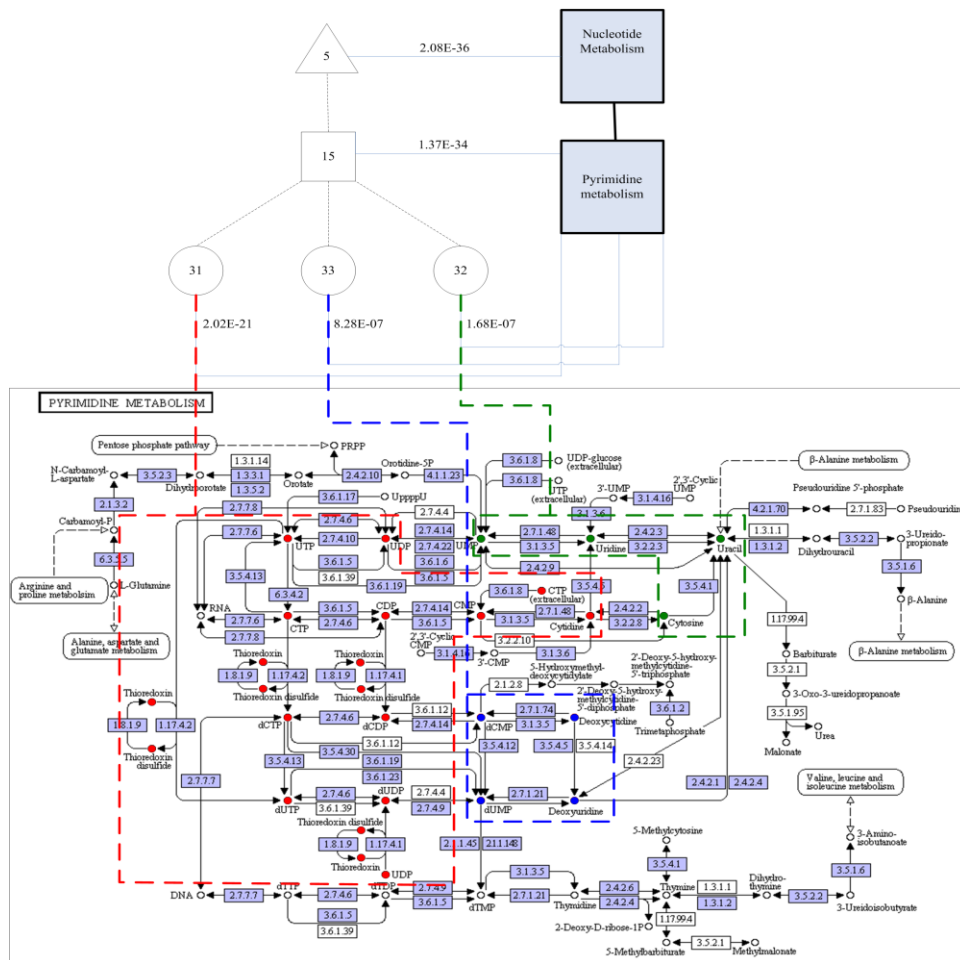


Fig. 56. Example of the vertical relationships in an abstraction pyramid disclosed

from a metabolic network.

