博士論文

The Network Generation Model based on Multilayer Networks

（マルチレイヤーネットワークに基づく
ネットワーク生成モデル）

范　超

The University of Tokyo

The Network Generation Model based on Multilayer Networks

By

Chao FAN

Supervised by

Associate Professor Fujio TORIUMI

A dissertation submitted to the Department of Systems Innovation

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Fall 2017

Abstract

Various networks can be defined in real world, such as social network, traffic network, scientific co-authorship network, etc. Many phenomena are observed on these complex networks and they have a strong influence on society. These phenomena are correlated with network structure and the underlying network generation mechanism. This research aims to clarify the mechanism of network generation and the relationship between phenomena occurring on the network and the network structure. In order to comprehend phenomena existing on networks, there is a need to explore the methodology of modeling real networks.

Since current models focusing on single layer cannot precisely represent high-modularity networks, this research proposed a high-modularity network generation model based on a multilayer network. As people belong to many communities in society, such as family, school, hobby group, and business organizations, each example is regarded as a community in a single layer of a multilayer network. However, measuring each relationship in each community is difficult. A network on SNSs that can be observed combines all communities. That is, a social network is generated from a multilayer network. A synthesized network in the model has either a community structure or a high-modularity structure. This research applied the proposed model to generate various types of networks and compared them with networks in real systems. Not only did it successfully represent real networks but it also predicted how real-world networks are generated from the model's parameters. Moreover, hidden structures were estimated by proposed multilayer model. This research elaborated the relationship between network

structures and social phenomena which has not been clarified so far. Finally, information diffusion experiments are carried out and the average influence degree (AID) is calculated. By enforcing spreading simulation, this research reproduced more realistic information diffusion over multilayer network.

Contents

List of Figures

List of Tables

Chapter 1. Introduction

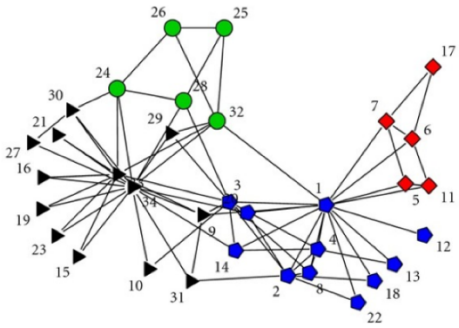
1.1 Background

In the real world and artificial systems, a substantial number of entities interact with each other, where entities and their relationships can be seen as nodes and links respectively. They are organized as complex networks such as social network, scientific co-authorship network, traffic network, financial network, etc. Fig. 1-1 shows some instances of different type of network.

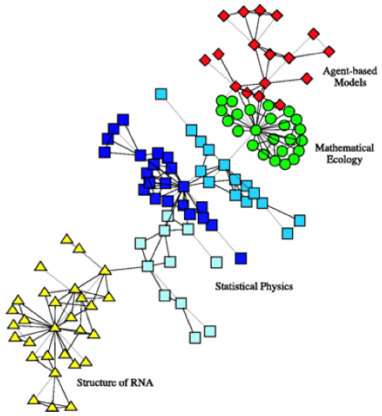
There are many advantages to analyze real system by employing complex network. For one thing, network theory is simple enough to handle complicated problems. Entities and their relationships on a complex system can be simplified as nodes and links. For another, when analyzing a real network, existing network tools are handy to measure the network structure.

1.1.1 Relationships between Network Structure and Phenomena

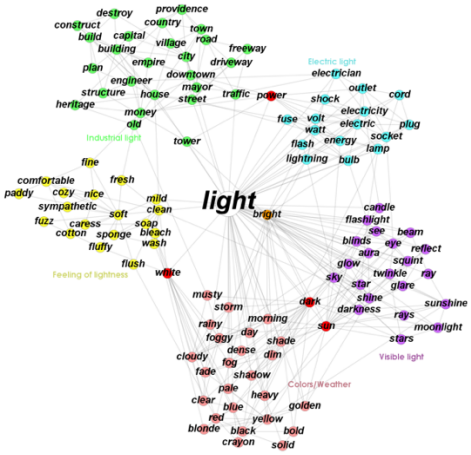
Many phenomena are observed on various types of complex networks in our daily lives, performing a strong influence on our society. For instance, there are information diffusion [1], [2], [3], [31], [32] and opinion formation [98], [99] on social networks. Errors and attacks [4], [5], [6] happen on financial networks or the Internet. Herd behavior [7], [8] occurs on an invention network. These phenomena have been proven to have a high correlation with network structure and the mechanism behind network generation. Therefore, knowing what kind of structures complex networks have and how these structures generate is crucial to the understanding of phenomena.



(1) Zachary's karate club network



(2) Coauthorship network



(3) Words network

(4) NAS Air traffic network

Fig. 1-1 Instances of various networks (Source: (1) and (2) are from Ref. [59] and [42]; (3) is from SNAP[*1]; (4) is from NAS[*2])

In social networks, the popularity of social network services (SNSs) or such social media as Facebook[*3], Twitter[*4], Instagram[*5], and LinkedIn[*6] continues to increase (Fig.

[*1] http://snap.stanford.edu/agm/

[*2] https://www.nsf.gov/

[*3] https://www.facebook.com/

[*4] https://twitter.com/

[*5] https://www.instagram.com/

[*6] https://www.linkedin.com/

1-2). For example, Facebook is the most popular online social network (OSN) worldwide. Fig. 1-3 presents the number of monthly active Facebook users from the third quarter 2008 to third quarter 2017. As of the third quarter of 2017, Facebook had 2.07 billion monthly active users. A great number of people, notably young generation, use it as a source of information. Also, Twitter has proven to be an effective mean of information propagation than traditional mass media during the Great Eastern Japan Earthquake in Japan 2011.

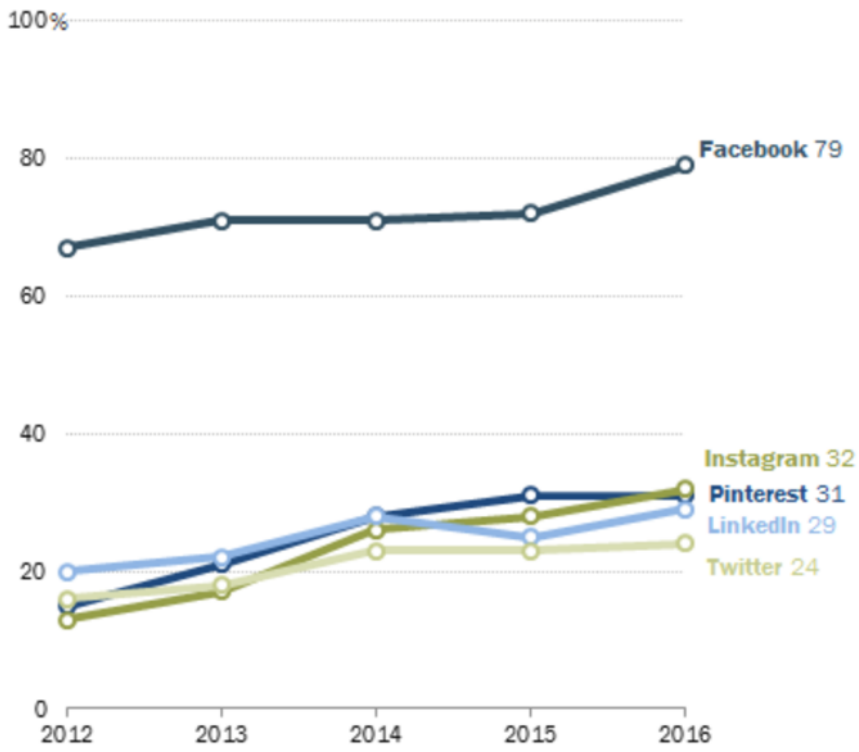


Fig. 1-2 Percentage of online adults who use SNSs

(Source[*1]: Social Media Update 2016)

[*1] http://www.pewinternet.org/2016/11/11/social-media-update-2016/

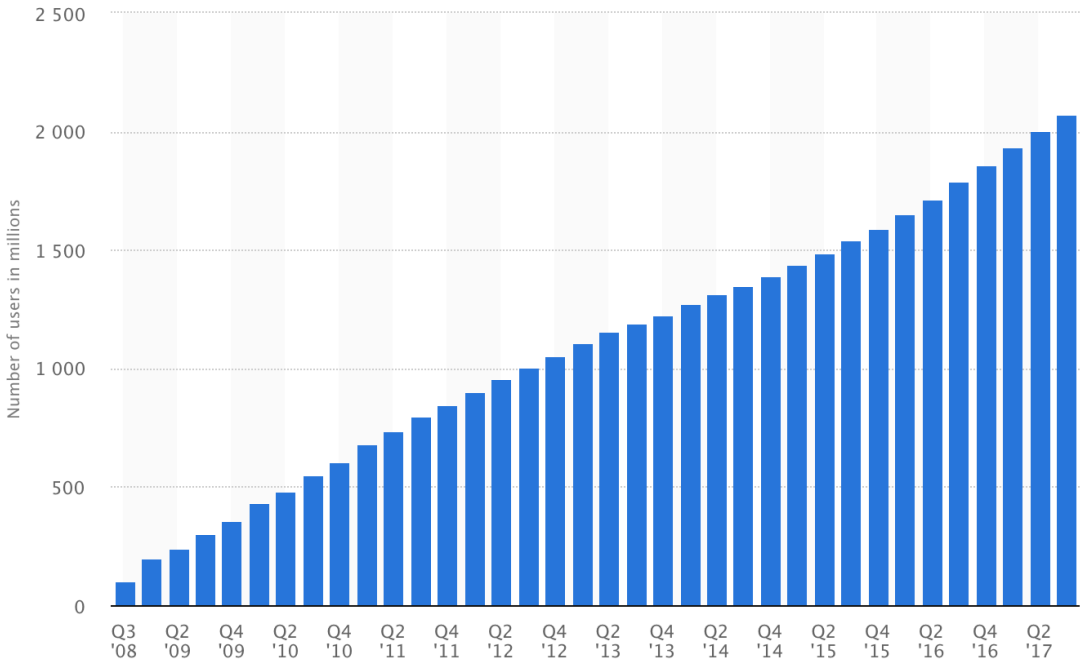


Fig. 1-3 Number of monthly active Facebook users, 3rd quarter 2008 to 3rd quarter 2017 (Source[*1]: Statista 2017)

Such phenomena as information sharing and propagation (Fig. 1-4) happen on online social networks every day and have a strong influence on society. However, these phenomena occurring on the network are strongly affected by the network structure. Some structures are easy to diffuse information, whereas the other is difficult to do so. The spreading results change significantly depending on the structure of network where information diffusion occurs. In catastrophic natural disasters, obtaining detailed information is beneficial to victims and rescuers, thus it is of importance to build network structure that is easy to spread meaningful rescue information and stop rumors. Most work considering this research area focuses on which type of network structure is easy to

[*1] https://www.statista.com/statistics/346167/facebook-global-dau/

diffuse [1], [2], [3].

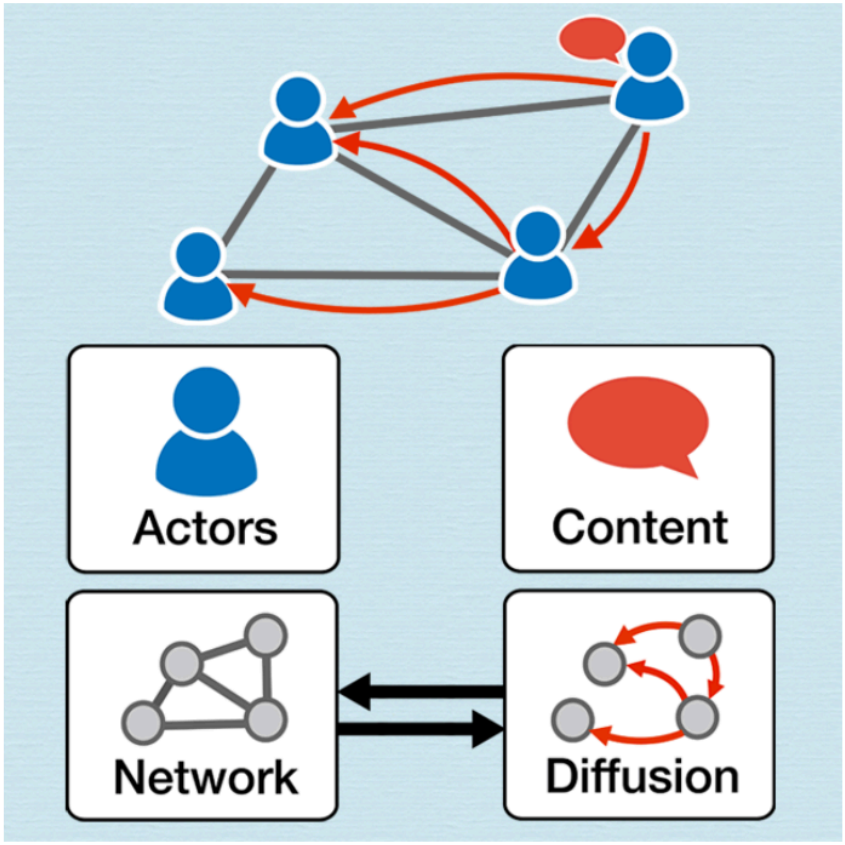


Fig. 1-4 Illustration of information diffusion on online social networks[*1]

Other phenomena like error and attack tolerance (Fig. 1-5) appear to highly correlate with the structure of network. Several researchers have analyzed different types of network structures and find out their relationship to the degree of tolerance against random or systematic failures [4], [5], [6]. When nodes and links undergo errors or intentional attacks, searching for a structure with high robustness and resiliency means a lot to the maintenance of functional complex systems.

[*1] http://cnets.indiana.edu/groups/nan/informationdiffusion/

In financial market, the underlying structure of interaction and relationship among investors can be described by a network. Analyses of herd behaviors and network structures have attracted attention of many researchers recently [9], [10]. Moreover, numerous studies also exist regarding cooperative behavior in the networked agent group [11], [10], [11].

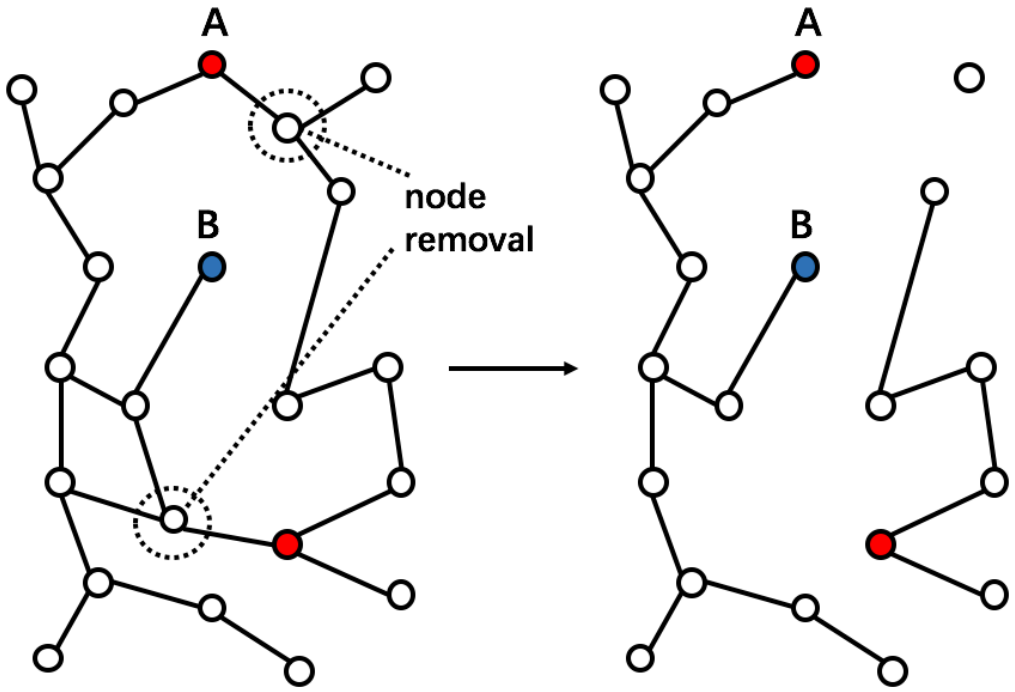


Fig. 1-5 Illustration of the effects of node removal

on an initially connected network [12]

All these social phenomena occurring in real-life situations are deeply influenced by network structure. By looking into how these structure features generate with network generation model, it helps to better understanding of complicated phenomena in complex systems.

1.1.2 Network Generation Model and Multilayer Network

A great deal of work has explored network structures utilizing network generation model. Different generation model is capable of creating a network with a structure of different properties. There are many simple generation models such as Erdős-Rényi (ER) model, Watts-Strogatz (WS) model, Barabàsi-Albert (BA) model, connecting nearest neighbor (CNN) model and so forth.

For instance, Erdős-Rényi model [13], [14], [15] generates random graphs using probabilistic methods. Watts and Strogatz [16] introduced a random graph generation model producing graphs with small world properties. It finally leads to graphs with short average path lengths and high clustering. Barabàsi and Albert [17] proposed another model to generate random scale-free networks. Their model is able to generate networks with power-law degree distributions, which is widely observed in many natural systems and some social networks. Vázquez [18] bring forth a connecting nearest neighbor model to represent network having scale-free property, small world property and a large clustering coefficient. However, these generation models are based on single layer and inadequate to precisely reproduce real-world networks.

Multilayer networks provide a rich representation of real-world interaction. Fig. 1-6 gives a visualization of multilayer dataset from social network called bank-wiring room network in the 1930s. Six different types of social interactions about 14 people was described by multilayer, such as friendship, job trade, helping, negative, etc. Fig. 1-7 shows a DBLP network, which is a co-authorship network with 16 layers in computer science bibliography website. Fig. 1-8 displays air transportation network of Europe. Four distinct airlines are shown in different layers, including Ryanair, Lufthansa, Vueling and

British Airways. The aggregation of all airline is also presented in the multilayer network. The last network illustrated in Fig. 1-9 is an multilayer of online social network. Each layer represents one type of social network in SNSs, e.g. Google, Twitter, and Myspace. It can be found that Twitter layer has a large number of nodes and links, which is contrary to the case of Google and Myspace.

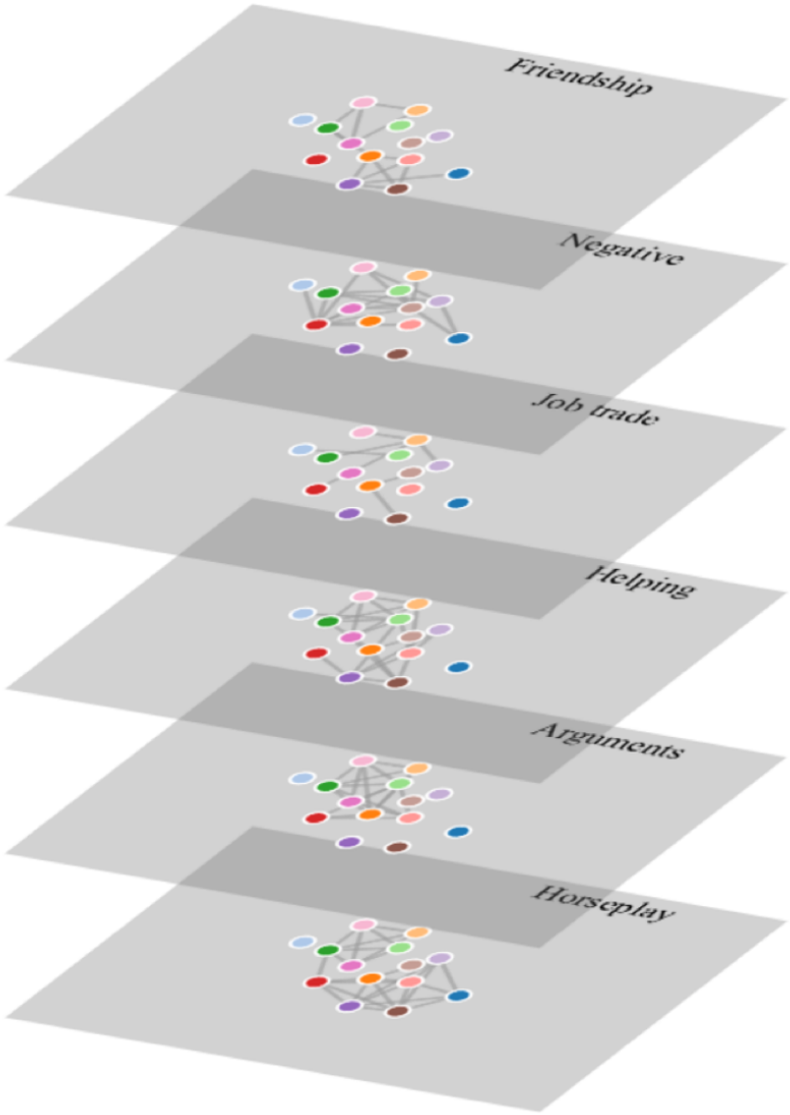


Fig. 1-6 Bank-wiring room network [19]

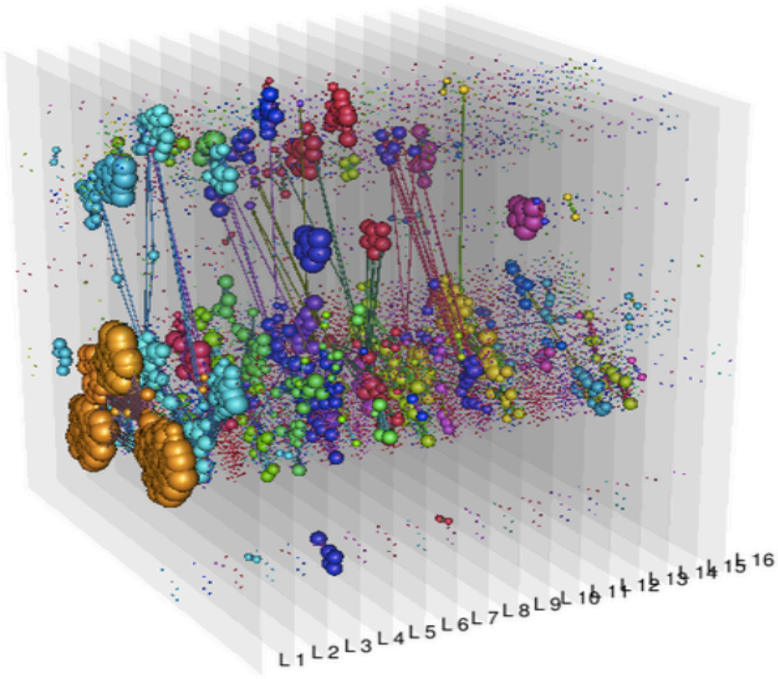


Fig. 1-7 DBLP co-authorship network (Source[*1]: muxviz)

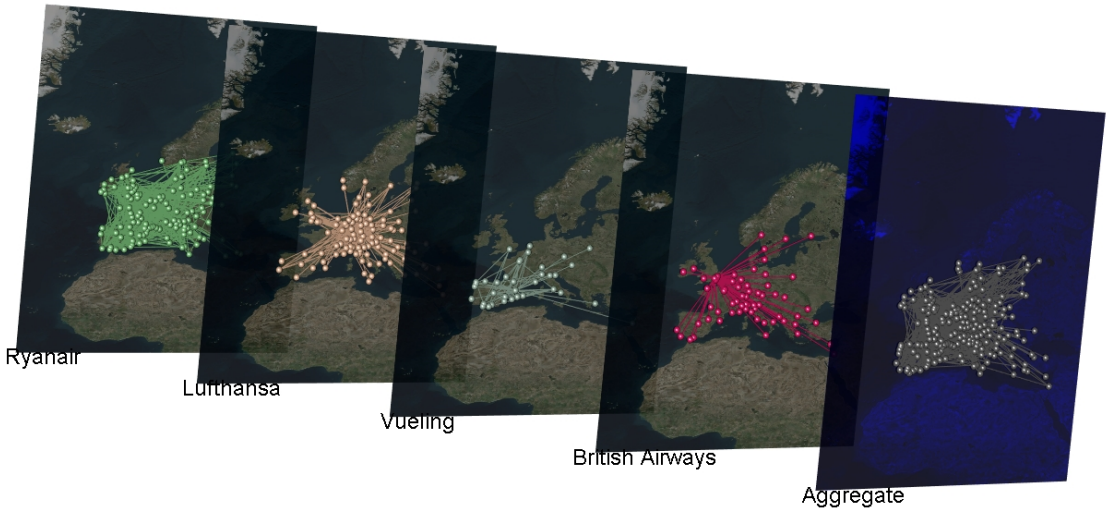


Fig. 1-8 Air transportation network of Europe (Source[*1]: muxviz)

[*1] http://muxviz.net

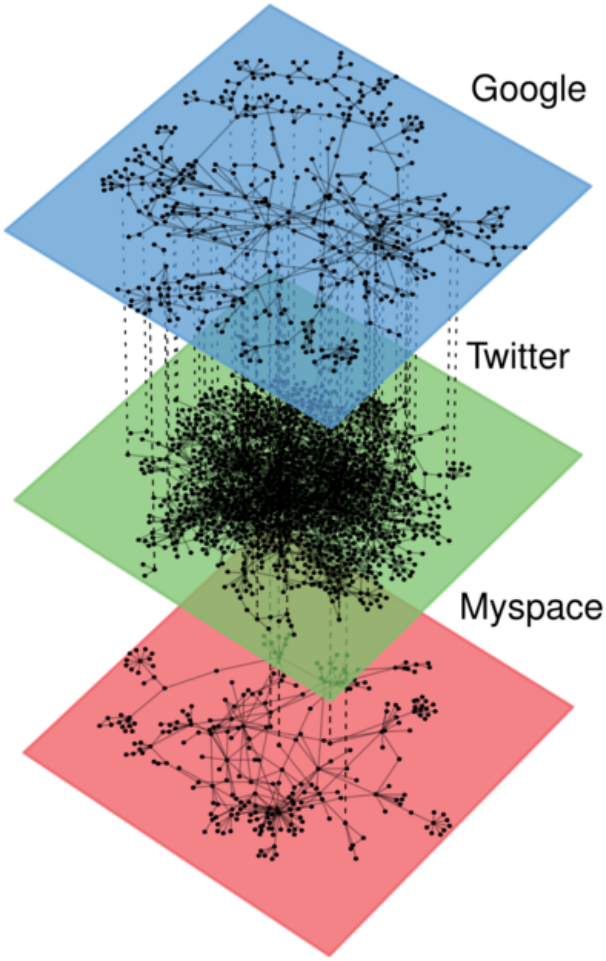


Fig. 1-9 Online social network (Source[*1]: Plexmath)

Some work has investigated network features of multilayer network in real data such as air transportation systems [20] and online games [21]. Work in Ref. [22] models the growth of multilayer network system, and work in Ref. [23] gives statistical mechanics of multiplex networks. Furthermore, there are a considerable amount of research on dynamics, such as epidemic, information spread, cooperation, diffusion process, and so forth [24], [25], [26], [27], [28], [29], [30]. Compared with traditional studies of single

[*1] http://www.plexmath.eu/

layer network, multilayer-based researches can uncover interaction relationships among entities which might not be easily identified in the real system.

1.2 Research Objectives

This research aims to investigate the mechanism of network generation based on multilayer network and tries to clarify the relationship between phenomena occurring on the network and the network structure. In order to comprehend social phenomena existing on complex systems, there is a need to explore the methodology of modeling real networks. Nevertheless, many current models focusing on the single layer cannot synthesize a high-modularity network or have a lower precision in producing networks. Moreover, as real society often has a multilayer structure, it is reasonable to introduce multilayer network in complex network modeling.