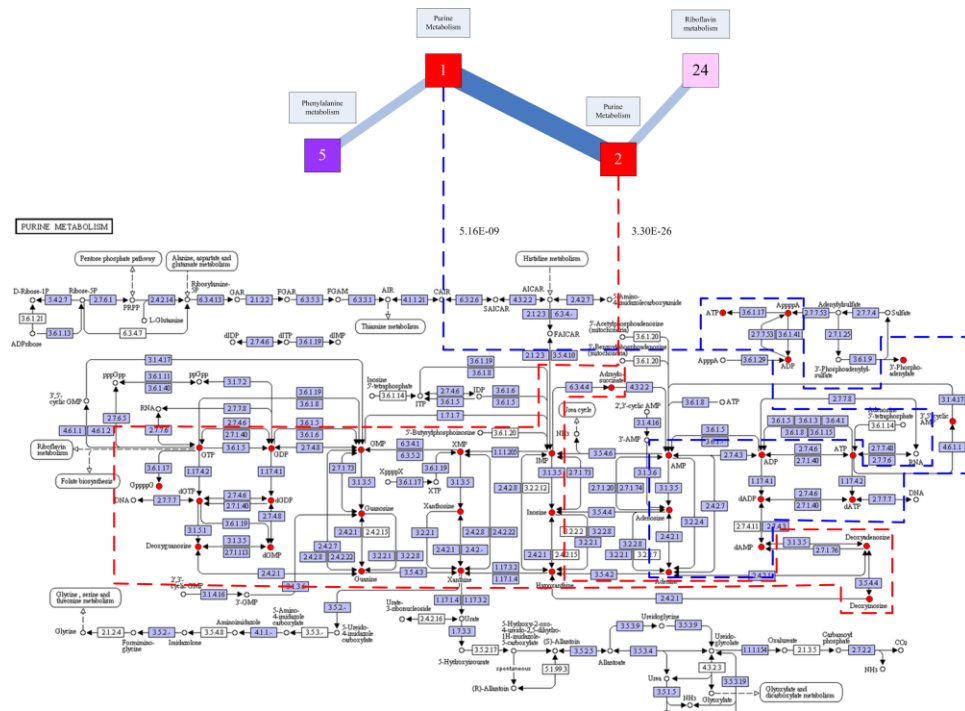


Fig. 57. Example of the horizontal relationship at the third level of an abstraction pyramid.

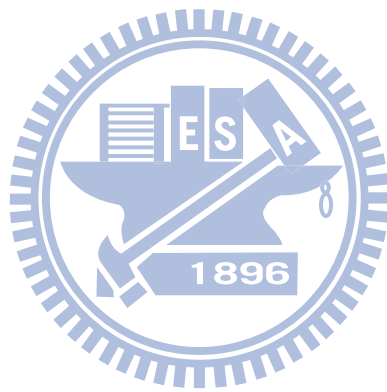


Chapter 7 Conclusion

The emergence of large and reliable network maps has driven the development of network theory during the past decade. For example, cell biologists use networks to understand signal transduction cascades and metabolism, computer scientists use them to map the Internet and World Wide Web, and epidemiologists use them to study virus transmission networks. However, the question remains whether their efforts can be used to form a theory of complexity or to determine the common characteristics of dynamic processes.

The shift from studying a small number of elements to studying the behavior of large-scale aggregates is equivalent to the shift from atomic and molecular physics to the physics of matter. Understanding how the same elements assembled in large numbers can give rise to different macroscopic and dynamical behaviors (according to the various forces and elements) opens potential paths to quantitative computational approaches and increased forecasting power.

Although many findings and studies have been offered and rejected, it is increasingly clear that interconnectivity and topology are fundamental to the behavior of complex systems. In other words, complex networks are here to stay.



Chapter 8 Reference

- [1] A. L. Barabasi and Z. N. Oltvai, "Network biology: Understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, pp. 101-U15, Feb 2004.
- [2] Z. N. Oltvai and A. L. Barabasi, "Life's complexity pyramid," *Science*, vol. 298, pp. 763-764, Oct 25 2002.
- [3] M. E. J. Newman, "The structure and function of complex networks," *Siam Review*, vol. 45, pp. 167-256, Jun 2003.
- [4] C. M. Song, S. Havlin, and H. A. Makse, "Origins of fractality in the growth of complex networks," *Nature Physics*, vol. 2, pp. 275-281, Apr 2006.
- [5] M. E. J. Newman, "The structure and function of complex networks," *Siam Review*, vol. 45, pp. 167-256, 2003.
- [6] M. E. J. Newman, "Scientific collaboration networks. I. Network construction and fundamental results," *Physical Review E*, vol. 64, Jul 2001.
- [7] C. M. Song, S. Havlin, and H. A. Makse, "Self-similarity of complex networks," *Nature*, vol. 433, pp. 392-395, Jan 2005.
- [8] H. Kitano, "Systems biology: A brief overview," *Science*, vol. 295, pp.