Specifically, this research attempts to do the followings:

(1) Propose a high-modularity network generation model based on the multilayer network to represent networks.

(2) Estimate the hidden structure in real network from network generation mechanism.

(3) By utilizing the proposed model to simulate spreading process, more realistic information diffusion can be reproduced.

The proposed multilayer-based model is applied to generate a number of networks and compare them with real-world networks. Not only did it successfully represent real-world data but we also found that we can predict how real-world networks are generated from the model's parameters. Accordingly, it is helpful to understand the essence of phenomena occurring on real networks.

1.3 Outline

This dissertation proceeds as follows. In chapter 2, previous studies are surveyed to give a brief understanding of the research on network generation model. Chapter 3 proposes a high-modularity network generation model based on multilayer network and clarifies how to build a network that resembles network in the real world. A validation is carried out to indicate the reliability and reproducibility of this model. Chapter 4 gives some experiments to estimate hidden structure of network in real data. In chapter 5, an application of multilayer-based model is brought forth by evaluating the accuracy of information diffusion. Finally, Chapter 6 makes a conclusion and provides a couple of prospective researches regarding this field.

Chapter 2. Previous studies

Many studies have analyzed the relationship between the structure of network and various phenomena. Phenomena that researchers are interested in involve information diffusion [1], [2], [3], [31], [32], error and attack tolerance [4], [5], [6], herd behavior [7], [8], cooperative behavior [9], [10], [11], and chain bankruptcy of financial institutions [33], [34], [35]. In these researches, simulations on diverse network structures are implemented in an effort to clarify the influence of network structures on phenomena.

The purpose of this section is to present fundamental knowledge of network generation model in this research. First, some basic generation models and their structure properties are introduced. Furthermore, different types of high-modularity network generation models used in existing experiments are discussed in detail. Also, the disadvantage of current work and necessity of multilayer properties are brought forward at the last part.

2.1 Basic network generation models

Network generation models have been studied by various scholars for a long period. Before discussion goes further, several classical networks and models will be introduced in order to exhibit some properties of network structures.

2.1.1 Classical networks and models

2.1.1.1 Complete network (CPN)

Complete network is a network whose nodes are linked one another. Fig. 2-1 is a

case in point where the complete network has six nodes. If the number of nodes is *N* and the number of links is equal to *M*, there exists an equation to display the relationship between *N* and *M* as follows.

$$M = \frac{N(N-1)}{2} \quad \textbf{(2.1)}$$

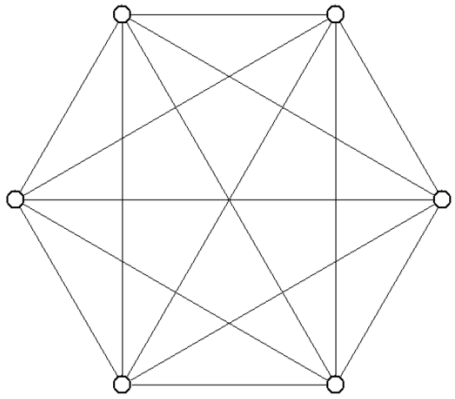


Fig. 2-1 A complete network with six nodes

2.1.1.2 Lattice network

A lattice network positions all nodes on square lattice. Each node is adjacent to its upper, lower, left and right nodes, which means a node is only linked to its four neighbor. A typical lattice network is shown in Fig. 2-2.

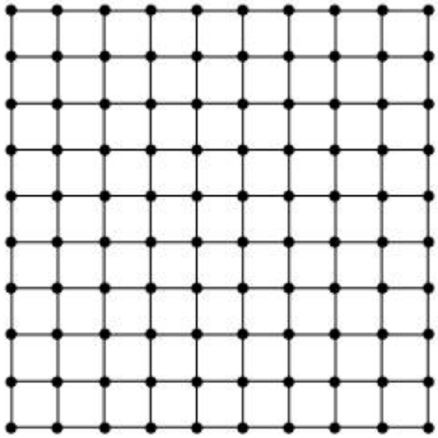


Fig. 2-2 A lattice network

2.1.1.3 Regular network

Regular networks are networks where each node has exactly the same number of neighbors. A regular network is highly ordered and every node has the same degree. For example, nodes are placed in a circle where every adjacent pair of nodes has two of neighbors in common in Fig. 2-3.

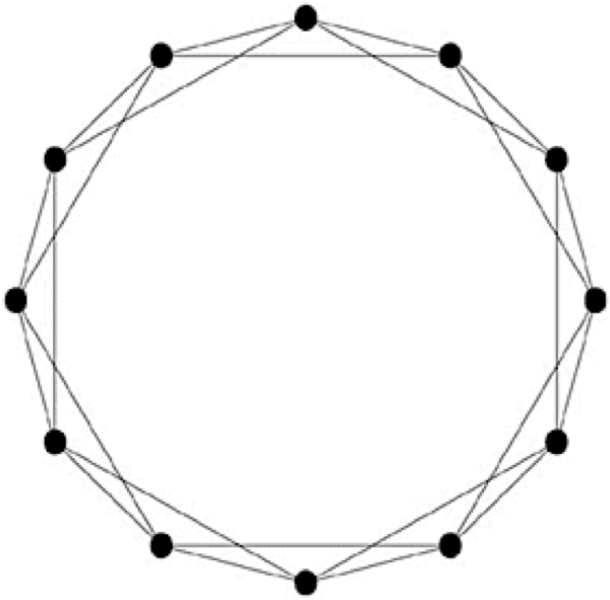


Fig. 2-3 A regular network

2.1.1.4 Random network and Erdős-Rényi model

Models for generating random graphs is prevalent in earlier research. Erdős-Rényi model (ER model) [13], [14], [15] is a famous model that generates random network utilizing probabilistic methods. In this model, a network is constructed by means of connecting nodes at random. Each link is present or absent with a fixed probability p, which is independent from the other links. For a network with N nodes and M links, the probability p can be expressed as:

$$p = \frac{M}{\binom{N}{2}} \tag{2.2}$$

As can be seen from Fig. 2-4, more links will be born and network tends to become denser as parameter p increase from 0 to 1. When producing a random network employing ER model, the shape of generated network changes every time.

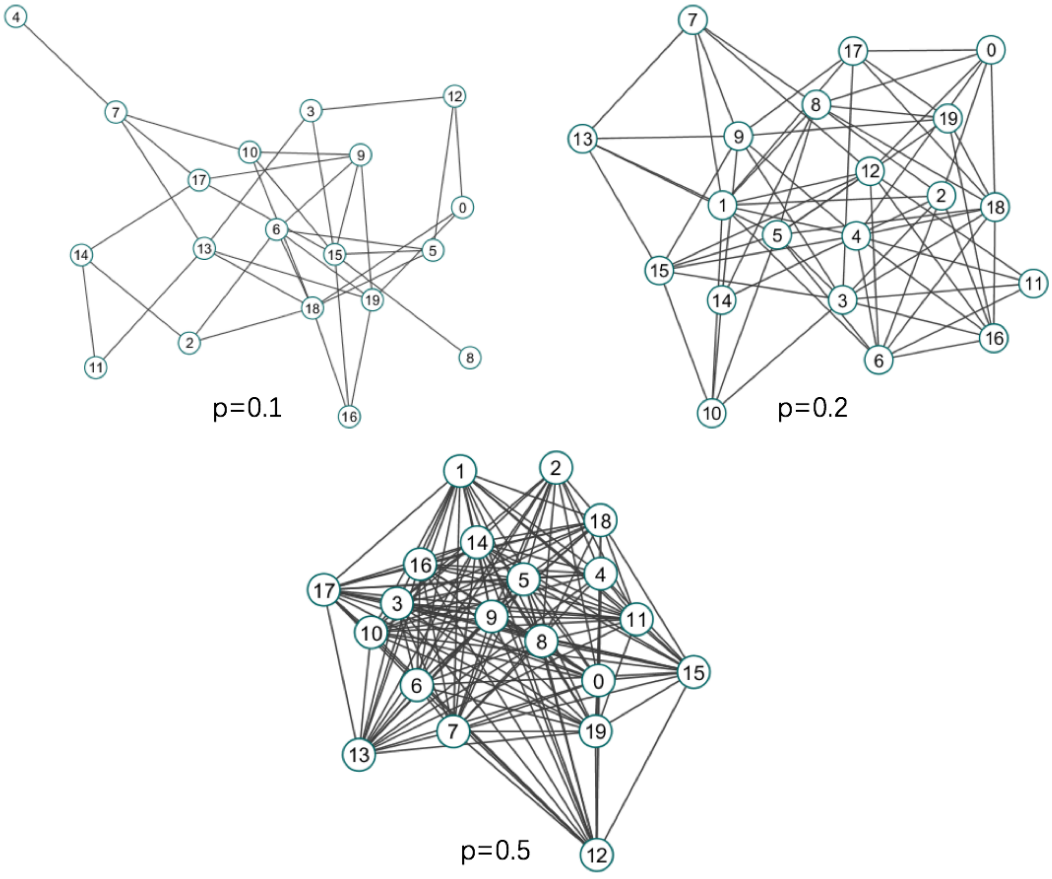


Fig. 2-4 Random graphs generated by ER model

2.1.2 Small-world network and Watts-Strogatz model

Stanley Milgram and other researchers devised a small-world experiment [36] around the 1960s to make a statistics of the relationship among people and investigate average path length for social networks. This classical experiment gives birth to a famous

conclusion called “six degrees of separation”. Although our world is big enough, the distance between people is really small when social interaction is taken into consideration.

In addition to short distance between any two persons, social network of people is usually highly clustered according to our common sense. As is depicted in Fig. 2-5, different types of clusters or communities exist in all aspects of social life, because friends of our friends will have a tendency to become friends.

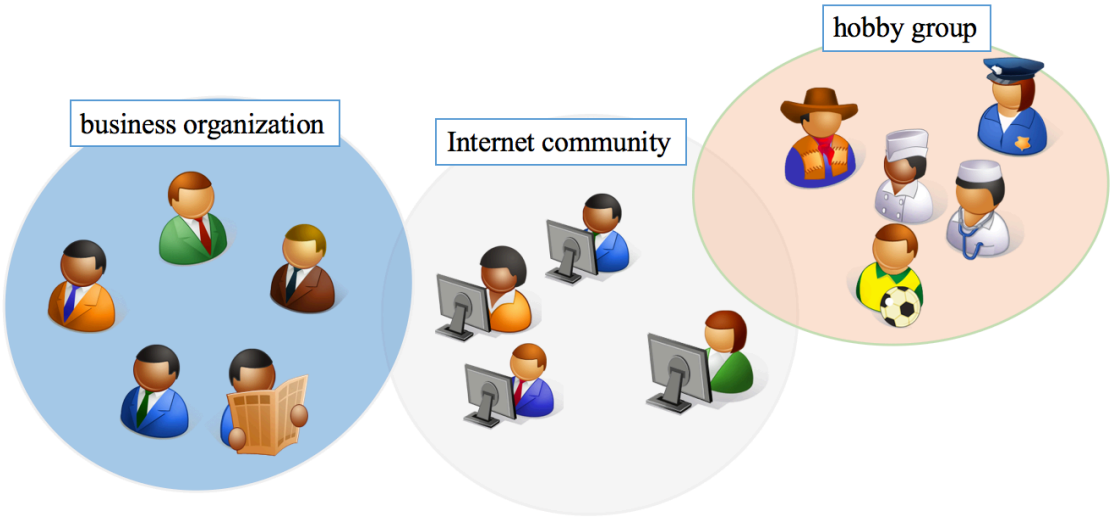


Fig. 2-5 Clustered social network of people

The small-world phenomenon is a significant discovery on the structure of complex networks. According to Duncan J. Watts [16], many real-world systems can be highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs. In other words, these sorts of networks retain properties of short average path lengths and high clustering coefficients.

So how can we construct a small-world network with these two properties using an artificial generation model? Watts and Strogatz [16] developed a Watts-Strogatz model (WS model). As shown in Fig. 2-6, this model starts with a regular ring lattice on the left side. Random rewiring of links is proceeded with a probability p. By increasing these

rewired links that are called “short cuts”, the path length between arbitrary two nodes becomes shorter. When p is equal to 1, network degenerates into a random network.

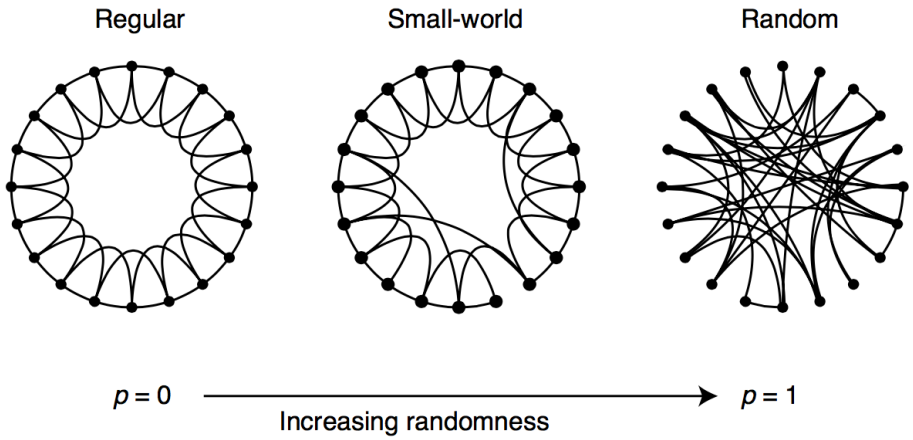


Fig. 2-6 Network generated by WS model [16]

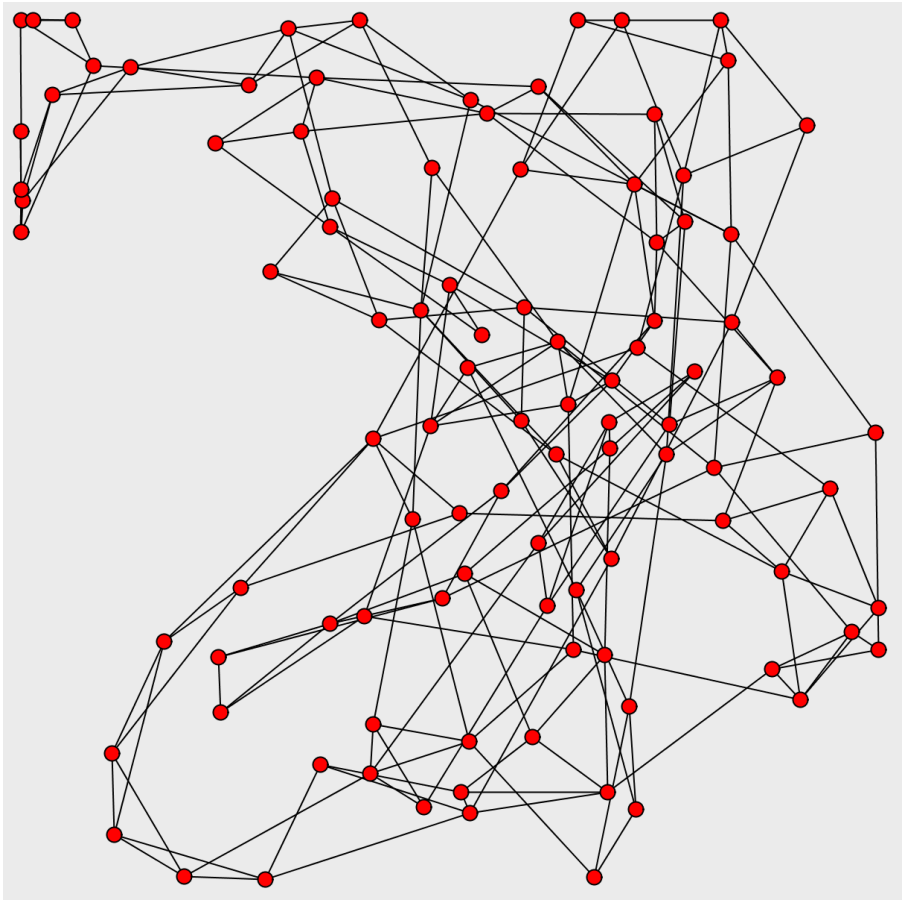


Fig. 2-7 A network generated by WS model

Fig. 2-7 shows a small-world network generated by the WS model. According to this figure, WS model produces a highly clustered structure.

2.1.3 Scale-free network and Barabási-Albert model

It is universally acknowledged that the rich get richer and the poor get poorer. Also, the top 20% of people own 80% of the world's wealth. The 80/20 rule, a.k.a. Pareto principle, is approximately followed by a power law distribution in mathematics. It gives an explanation of many natural phenomena that roughly 80% of the effects stem from 20% of the causes. Barabási called this property "scale-free" [17], which is widely observed in many natural systems and some social networks (see Fig. 2-8).

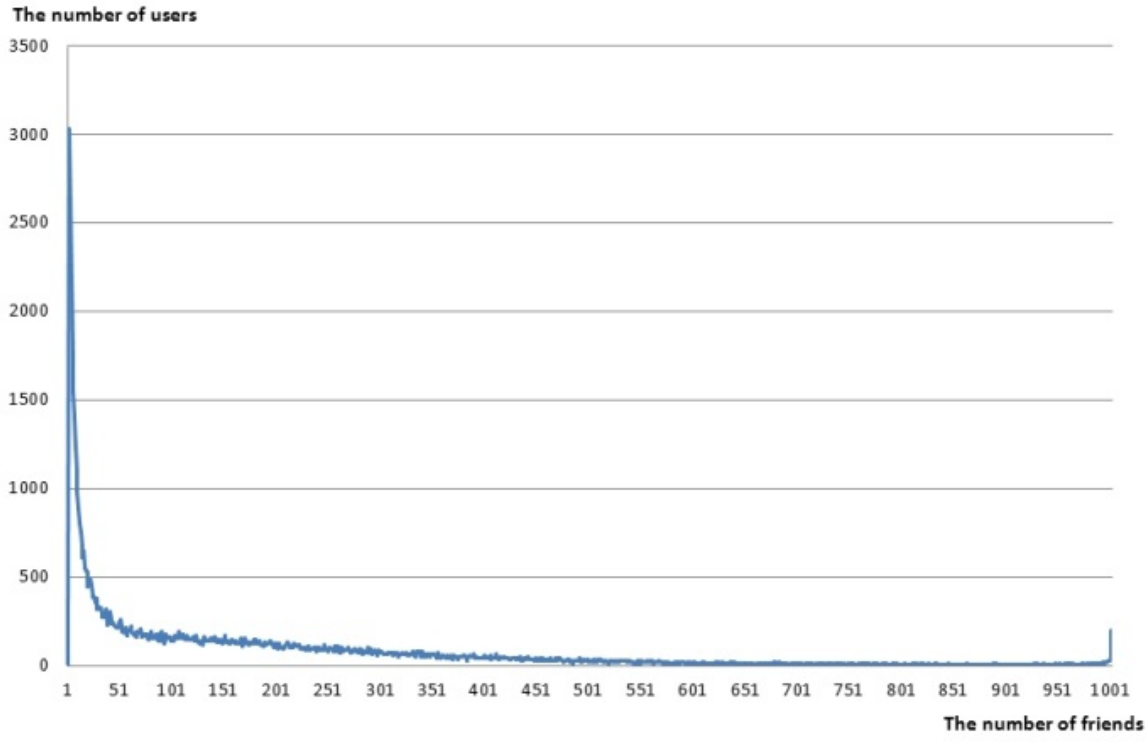


Fig. 2-8 Power law distribution in Renren network [47]

The scale-free property is another great discovery about the structure of a complex network. The work of Albert-László Barabási and colleagues in 1999 triggers a boom of scale-free network research in the following years. A scale-free network is defined as a network whose degree distribution follows a power law. This distribution *p(k)* represents the existence probability of nodes having degree *k*. The relationship between *p(k)* and *k* can be approximated by a power function as follows:

$$p(k) \sim k^{-\gamma} \tag{2.3}$$

where γ is a parameter whose value is typically in the range between 2 and 3.

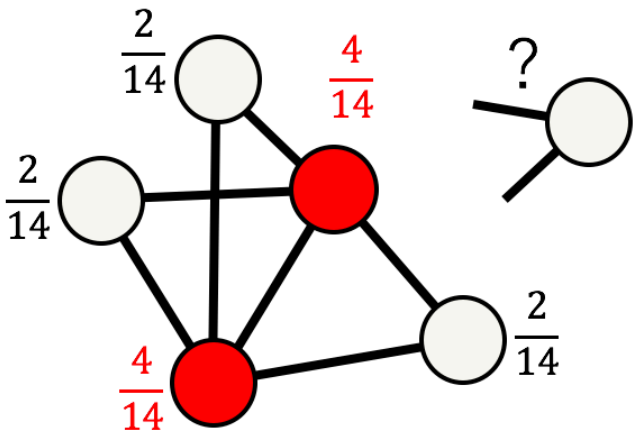


Fig. 2-9 Preferential attachment of BA model

Further, in order to simulate the generation of scale-free network, Barabási et al. proposed a Barabási-Albert model (BA model) utilizing a mechanism of growth and preferential attachment. When linking a new node, high degree nodes are easy to be selected. The basic idea of preferential attachment can be illustrated by Fig. 2-9. The red nodes have a high degree than others, thus they will be selected with a high probability.

The detailed algorithm of BA model can be elaborated in the following steps.

(1) Create a complete network where the degree of each node is m;

(2) Add a new node;

(3) Attach the new node to m nodes which is selected by preferential selection;

(4) Repeat (2) and (3).

The probability that node i is selected can be calculated as follows.

$$p_i = \frac{k_i}{\sum_j k_j} \tag{2.4}$$

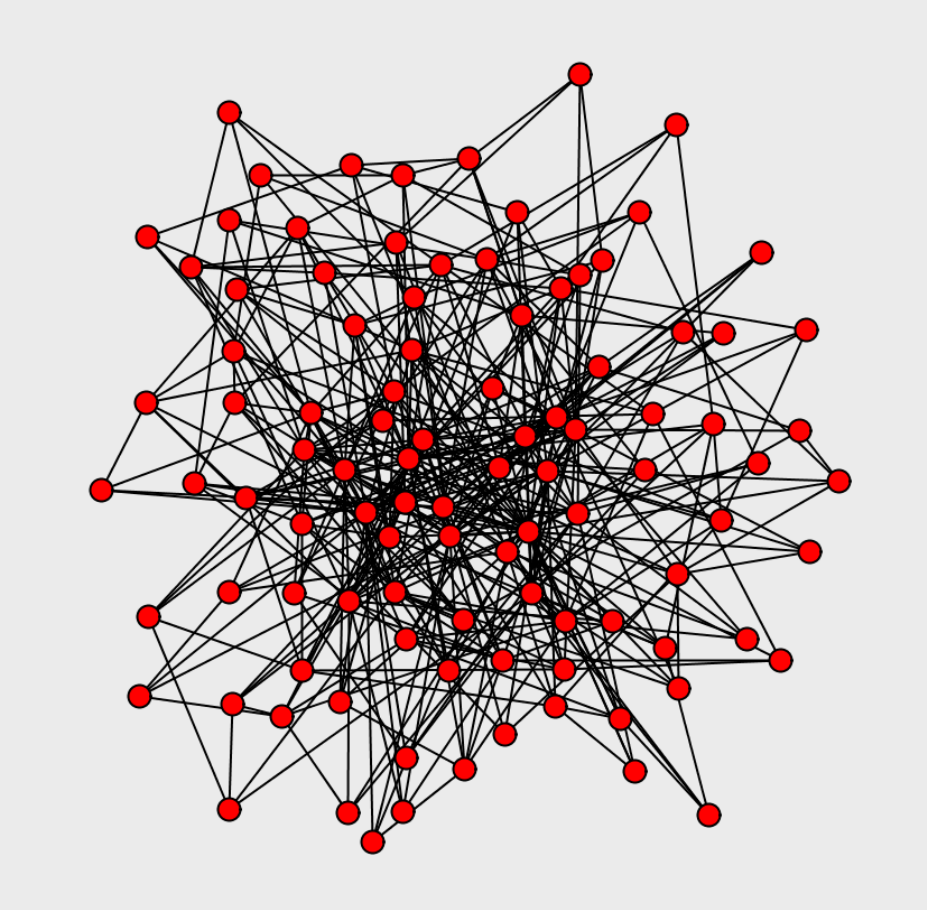


Fig. 2-10 A network generated by BA model

A network constructed by BA model is presented in Fig. 2-10. Some hub nodes with

many links are present in the generated network.

2.1.4 Connecting Nearest Neighbor (CNN) model

The small-world network generated by WS model does not exhibit a power law degree distribution. On the other hand, a standard BA model cannot produce a small-world network. Can we find out an algorithm that builds a network with both small-world and scale-free property?

Vázquez bring forth a connecting nearest neighbor model (CNN model) [18] to represent such network. Based on an idea that friends of one's friends will tend to become one's friends, this model repeats two steps to create a network.

(1) Add a new node to the preferentially selected node i with probability p and create potential links between the new node and all the neighbor of i;

(2) With probability $1-p$, convert one potential link selected at random into a link.

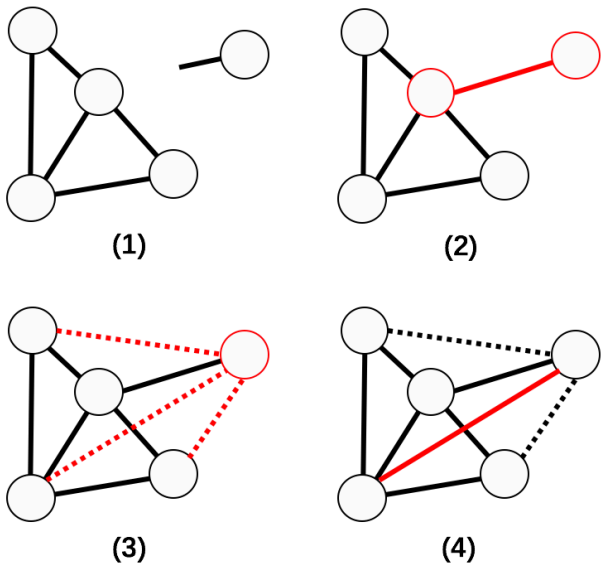


Fig. 2-11 Procedure of network generation by CNN model

Fig. 2-11 depicts the procedure of network generation in detail, whereas Fig. 2-12

gives a visualization of network produced by CNN model. This algorithm can yeild a network with a power law degree distribution, a short average path length and a large clustering coefficient.