1662-1664, Mar 2002.

- [9] D. K. Smith, "NETWORK FLOWS - THEORY, ALGORITHMS, AND APPLICATIONS - AHUJA,RK, MAGNANTI,TL, ORLIN,J," *Journal of the Operational Research Society*, vol. 45, pp. 1340-1340, Nov 1994.
- [10] R. Dobrin, Q. K. Beg, A. L. Barabasi, and Z. N. Oltvai, "Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network," *BMC Bioinformatics*, vol. 5, pp. -, Jan 30 2004.
- [11] N. Shadbolt and T. Berners-Lee, "Web science emerges," *Scientific American (International Edition)*, pp. 60-5, 2008.
- [12] E. J. Deeds, N. V. Dokholyan, and E. I. Shakhnovich, "Protein evolution within a structural space," *Biophysical Journal*, vol. 85, pp. 2962-2972, 2003.
- [13] K. M. Bryden, D. A. Ashlock, S. Corns, and S. J. Willson, "Graph-based evolutionary algorithms," *Ieee Transactions on Evolutionary Computation*, vol. 10, pp. 550-567, 2006.
- [14] F. C. Santos, J. F. Rodrigues, and J. M. Pacheco, "Graph topology plays a determinant role in the evolution of cooperation," *Proceedings of the Royal Society B-Biological Sciences*, vol. 273, pp. 51-55, 2006.
- [15] E. Lieberman, C. Hauert, and M. A. Nowak, "Evolutionary dynamics on

- graphs," *Nature*, vol. 433, pp. 312-316, 2005.
- [16] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, pp. 824-827, Oct 2002.
- [17] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Physical Review E*, vol. 64, Aug 2001.
- [18] D. S. Callaway, M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Network robustness and fragility: Percolation on random graphs," *Physical Review Letters*, vol. 85, pp. 5468-5471, Dec 2000.
- [19] D. Garlaschelli, "The weighted random graph model," *New Journal of Physics*, vol. 11, 2009.
- [20] M. E. J. Newman, "Detecting community structure in networks," *European Physical Journal B*, vol. 38, pp. 321-330, 2004.
- [21] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440-442, Jun 1998.
- [22] E. Estrada, "Topological structural classes of complex networks," *Physical Review E*, vol. 75, 2007.

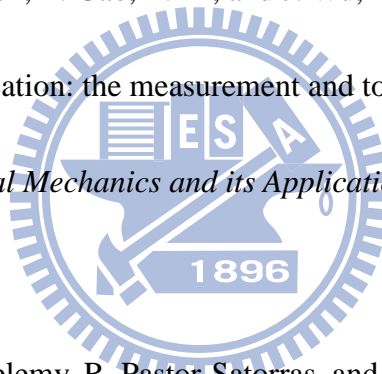
- [23] H. Ohtsuki, C. Hauert, E. Lieberman, and M. A. Nowak, "A simple rule for the evolution of cooperation on graphs and social networks," *Nature*, vol. 441, pp. 502-505, 2006.
- [24] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509-512, Oct 1999.
- [25] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, "Complex networks: Structure and dynamics," *Physics Reports-Review Section of Physics Letters*, vol. 424, pp. 175-308, 2006.
- [26] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon, "Superfamilies of evolved and designed networks," *Science*, vol. 303, p. 1538, 2004.
- [27] M. E. J. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, Nov 2002.
- [28] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna, "A tool for filtering information in complex systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 10421-10426, Jul 2005.
- [29] M. Girvan and M. E. J. Newman, "Community structure in social and

- biological networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 7821-7826, Jun 2002.
- [30] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, pp. 910-913, May 3 2002.
- [31] J. Schiff, "Cellular automata: a discrete view of the world," *Discrete Mathematics and Optimization, John Wiley and Sons, Inc, Hoboken, New Jersey*, 2007.
- [32] S. Fortunato and M. Barthelemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 36-41, Jan 2007.
- [33] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, vol. 3, Mar 2007.
- [34] J. A. Freyre-Gonzalez, J. A. Alonso-Pavon, L. G. Trevino-Quintanilla, and J. Collado-Vides, "Functional architecture of Escherichia coli: new insights provided by a natural decomposition approach," *Genome Biology*, vol. 9, p. 40, 2008.
- [35] Z. Wang and J. Z. Zhang, "In search of the biological significance of modular structures in protein networks," *Plos Computational Biology*, vol. 3, pp.

1011-1021, Jun 2007.

- [36] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C47-C52, Dec 1999.
- [37] C. Huang, C. Cheng, and C. Sun, "Bridge and brick network motifs: Identifying significant building blocks from complex biological systems," *Artificial Intelligence In Medicine*, vol. 41, pp. 117-127, 2007.
- [38] J. Bascompte, C. Melian, and E. Sala, "Interaction strength combinations and the overfishing of a marine food web," *Proceedings of the National Academy of Sciences*, vol. 102, p. 5443, 2005.
- [39] M. Newman, "The structure and function of complex networks," *Arxiv preprint cond-mat/0303516*, 2003.
- [40] M. Granovetter, "The strength of weak ties: A network theory revisited," *Sociological theory*, vol. 1, pp. 201-233, 1983.
- [41] J. Davidsen, H. Ebel, and S. Bornholdt, "Emergence of a small world from local interactions: Modeling acquaintance networks," *Physical Review Letters*, vol. 88, p. 128701, 2002.
- [42] S. Teichmann and M. Babu, "Gene regulatory network growth by duplication," *Nature Genetics*, vol. 36, pp. 492-496, 2004.

- [43] M. Girvan and M. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, p. 7821, 2002.
- [44] H. Zhu and Z. Huang, "Navigation in a small world with local information," *Physical Review E*, vol. 70, p. 36117, 2004.
- [45] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, Feb 2004.
- [46] M. Li, Y. Fan, J. Chen, L. Gao, Z. Di, and J. Wu, "Weighted networks of scientific communication: the measurement and topological role of weight," *Physica A: Statistical Mechanics and its Applications*, vol. 350, pp. 643-656, 2005.
- [47] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 101, p. 3747, 2004.
- [48] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions - Two methods for assessment of the reliability of high throughput observations," *Molecular & Cellular Proteomics*, vol. 1, pp. 349-356, May 2002.



- [49] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone, "Comment on "network motifs: Simple building blocks of complex networks" and "superfamilies of evolved and designed networks", " *Science*, vol. 305, 2004.
- [50] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, and U. Alon, "Response to comment on "network motifs: Simple building blocks of complex networks" and "superfamilies of evolved and designed networks", " *Science*, vol. 305, 2004.
- [51] D. Goldberg and F. Roth, "Assessing experimentally derived interactions in a small world," *Proceedings of the National Academy of Sciences*, vol. 100, p. 4372, 2003.
- [52] G. F. Davis, Y. Mina, and W. E. Baker, "The small world of the American corporate elite, 1982-2001," *Strategic Organization*, vol. 1, pp. 301-26, 2003.
- [53] S. Qin and G. Z. Dai, "A new local-world evolving network model," *Chinese Physics B*, vol. 18, pp. 383-390, 2009.
- [54] M. Boguna, R. Pastor-Satorras, and A. Vespignani, "Epidemic spreading in complex networks with degree correlations," *Statistical Mechanics of Complex Networks. 18th Sitges Conference*, pp. 127-47|xii+206, 2003.
- [55] J. Gomez-Gardenes, P. Echenique, and Y. Moreno, "Immunization of real

- complex communication networks," *European Physical Journal B*, vol. 49, pp. 259-264, 2006.
- [56] Y. Moreno, J. B. Gomez, and A. F. Pacheco, "Epidemic incidence in correlated complex networks," *Physical Review E*, vol. 68, 2003.
- [57] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "Critical phenomena in complex networks," *Reviews of Modern Physics*, vol. 80, pp. 1275-1335, 2008.
- [58] M. Simoes, M. M. T. da Gama, and A. Nunes, "Stochastic fluctuations in epidemics on networks," *Journal of the Royal Society Interface*, vol. 5, pp. 555-566, 2008.
- [59] Y. Moreno, J. B. Gomez, and A. F. Pacheco, "Epidemic incidence in correlated complex networks," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 68, p. 035103, 2003.
- [60] Z. H. Liu, Y. C. Lai, and N. Ye, "Propagation and immunization of infection on general networks with both homogeneous and heterogeneous components," *Physical Review E*, vol. 67, 2003.
- [61] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The Computer Journal*, vol. 26, p. 354, 1983.