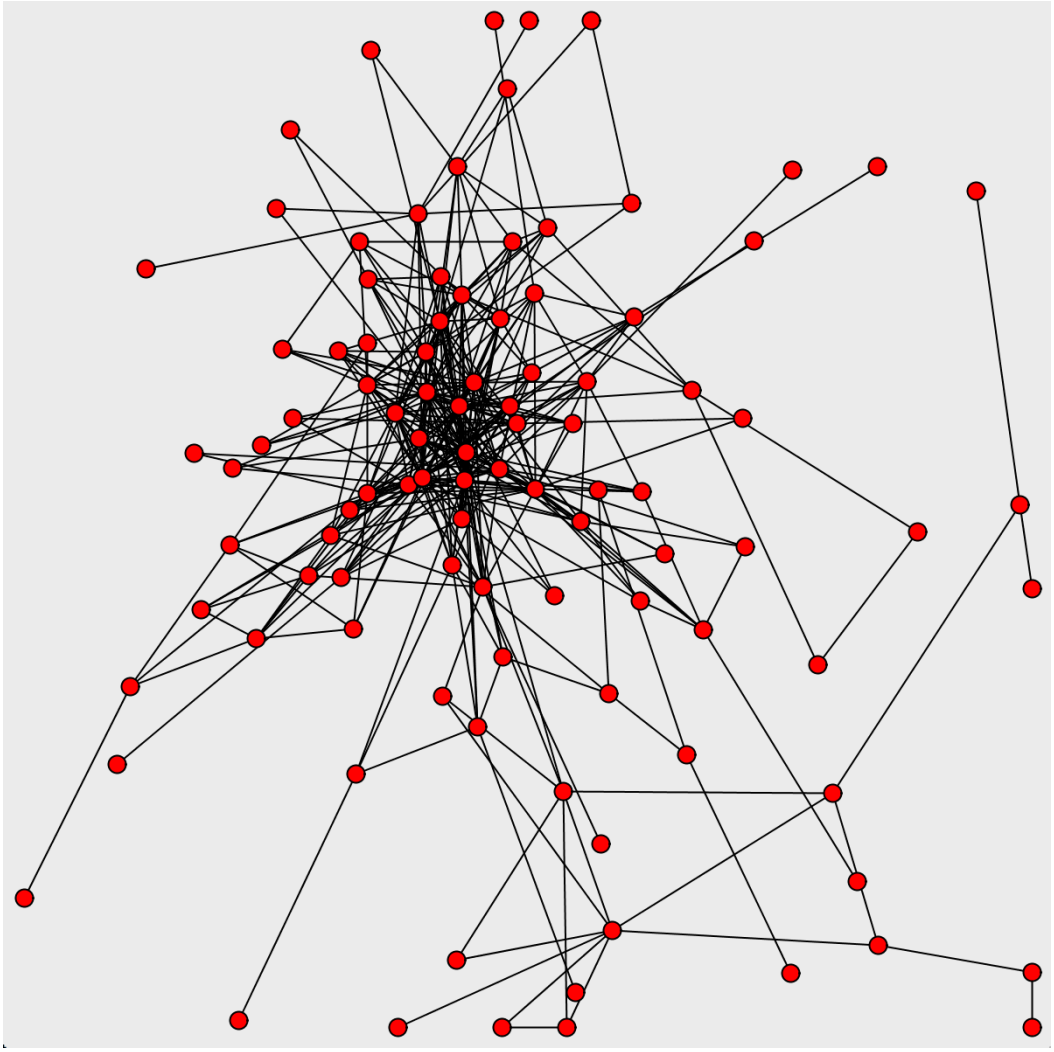


Fig. 2-12 A network generated by CNN model

2.1.5 Other generation models

Many generation models given above will be utilized in this research, e.g. complete network (CPN), ER model, WS model, BA model and CNN model， which will be discussed in detail later. In addition to the discovery of classical models, a great number

of new models [37], [38], [41], [40], [41] are proposed to discribe the generation or growth of network. Some variants extend the standard model and include such important properties as small-world effect and scale-free distribution of degree.

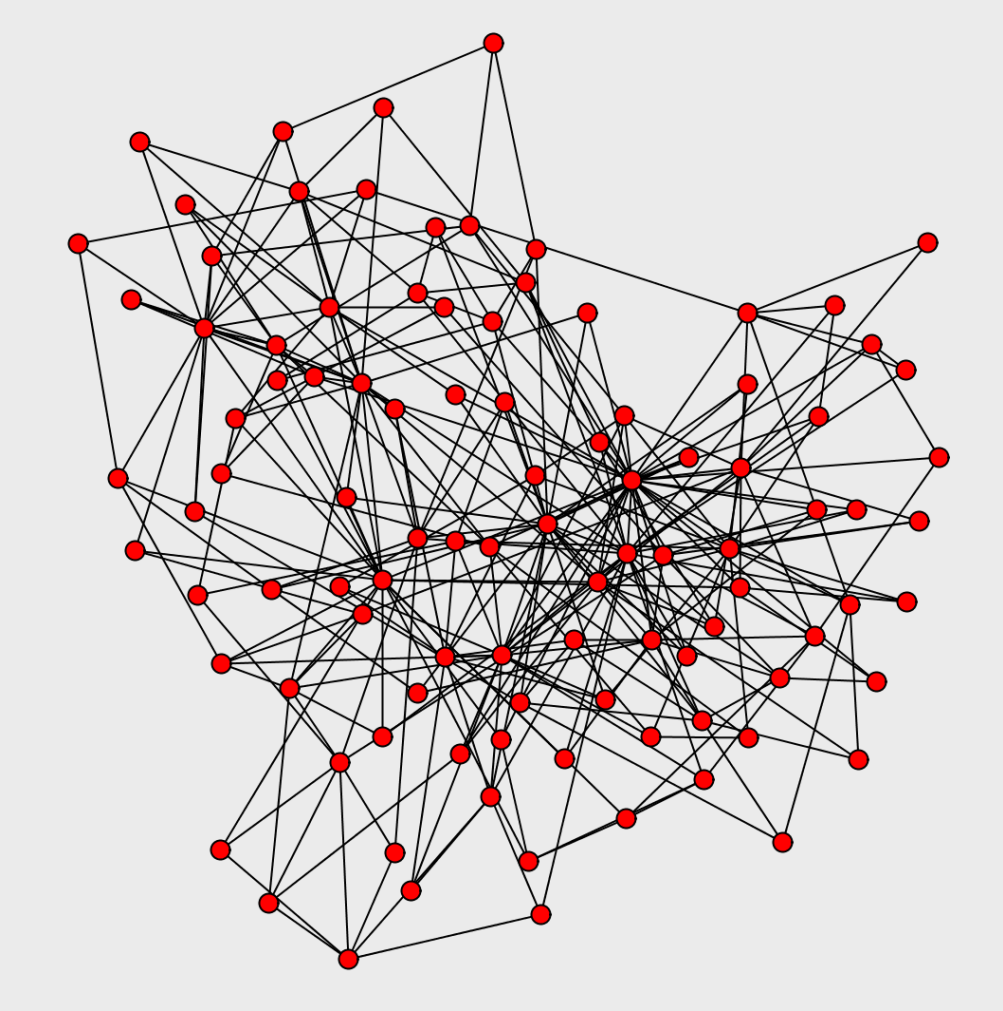


Fig. 2-13 A network generated by HK model

For instance, Holme-Kim model (HK model) [37] added an additional "triad formation" step to BA model in order to incorporate the high clustering. If a new node is connected to node *i* in the previous preferential attachment step, then one more link from the new node to a randomly chosen neighbor of node *i* will be added. The algorithm executes preferential attachment at first, and then performs triad formation with

probability p. Similar to CNN model, HK model has an ability to obtain networks with high clustering coefficient. Besides, the clustering coefficient can be adjusted by adjusting the value of p. Fig. 2-13 gives an exhibition of network generated by HK model. From this graph, it can be noticed that many nodes are connected with a triangle forming some local clusters, which feature can hardly be found in the network produced by BA model.

Klemm and Eguiluz [38] device a dynamical model that extends scale-free network with small world behavior. Newman et al. [39] employs a bipartite structure to generate affiliation networks with both small-world and scale-free properties. Wang's model [40] is also a hybrid of BA model and WS model to mimic real-world systems. Another network generation model, introduced by Catanzaro et al. [41], present a growth model producing a degree-assortative network for the purpose of describing the behavior of social networks.

2.2 High-modularity network generation models

The terms, small-world and scale-free are commonly accepted as significant statistical properties of network strucutre. Besides the two properties, community is another structural features of importance to comprehend social phenomena on society. Community structure reveals the internal organization of nodes, therefore, it is salient in determining the characteristics of people in same group or circle. From the 2000s, it gains a large quantity of attentions from many researchers [42], [43], [43], [45], [46], [47]. Before going into further, the relationship between modularity and community structure will be illuminated at first.

2.2.1 Modularity

Modularity, proposed by Newman [44], [45], is one measure of the community structure. It measures the strength of a division of a network into communities.

An undirected graph is defined as $G = (V, E)$, where V and E represent nodes and links respectively. Then the number of nodes and links is denoted as N and M ($|V| = N$ and $|E| = M$). A_{ij} is an element of the adjacent matrix of G.

$$A_{ij} = \begin{cases} 1 & (i,j) \in E \\ 0 & otherwise \end{cases} \quad \textbf{(2.5)}$$

The degree of a node i can be defined as $k_i = \sum_j A_{ij}$. Thus, the modularity Q is defined as follows.

$$Q = \frac{1}{2M}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2M}\right]\phi(C_i, C_j)$$

$$= \frac{1}{2m}\sum_{ij}\left[A_{ij} - P_{ij}\right]\phi(C_i, C_j) \quad \textbf{(2.6)}$$

where C_i represents node i that belongs to community C_i. $\phi(C_i, C_j)$ is 1 if i and j are in the same community and 0 otherwise. P_{ij} is the probable node degrees between node i and j at random. In this studies, the value of Q is acquired by performing community detection using Newman's method.

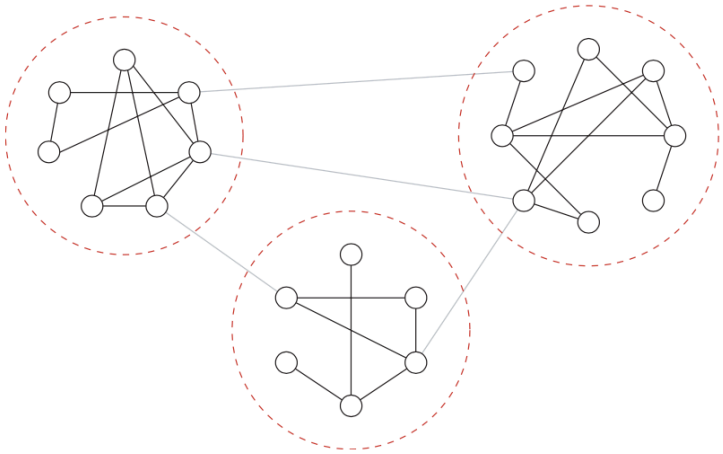


Fig. 2-14 A network with high modularity

Modularity is an important characteristic of real-world networks and reflects whether a network exhibits the property of a community. High-modularity implies dense intra-community links and fewer inter-community links and it measures a good division of a network into communities, or a better community structure. As is clearly shown in the Fig. 2-14, the links within each community are more than the ones between communities, thus this network has a high-modularity property or an apparent community strucutre.

The following sections will introduce some baseline models producing network with community structure, including LFR benchmeark, Kronecker graph and Pasta's model. The multilayer-based model proposed in this research will make a comparision to them from the perspective of network features.

2.2.2 Lancichinetti-Fortunato-Radicchi benchmark

The Lancichinetti-Fortunato-Radicchi (LFR) benchmark [48] is derived from the issue of testing a community detection algorithm. Before the introduction of LFR

benchmark, standard tests involve constructing simple artificial graphs with a built-in community structure. However, these graphs fail to reflect some real features of network in reality.

The most famous benchmark before LFR benchmark is an algorithm designed by Girvan and Newman [42]. The procedure of Girvan-Newman (GN) benchmark begins as this:

(1) Each network was constructed with 128 nodes, divided into four groups with 32 nodes each.

(2) Links were connected between node pairs independently at random, with a probability P_{in} within communities and P_{out} between communities where $P_{in} < P_{out}$. The average degree of the network is keeped as 16 by adjusting probability.

However, some obvious drawback concerning this benchmark can be discerned. On one hand, degree of all nodes is approximately the same. On the other hand, all communities of generated network have the same size. Both two features suggest that GN benchmark cannot be considered as a effective model to generate a real network with community structure.

The characteristics of real networks lie in the heterogeneity of both degree and community size. According to previous studies [43], [49], [50], [51], the distribution of community sizes follows a power law distribution in real systems. Further, the distribution of node degree of real networks can also be approximated by a powe law [4], [52], [53]. Hence, both the degree and community size distribution are assumed as power laws in LFR benchmark with exponents τ_1 and τ_2, respectively. The implement of algorithm for LFR benchmark are presented as follows.

(1) The degree of each node is selected from a power law distribution with exponent

τ_1. The maximal and minimal degrees k_{min} and k_{max} are chosen to ensure the average degree k. Linking of nodes is carried out by configuration model.

(2) A mixing parameter μ is utilized to control the fractions of links among communities, which is the ratio of the number of external neighbors of a node by a node's total degree.

(3) The community sizes are taken from a power law distribution with exponent τ_2. The two extremes of community size s_{min} and s_{max} are set to meet some constrains: $s_{min} > k_{min}$ and $s_{max} > k_{max}$.

(4) All nodes are homeless at first. A node enters a randomly selected community if there is enough capacity, otherwise it remains homeless. In latter iterations, a homeless node is assigned to a community at random. If the community is complete, a randomly chosen node will be kicked out and become homeless. Iteration stops when there is no homeless node.

(5) Rewiring steps are executed to ensure the mixing parameter μ.

LFR benchmark can synthesize networks with planted community structures. A network constructed by LFR benchmark with 500 nodes is shown in Fig. 2-15. LFR benchmark generalizes GN benchmark by introducing features of real network. GN benchmark is actually an instance of LFR benchmark with community sizes equal to 32 nodes, and the degree of each node equal to 16. Later on Bródka proposed an extension of LFR benchmark called mLFR benchmark [54] which is able to generate multi-layered social networks. The algorithm organizes link distribution following power law distribution, changes degrees of nodes through the layers, and change membership of node on layers.

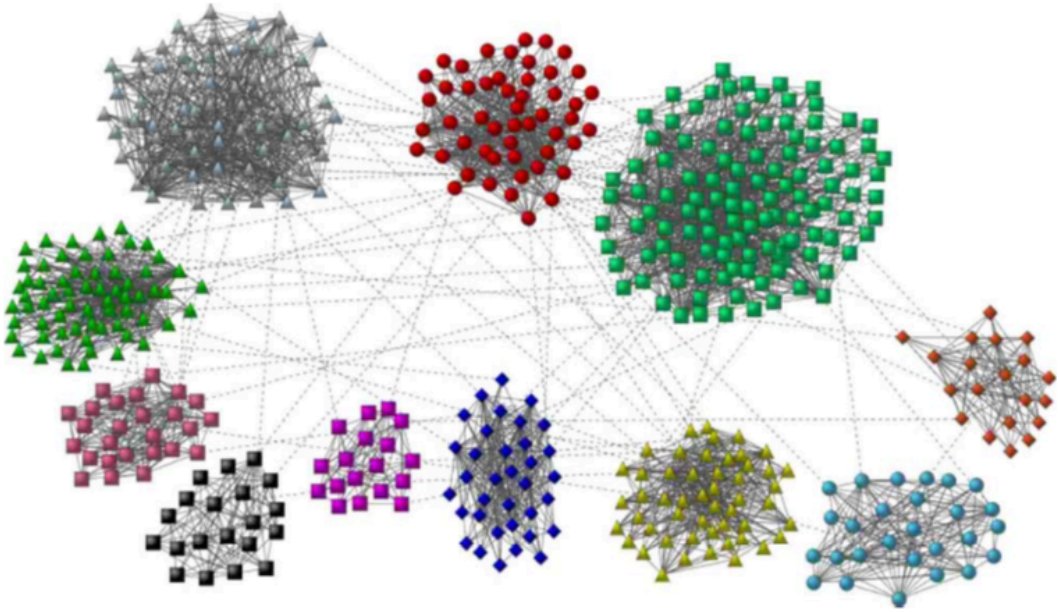


Fig. 2-15 A network generated by LFR benchmark [48]

2.2.3 Kronecker graph model

Kronecker graph [57] is another approach to modeling real networks with communities. The foundation of Kronecker graph model is hierarchical network. It is a common intuition that real networks are often hierarchically organized into communities. Then these communities grow recursively, creating similar copies of themselves. For hierarchical network model, its algorithm synthesizes networks by recursive way. It also enables the reproduction of some properties in real networks, such as scale-free topology and high clustering of the nodes. However, it differs from WS model in the distribution of nodes' clustering coefficients, because high degree nodes tend to have a small clustering coefficient rather than a constant value.

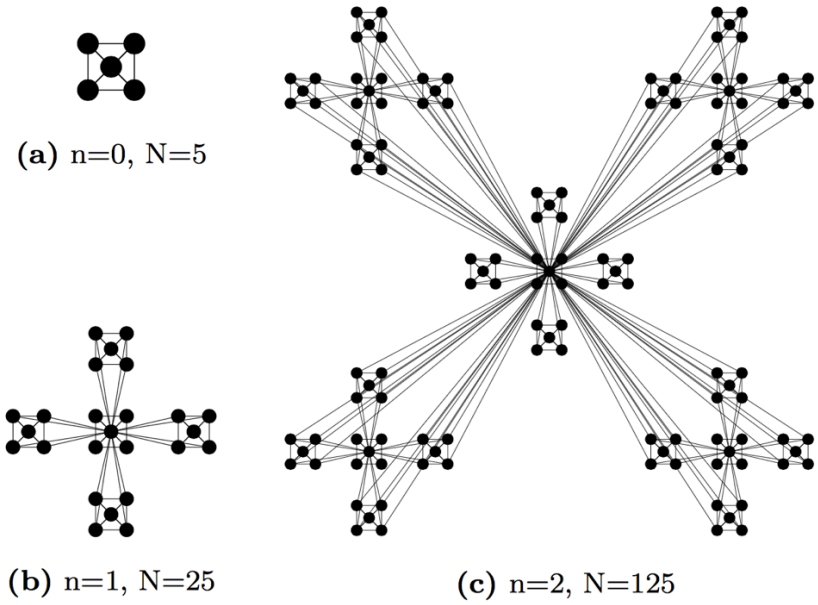


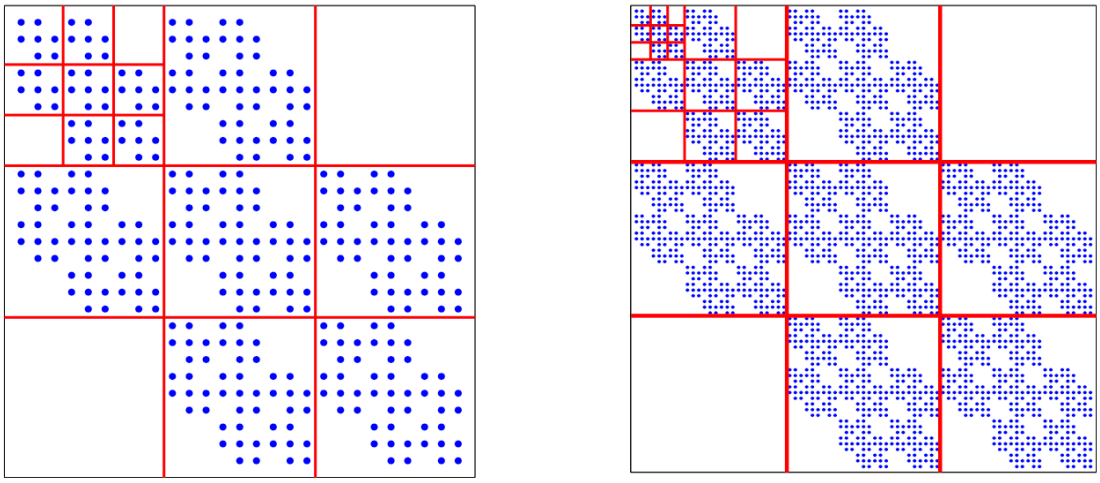
Fig. 2-16 The iterative construction of Ravasz's model [56]

Ravasz et al. [55], [56] identified some inherent natures of real networks including scale-free, hierarchical, and modular structure. They proposed a simple hierarchical network model, which starts with several small clusters of five densely linked nodes. These clusters combine to form larger and self-similar clusters. Conducting this self-similar nesting process eventually results in a strict, fine structure with hierarchy and community as well as scale-free topology. The figure in Fig. 2-16 gives the iterative construction process of this model.

Leskovec and his coauthers raised a more general Kronecker graph model [57] than Ravasz's model to produce a hierarchical structure in a real network. Their algorithm begins with initiator graph K_1 with N_1 vertices and E_1 edges. The algorithm recursively generates successively larger self-similar graphs K_2, K_3, K_4, K_5, ... such that *k-th* graph K_k has $N_k = {N_1}^k$ vertices. To do this, they construct a sequence of graphs from initiator graph K_1 by iterating the Kronecker product.

Given two matrices A and B with sizes $m\times n$ and $p\times q$ respectively, then the Kronecker product $A\otimes B$ is the $mp\times nq$ block matrix.

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \tag{2.7}$$



(a) K_3 adjacency matrix (27×27) (b) K_4 adjacency matrix (81×81)

Fig. 2-17 A case of adjacent matrix of Kronecker graph [57]

In order to obtain K_k from K_{k-1}, it needs to expand each vertex in K_{k-1} in by converting it into a copy of K_1 and joing the copies according to the adjacencies in K_{k-1}. An adjacent matrix of K_4 is given in Fig. 2-17.

Through the steps above, the Kronecker graph model can model the self-similar structure and produce hierarchical networks that exhibit properties of hierarchy, scale-free and community structure in real networks.

2.2.4 Pasta's model

Model proposed by Pasta et al. [58] is a tunable and growing network generation

model. Not only does it form small-world and scale-free properties, but it constructs a network with hierarchical community structure. Specifically, node degree distribution within a community follows a power law. Moreover, clustering coefficients within each of the communities is fairly high. Finally, hierarchy of internal structure of communities is crucial topological characteristics for creating synthetic networks that resemble networks in real systems.

Five parameters of this model are listed below. They are required as input at the beginning of algorithm.

(1) The number of nodes N;

(2) The number of links for each newly added node m;

(3) The minimum number of communities c;

(4) The probability of triad formation P_t;

(5) The probability of having inter-cluster links P_c.

The procedure of constructing a network by Pasta's method can be elaborated through the following steps. The illustration in Fig. 2-18 reveals some crucial steps in the creation of network.

(1) c triads denoting c communities are added to network. A link is randomly chosen between every two different communities.

(2) A new node n_1 is added and attached to an existing node n_2 which is selected based on preferential attachment.

(3) With probability P_t, n_1 is connected to $m-1$ other nodes preferentially chosen from the community of n_2, resulting in a triad structure. $m-1$ nodes can be either neighbors

or non-neighbors of n_2.

(4) A new node n_3 is added and attached to an existing node n_4 which is again selected based on preferential attachment. In this step, a limitation is imposed that n_3 does not belong to the community of node added in the previous step.

(5) With probability P_t, n_3 is connected to $m-1$ other nodes which are preferentially chosen from the community of n_4, forming a triad structure. $m-1$ nodes can be either neighbors or non-neighbors of n_4.

(6) The communities of n_2 and n_4 in the previous steps are connected by a link with probability P_c. Two endpoints of the added link are preferentially chosen in both communities.

(7) Repetition is performed from step (4) till the network evolves into a size of n.

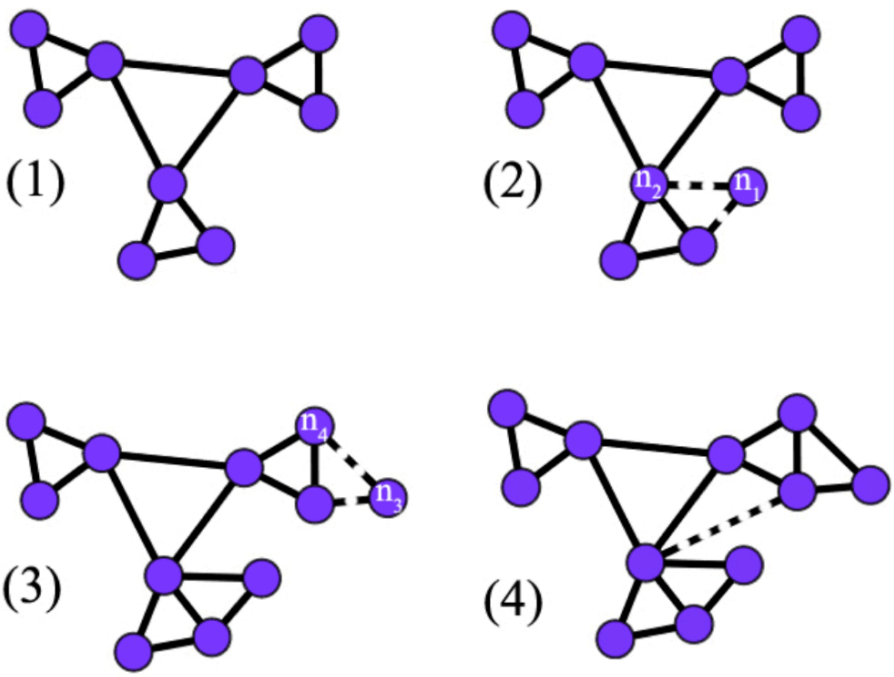


Fig. 2-18 The constructing process of Pasta's model [58]

By enforcing the algorithm above, Pasta's model is able to synthesize networks with hierarchical community structure as well as small-world and scale-free properties. This model is tunable and flexible by modifying different values of parameter. A generated network of Pasta's model is presented in Fig. 2-19. Different colors show a variety of communities, where a hierarchical structure is present.

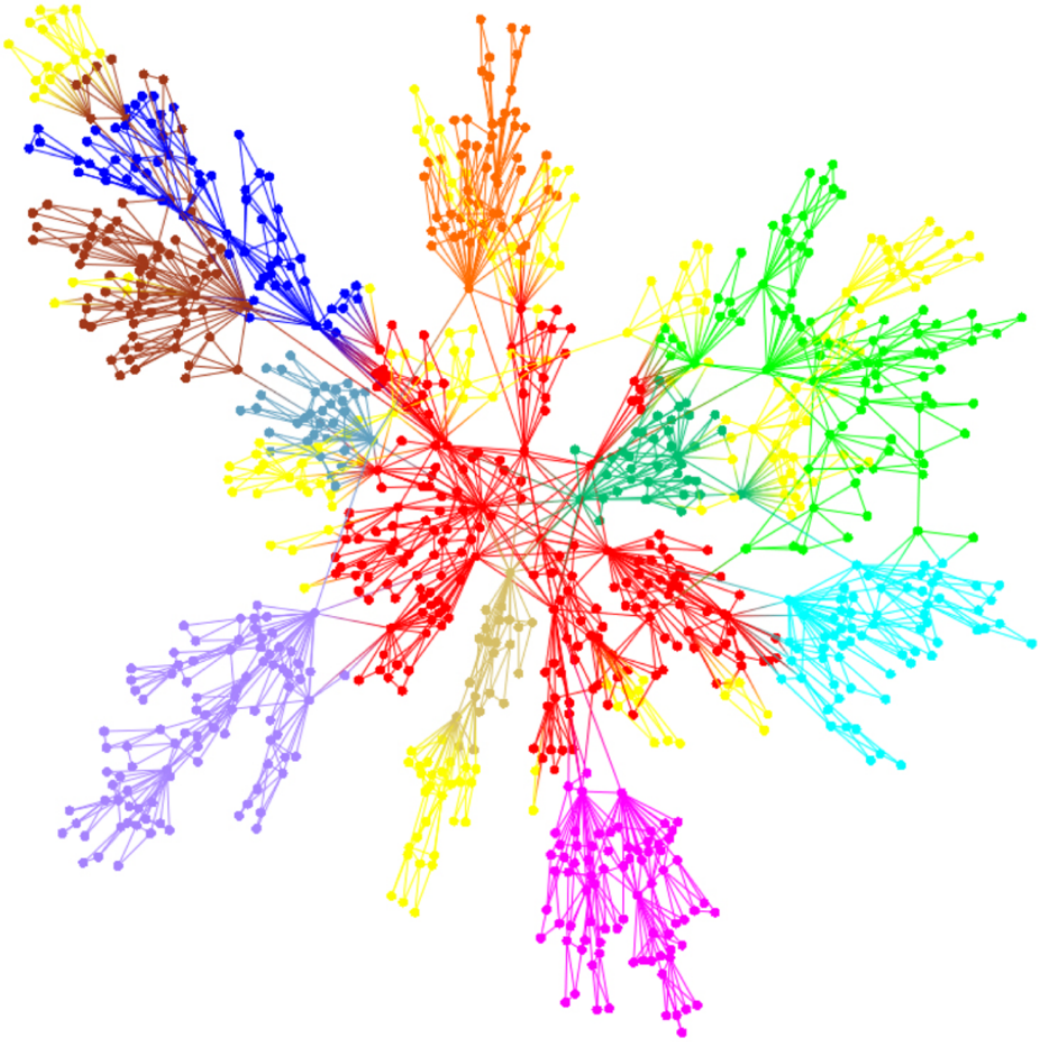


Fig. 2-19 A network generated by Pasta's model [58]

2.3 Summary

In this chapter, some previous work regarding network generation model, especially high-modularity network generation models were presented.