- [62] P. Macnaughton-Smith, W. Williams, M. Dale, and L. Mockett, "Dissimilarity analysis: A new technique of hierarchical sub-division," 1964.
- [63] M. A. Serrano, M. Boguna, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 6483-6488, Apr 2009.
- [64] F. Luo, Y. F. Yang, C. F. Chen, R. Chang, J. Z. Zhou, and R. H. Scheuermann, "Modular organization of protein interaction networks," *Bioinformatics*, vol. 23, pp. 916-916, Apr 2007.
- [65] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 8577-8582, Jun 2006.
- [66] V. Arnau, S. Mars, and I. Marin, "Iterative cluster analysis of protein interaction data," *Bioinformatics*, vol. 21, pp. 364-378, Feb 2005.
- [67] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551-1555, Aug 2002.
- [68] H. N. Chua, W. K. Sung, and L. S. Wong, "Exploiting indirect neighbours and

- topological weight to predict protein function from protein-protein interactions," *Bioinformatics*, vol. 24, pp. 452-452, Feb 2008.
- [69] M. Sales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, "Extracting the hierarchical organization of complex systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 15224-15229, Sep 2007.
- [70] A. M. Yip and S. Horvath, "Gene network interconnectedness and the generalized topological overlap measure," *Bmc Bioinformatics*, vol. 8, Jan 2007.
- [71] C. T. Zahn, "Graph-theoretical methods for detecting and describing Gestalt clusters," *IEEE Transactions on Computers*, vol. C-20, pp. 68-86, 1971.
- [72] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 2658-2663, Mar 2 2004.
- [73] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814-818, Jun 9 2005.

- [74] M. P. Samanta and S. Liang, "Predicting protein functions from redundancies in large-scale protein interaction networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 12579-12583, Oct 2003.
- [75] S. Strogatz, "Exploring complex networks," *Nature*, vol. 410, pp. 268-276, 2001.
- [76] M. Newman and D. Watts, "Scaling and percolation in the small-world network model," *Physical Review E*, vol. 60, pp. 7332-7342, 1999.
- [77] R. Cancho, C. Janssen, and R. Sole, "Topology of technology graphs: Small world patterns in electronic circuits," *Physical Review E*, vol. 64, p. 46119, 2001.
- [78] A. Barrat and M. Weigt, "On the properties of small-world network models," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 13, pp. 547-560, 2000.
- [79] R. Albert, "Scale-free networks in cell biology," *Journal of cell science*, vol. 118, p. 4947, 2005.
- [80] S. Mangan and U. Alon, "Structure and function of the feed-forward loop network motif," *Proceedings of the National Academy of Sciences of the*

- United States of America*, vol. 100, p. 11980, 2003.
- [81] J. Rice, A. Kershenbaum, and G. Stolovitzky, "Lasting impressions: Motifs in protein–protein maps may provide footprints of evolutionary events," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, p. 3173, 2005.
- [82] S. Mangan, A. Zaslaver, and U. Alon, "The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks," *Journal of Molecular Biology*, vol. 334, pp. 197-204, 2003.
- [83] J. Stelling, U. Sauer, Z. Szallasi, F. Doyle, and J. Doyle, "Robustness of cellular functions," *Cell*, vol. 118, pp. 675-685, 2004.
- [84] R. Dobrin, Q. Beg, A. Barabasi, and Z. Oltvai, "Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network," *BMC bioinformatics*, vol. 5, p. 10, 2004.
- [85] Y. Artzy-Randrup, S. Fleishman, N. Ben-Tal, and L. Stone, "Comment on" Network Motifs: Simple Building Blocks of Complex Networks" and" Superfamilies of Evolved and Designed Networks"," *Science*, vol. 305, p. 1107c, 2004.
- [86] A. Vazquez, R. Dobrin, D. Sergi, J. Eckmann, Z. Oltvai, and A. Barabasi, "The

topological relationship between the large-scale attributes and local interaction patterns of complex networks," *Proceedings of the National Academy of Sciences*, vol. 101, p. 17940, 2004.

[87] H. Moon, J. Bhak, K. Lee, and D. Lee, "Architecture of basic building blocks in protein and domain structural interaction networks," *Bioinformatics*, vol. 21, p. 1479, 2005.

[88] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of *Escherichia coli*," *Nature Genetics*, vol. 31, pp. 64-68, 2002.

[89] J. Berg and M. Lassig, "Local graph alignment and motif search in biological networks," *Proceedings of the National Academy of Sciences*, vol. 101, p. 14689, 2004.

[90] M. Middendorf, E. Ziv, and C. Wiggins, "Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network," *Proceedings of the National Academy of Sciences*, vol. 102, p. 3192, 2005.

[91] H. Hethcote, "The mathematics of infectious diseases," *SIAM review*, vol. 42, pp. 599-653, 2000.

[92] J. Eckmann and E. Moses, "Curvature of co-links uncovers hidden thematic

- layers in the world wide web," *Proceedings of the National Academy of Sciences*, vol. 99, p. 5825, 2002.
- [93] S. Manrubia and J. Poyatos, "Motif selection in a model of evolving replicators: The role of surfaces and limited transport in network topology," *Europhysics Letters*, vol. 64, pp. 557-563, 2003.
- [94] R. van der Hofstad and G. Hooghiemstra, "Universality for distances in power-law random graphs," *Journal of Mathematical Physics*, vol. 49, 2008.
- [95] S. Kundu, L. Huisman, I. Nair, V. Iyengar, and L. Reddy, "A small test generator for large designs," 1992, pp. 30-40.
- [96] D. MacRae Jr, "Direct factor analysis of sociometric data," *Sociometry*, vol. 23, pp. 360-371, 1960.
- [97] L. Zeleny, "Adaptation of research findings in social leadership to college classroom procedures," *Sociometry*, pp. 314-328, 1950.
- [98] H. W. Ma, J. Buer, and A. P. Zeng, "Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach," *BMC Bioinformatics*, vol. 5, Dec 2004.
- [99] H. Y. Yu and M. Gerstein, "Genomic analysis of the hierarchical structure of regulatory networks," in *Colloquium on Frontiers in Bioformatics - Unsolved*

Problem and Challenges, Irvine, CA, 2004, pp. 14724-14731.

- [100] R. Williams and N. Martinez, "Simple rules yield complex food webs," *Nature*, vol. 404, pp. 180-183, 2000.
- [101] J. Shurin, E. Borer, E. Seabloom, K. Anderson, C. Blanchette, B. Broitman, S. Cooper, and B. Halpern, "A cross-ecosystem comparison of the strength of trophic cascades," *Ecology Letters*, vol. 5, pp. 785-791, 2002.
- [102] M. Costanzo, M. Crawford, J. Hirschman, J. Kranz, P. Olsen, L. Robertson, M. Skrzypek, B. Braun, K. Hopkins, and P. Kondu, "YPDTM, PombePDTM and WormPDTM: model organism volumes of the BioKnowledge™ Library, an integrated resource for protein information," *Nucleic Acids Research*, vol. 29, p. 75, 2001.
- [103] C. V. Robinson, A. Sali, and W. Baumeister, "The molecular sociology of the cell," *Nature*, vol. 450, pp. 973-982, Dec 13 2007.
- [104] S. H. Zhang, G. X. Jin, X. S. Zhang, and L. N. Chen, "Discovering functions and revealing mechanisms at molecular level from biological networks," *Proteomics*, vol. 7, pp. 2856-2869, Aug 2007.
- [105] G. Conant and A. Wagner, "Convergent evolution of gene circuits," *Nature Genetics*, vol. 34, pp. 264-266, 2003.