To begin with, some important properties of network were introduced, such as small-world phenomenon and scale-free topology. Later on, some basic networks and network generation models are listed as follows:

(1) Classical networks and models: Complete network (CPN), Erdős-Rényi (ER) model, etc.

(2) Watts-Strogatz (WS) model;

(3) Barabási-Albert (BA) model;

(4) Connecting Nearest Neighbor (CNN) model;

(5) Other models: Holme-Kim (HK) model, etc.

Moreover, the definition of modularity is brought forth to explain the term of "high-modularity", which is nearly equal to the meaning of good community structure. It measures the strength of a division of a network into communities. After that, three types of network generation model synthesizing networks with communities were disscussed, including:

(1) Lancichinetti-Fortunato-Radicchi (LFR) benchmark;

(2) Kronecker graph model;

(3) Pasta's model.

LFR benchmark synthesizes networks with planted community structures. Kronecker graph model employs a self-similar nesting process to produce a hierarchical

structure, which constructs a sequence of graphs from initiator graph K_1 by iterating the Kronecker product. And Pasta's model is a tunable and growing network generation model with community structures that ensures three properties of communities: internal structure, power-law degree distribution, and high clustering coefficients.

The three models and some basic model like CPN, ER, WS, BA and CNN models are assumed as baseline of experiments in following chapters.

Chapter 3. High-modularity network generation model based on the multilayer network

As discussed in Chapter 2, a network with good community structure has a high modularity. From the viewpoint of modeling real properties of networks, it is significant to develop a reliable high-modularity network generation model that can show good reproducibility. Unfortunately, network generation models presented in Chapter 2 are inadequate to precisely reproduce real networks. Therefore, in order to synthesize networks with communities, the development of high-modularity network generation model with high degree of accuracy is very urgent.

A multilayer structure is introduced to construct a network because it provides a rich representation of real-world interaction. For example, people in daily life belong to many communities in society, such as family, school, hobby group, and business organizations (Fig. 2-5). Each example is regarded as a community in a single layer of a multilayer network. However, a network that can be observed on social network services (SNSs) does not reflect such multilayer structure, because it combines all communities. Measuring each relationship in each community is fairly difficult. Hence, the utilization of a multilayer structure can uncover such hidden properties of networks and other structure features in real data.

The high-modularity network generation model proposed in this research exploits the multilayer structure behind a real network, because a social network is actually a superimposed multilayer network. With such critical structure property, proposed model can reproduce real networks with higher precision than other high-modularity network generation models. By introducing this multilayer-based model, we have reason to believe

that a statistically more accurate and more realistic representation of real networks can be obtained.

As baseline models, LFR model (LFR), Kronecker graph model (Kronecker) and Pasta's model (Pasta) are talked about in Chapter 2. More details of proposed model are elaborated in next section.

3.1 Overview of model

A social network in real world is regarded as a multilayer network [19], because people are usually members of different social relationships: family, school, hobby group, business organizations, etc. A layer represents one kind of relationship in a social interaction and each layer contains many communities. The networks obtained from social networks are superimposed networks of all communities.

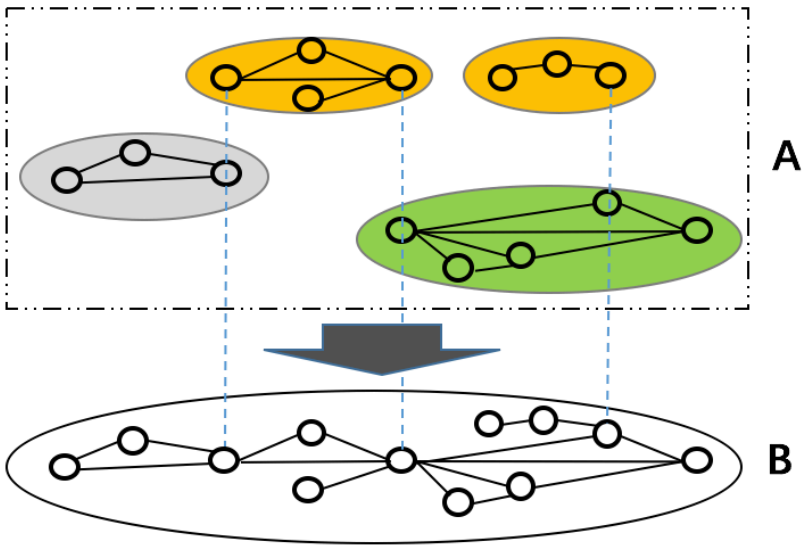


Fig. 3-1 Generation process of proposed model

The generation process of proposed model is shown in Fig. 3-1. Each node in the superimposed network (the inferior network *B* in Fig. 2) belongs to a few communities.

Nodes are connected to other nodes that belong to the shared communities. The link between two nodes in the superimposed network can only appear when there is a link between two nodes in some communities.

Moreover, different communities in the same layer are assumed to be exclusive. They have different sizes and network structures. Each node can belong to only one community in each layer for the simple reason that most people only belong to one family, one university, or one company. In Fig. 3-1, the same colored communities represent communities in the same layer. To imitate a real-world network, our model can change parameters to control the network structures of communities, the size of communities, the numbers of member communities, and so on.

3.2 Model parameters

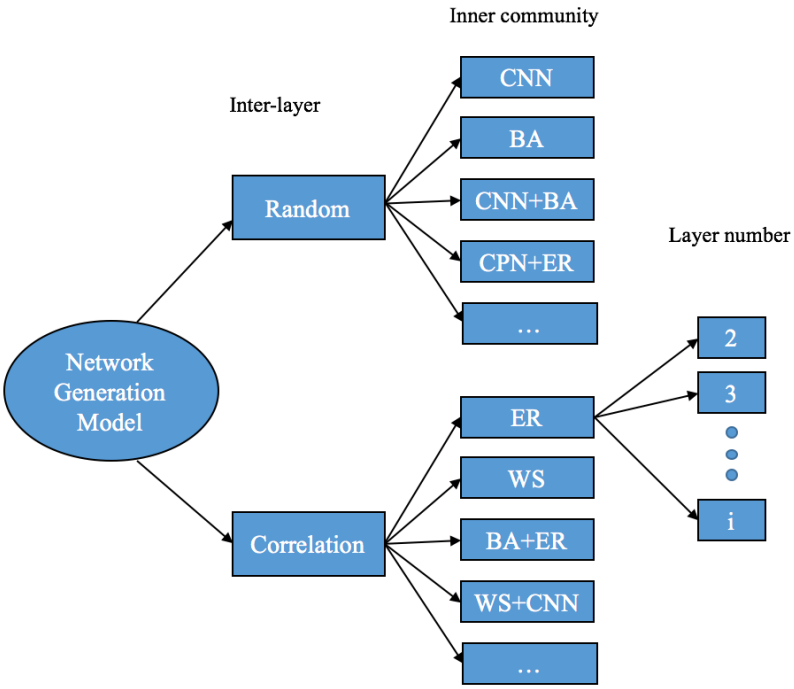


Fig. 3-2 Main parameters of our proposed model

In my study, the model has four critical parts to control the generating process: community size distribution, number of layers, an inner community network model, and inter-layer degree correlation. Fig. 3-2 depicts the simple process of selection.

3.2.1 Community size distribution

The distribution of the community sizes is a functional relationship between two quantities: community size (number of nodes in a community) and the frequency of the same size.

Real networks possess a broad distribution of community sizes. According to some studies, the distribution of real networks can be fairly well approximated by a power law [43], [48], [49], [50], [60], where small-sized communities are numerous but large-sized ones are scarce in the real world.

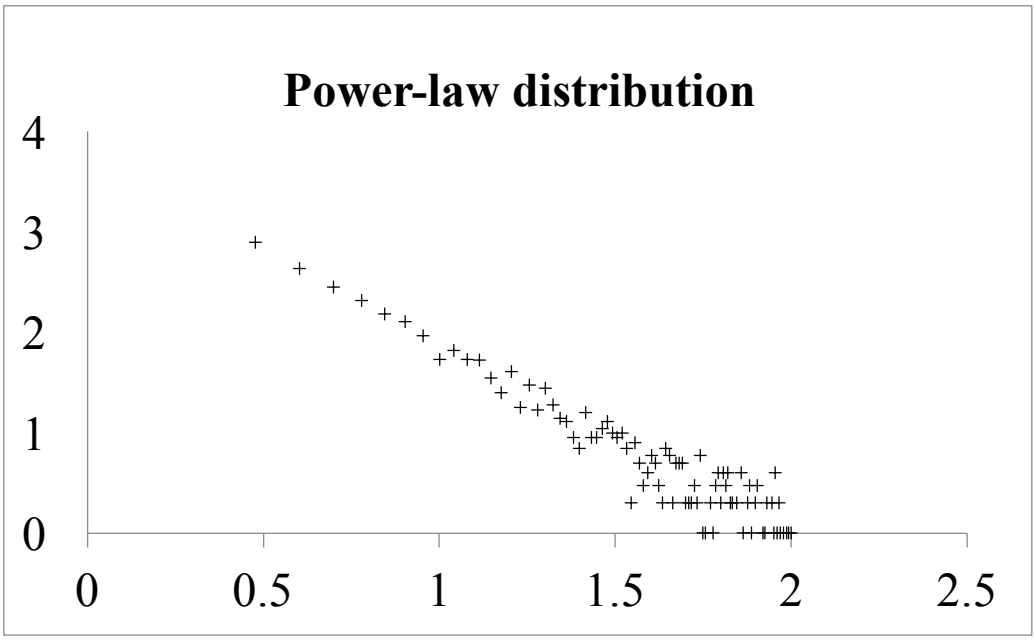


Fig. 3-3 A log-log plot of community size distribution in proposed model

In my model, the size of the communities is taken from power-law distribution $p(k) \sim k^{-\beta}$ with exponent β. We chose typical values of real networks with $1 \leq \beta \leq 2$ [48]. A log-log plot of community size distribution used in proposed model is shown in Fig. 3-3. It follows a power law distribution.

3.2.2 Number of layers

How many layers are necessary to accurately describe the structure of a multilayer network? Recent research has concluded that real networks have a small number of layers [61]. In my model, this value s is selected within a range between 2 and 10. It can be optimized by maximizing the similarity between synthesized networks and target networks in real systems. In Fig. 3-5, the proposed multi-layer model include three layers, where the layer at the bottom of three layers involve two communities.

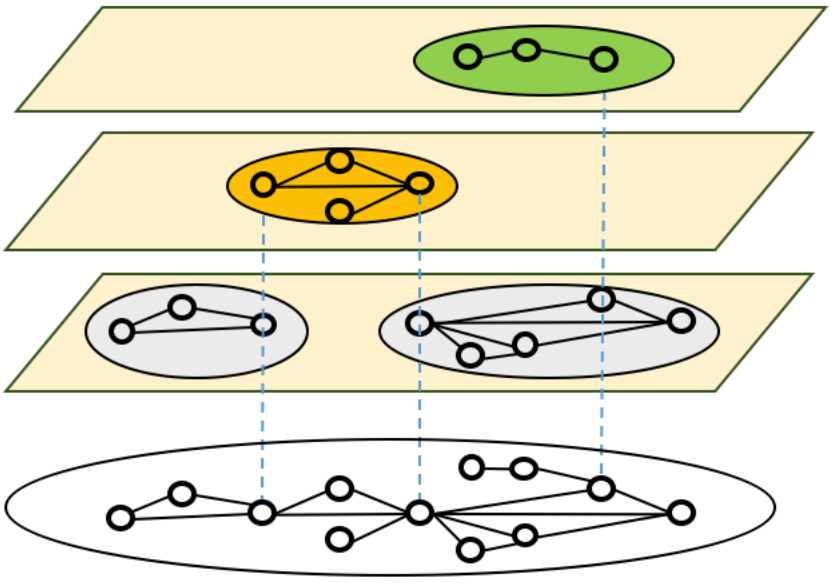


Fig. 3-4 A case of proposed model with three layers

3.2.3 Inner community network model

Nodes can be connected within a community in many ways. One simple idea is to connect all the nodes in one layer with each other by the links as a complete network. But in the social interaction of human society, not all the members in a group are friends. In this simulation, I used five basic models and their combinations to build links among nodes within a community. The basic models include:

(1) Erdős-Rényi (ER) model [13], [14], [15];

(2) Watts-Strogatz (WS) model [16];

(3) Barabási-Albert (BA) model [17];

(4) Connecting Nearest Neighbor (CNN) model [18];

(5) Complete network (CPN).

Two of schemes above are combined into an inner community model. For example, BA+CNN is a hybrid model where layers with small-sized communities are created by the BA model and the others with large-sized communities are generated by the CNN model. The algorithm simulates the h percent of the layers with the first model and the $1-h$ percent with the second one so as to secure the best generated network. However, complete network cannot be the second model because the largest communities, which are all created by complete networks, fail to satisfy real-world data (due to too much links). Consequently, 20 combination models are used in the construction of an inner community (see Tables 6 to 9).

3.2.4 Inter-layer degree correlation

The way of connecting nodes in different layers with various degrees will lead to a completely different network structure. When performing layer aggregation, it is considered whether the same node entity has a degree correlation in different layers. The inter-layer degree correlation controls connection mode of links among communities in different layers.