- [106] S. Zhao, Q. Zhu, and R. Somerville, "The sigma 70 transcription factor TyrR has zinc-stimulated phosphatase activity that is inhibited by ATP and tyrosine," *Journal of Bacteriology*, vol. 182, p. 1053, 2000.
- [107] J. Sarsero, P. Wookey, and A. Pittard, "Regulation of expression of the *Escherichia coli* K-12 mtr gene by TyrR protein and Trp repressor," *Journal of Bacteriology*, vol. 173, p. 4133, 1991.
- [108] R. Prill, P. Iglesias, and A. Levchenko, "Dynamic properties of network motifs contribute to biological network organization."
- [109] D. H. Zanette, "Models of social processes on small-world networks," *Modern Challenges in Statistical Mechanics: Patterns, Noise, and the Interplay of Nonlinearity and Complexity*, vol. 658, pp. 187-203, 2003.
- [110] M. E. J. Newman and D. J. Watts, "Renormalization group analysis of the small-world network model," *Physics Letters A*, vol. 263, pp. 341-346, 1999.
- [111] E. Bruce, *Biomedical signal processing and signal modeling*: Wiley New York:, 2001.
- [112] G. Davis, M. Yoo, and W. Baker, "The small world of the American corporate elite, 1982-2001," *Strategic organization*, vol. 1, p. 301, 2003.
- [113] D. Zanette, "Dynamics of rumor propagation on small-world networks,"

- Physical Review E*, vol. 65, p. 41908, 2002.
- [114] P. Erdos and A. Renyi, "On the evolution of random graphs," *Publication*, vol. 5, pp. 17–61.
- [115] R. Pastor-Satorras and A. Vespignani, "Immunization of complex networks," *Physical Review E*, vol. 65, p. 36104, 2002.
- [116] R. Pastor-Satorras and A. Vespignani, "Epidemics and immunization in scale-free networks," *Arxiv preprint cond-mat/0205260*, 2002.
- [117] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical Review Letters*, vol. 86, pp. 3200-3203, 2001.
- [118] C. Huang, C. Sun, and H. Lin, "Influence of local information on social simulations in small-world network models," *Journal of Artificial Societies and Social Simulation*, vol. 8, p. 8, 2005.
- [119] D. Bray, "Molecular networks: The top-down view," *Science*, vol. 301, pp. 1864-1865, Sep 26 2003.
- [120] J. Gagneur, R. Krause, T. Bouwmeester, and G. Casari, "Modular decomposition of protein-protein interaction networks," *Genome Biology*, vol. 5, 2004.
- [121] J. Zhao, G. H. Ding, L. Tao, H. Yu, Z. H. Yu, J. H. Luo, Z. W. Cao, and Y. X.

- Li, "Modular co-evolution of metabolic networks," *Bmc Bioinformatics*, vol. 8, pp. -, Aug 27 2007.
- [122] J. Yoon, Y. G. Si, R. Nolan, and K. Lee, "Modular decomposition of metabolic reaction networks based on flux analysis and pathway projection," *Bioinformatics*, vol. 23, pp. 2433-2440, Sep 2007.
- [123] R. Xu and I. D. Wunsch, *Clustering*: IEEE Press, Wiley, 2009.
- [124] L. Kaufman and P. Rousseeuw, "Finding groups in data. An introduction to cluster analysis," 1990.
- [125] W. W. Macnaughton-Smith P, Dale M, "Dissimilarity analysis: A new technique of hierarchical sub-division," *Nature*, vol. 202, pp. 1034-1035, 1964.
- [126] W. W. Zachary, "Information-flow model for conflict and fission in small-groups," *Journal of Anthropological Research*, vol. 33, pp. 452-473, 1977.
- [127] I. Xenarios, L. Salwinski, X. Q. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, pp. 303-305, Jan 2002.

- [128] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, "The KEGG databases at GenomeNet," *Nucleic Acids Research*, vol. 30, pp. 42-46, Jan 2002.