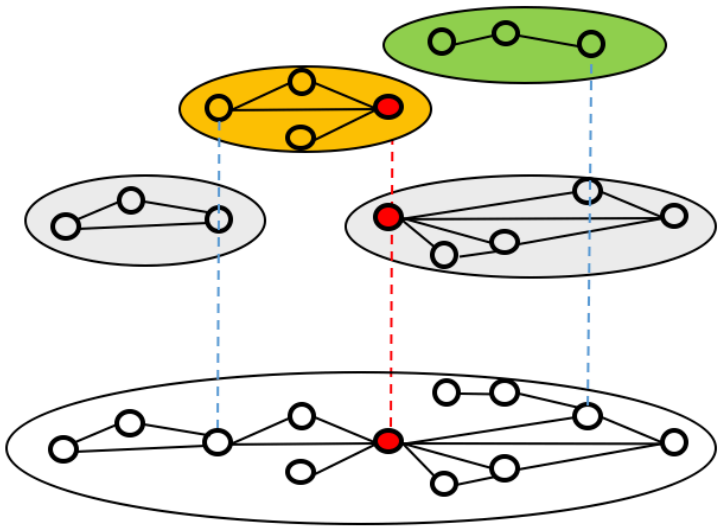


Fig. 3-5 Inter-layer degree correlation in proposed model

Two correlation methods are implemented in proposed model: correlation and random. If a node has a degree correlation with each layer, the node with the higher degree in one layer is also the high-degree node in the other. When there is no degree correlation between inter-layers, the degree of each node is selected randomly in each layer. As is described in Fig. 3-5, the red node has a highest degree in gray community of one layer. At the same time, it is a high-degree node in orange community of the other layer. It can

be concluded that there exists an inter-layer degree correlation for the two communities in different layers.

3.3 Procedure of Algorithm

My algorithm of multilayer-based model begins with the number of nodes and links (N and M) as input data. The proposed model's procedure is elucidated through the following steps:

(1) Create s layers ($2 \leq s \leq 10$) and determine the number of nodes at each layer N_l by N and parameter t ($0 < t < 1$):

$$N_l = N \cdot t \tag{3.1}$$

Algorithm will traverse t from 0.1 to 1 by every 0.1 step and select the best value for each layer. This step is shown in Fig. 3-6.

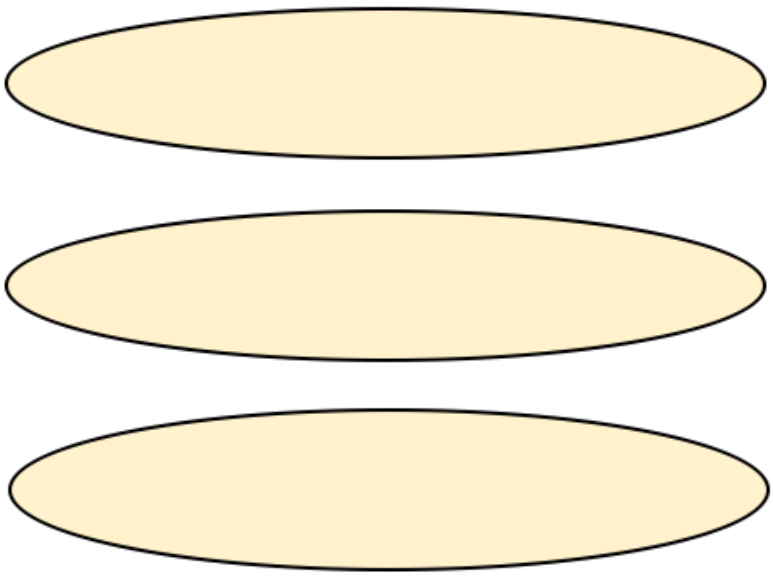


Fig. 3-6 The creation of layers

(2) Create a number of communities whose sizes are selected from a power-law distribution with exponent β ($1 \leq \beta \leq 2$). The number of nodes in all the communities is calculated by formula as follows:

$$\sum N_c = \sum_{l=1}^{s} N_l \quad \textbf{(3.2)}$$

where N_c is the community's size. At this time, no links are constructed within communities. The creation of communities can be expressed by Fig. 3-7.

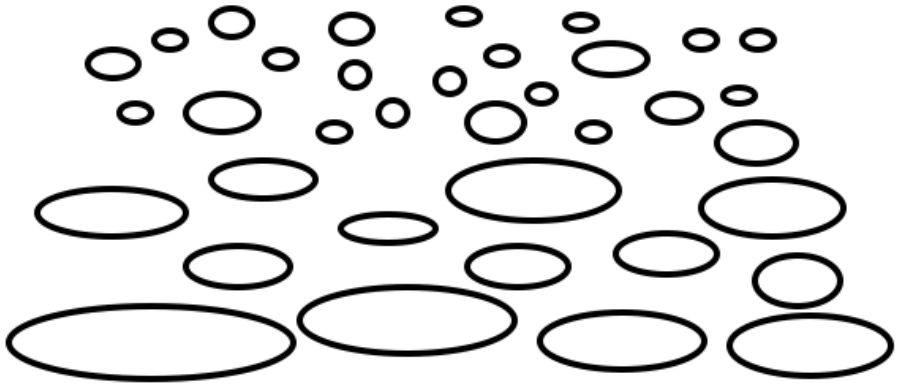


Fig. 3-7 The creation of communiites

(3) Assign communities to each layer in ascending order by size. The smallest community is placed first and then the next larger size. If the number of nodes of the current community exceeds the capacity of the current layer, we turn to the next layer. Fig. 3-8 gives the assigning process of communities.

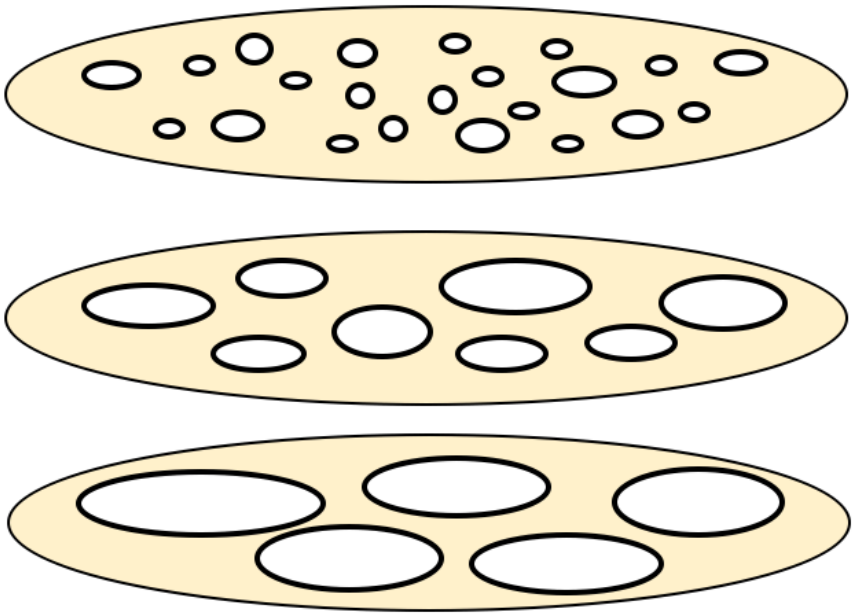


Fig. 3-8 The placement of communities on layers

(4) Connect the nodes in each community by the selected inner community network model. A link exists between two nodes if they are connected by a corresponding model. All possible parameter values of model are traversed. Here, the inter-layer degree correlation is considered, too. If there is a degree of correlation, the node with a higher degree in one layer will also be a high-degree node in the other. Such formation process of inner-community structure is sketched in Fig. 3-9.

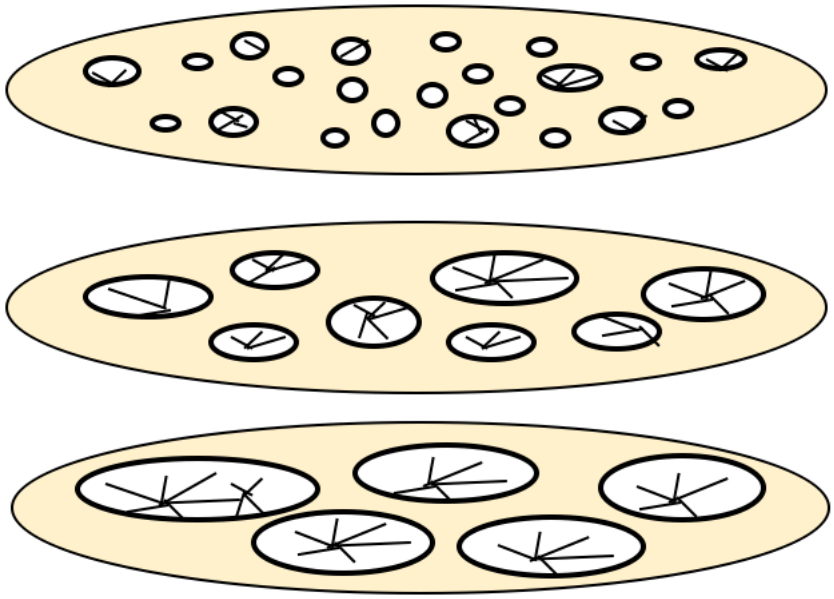


Fig. 3-9 Connection of nodes within communities

(5) Aggregate all the layers of the multilayer network into one layer. Nodes $x_i(i = 1, \dots, q)$, which represent the same entity in different layers of the multilayer network, are integrated into one node x. Nodes x and y in the aggregated network are linked if corresponding x_i and y_j in a multilayer network are connected in either of their respective layer communities. When the aggregated network consists of too many or too few links, we reject it (the number of links of generated network M_0 with $M_0 < M - 5\%M$ or $M_0 > M + 5\%M$), because it cannot reproduce a real-world network with a proper number of links. Fig. 3-10 displays the operation of aggregating all communities in different layers.

(6) Traverse every possible parameter value and select the best ones by optimizing distance D (D will be defined in next section).

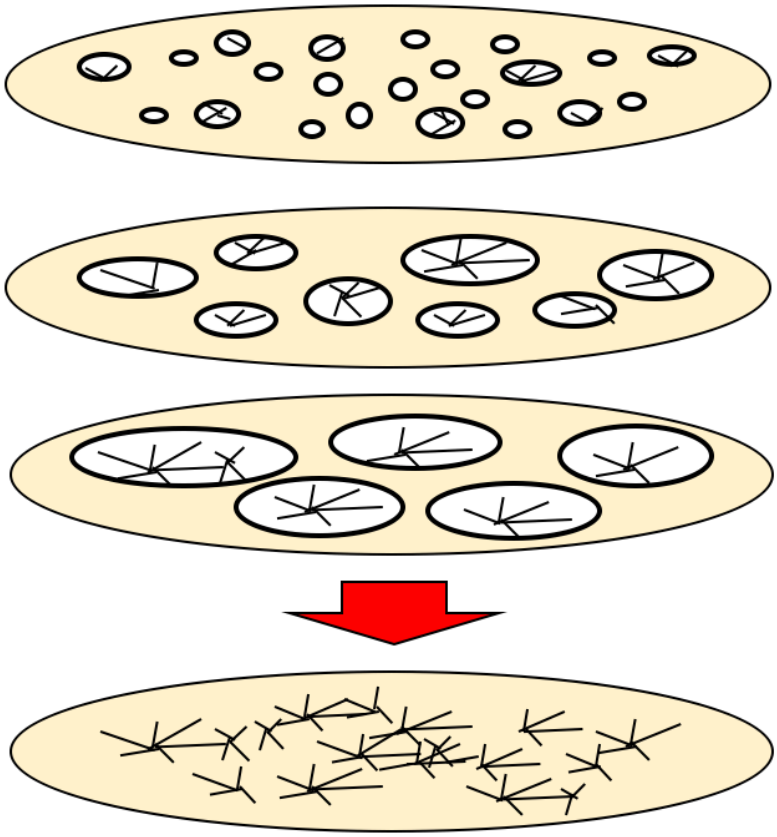


Fig. 3-10 Aggregation of all communities in different layers

Some discrete values of parameters used in generation algorithm above are listed in Table. 3-1.

Table. 3-1 Values of parameters in algorithm

Parameter	Discrete values of each parameter
s	2, 3, 4, 5, 6, 7, 8, 9, 10
t	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
β	1, 2
h	0.2, 0.4, 0.6, 0.8

3.4 Measurement

The values of parameter are optimized by minimizing the distance between the real-world data and the generated data in the proposed model. Distance D is defined by the normalized Euclidean distance of the network features between two networks. In this paper, five representative network features below are utilized to quantitatively measure the whole network.

(1) Clustering coefficient C;

(2) Assortativity r;

(3) Modularity Q;

(4) Power index of degree distribution γ;

(5) Coefficient of determination of degree distribution R^2.

3.4.1 Clustering coefficient *C*

A clustering coefficient [16] measures the extent to which a network's nodes tend to cluster together. In social networks, if most friends of a person are also friends of each other, the clustering coefficient is relatively high. It measures the clannishness of a typical circle of friends. For instance, networks produced by WS model have a high clustering coefficient.

For node i in a network, clustering coefficient C_i, which is defined by the ratio of the existing links between the nodes within its neighborhood to the number of potential links, can be described as the following formulas:

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{3.3}$$

$$C = \frac{1}{N}\sum_i C_i \tag{3.4}$$

where k_i is the degree of node i. $k_i(k_i-1)/2$ is the number of edges that might exist near its neighborhood. N is the number of nodes and E_i is the number of connected links among the nodes within the neighborhood of node i. The overall clustering coefficient of the network is the average of each clustering coefficient C_i. This network index shows how well the local nodes of a network cluster together.

3.4.2 Assortativity *r*

Assortativity [62], [63], or assortative mixing, reflects the preference that nodes in a network tend to be connected to others in a similar way. It indicates the degree correlation relationship between nodes. As the specific measure of assortativity varies, one can be defined as following formula:

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i+k_i)\right]^2}{M^{-1}\sum_i (j_i{}^2+k_i{}^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i+k_i)\right]^2} \tag{3.5}$$

where j_i and k_i are the degrees of two nodes that are connected by link i, and M is the number of links.

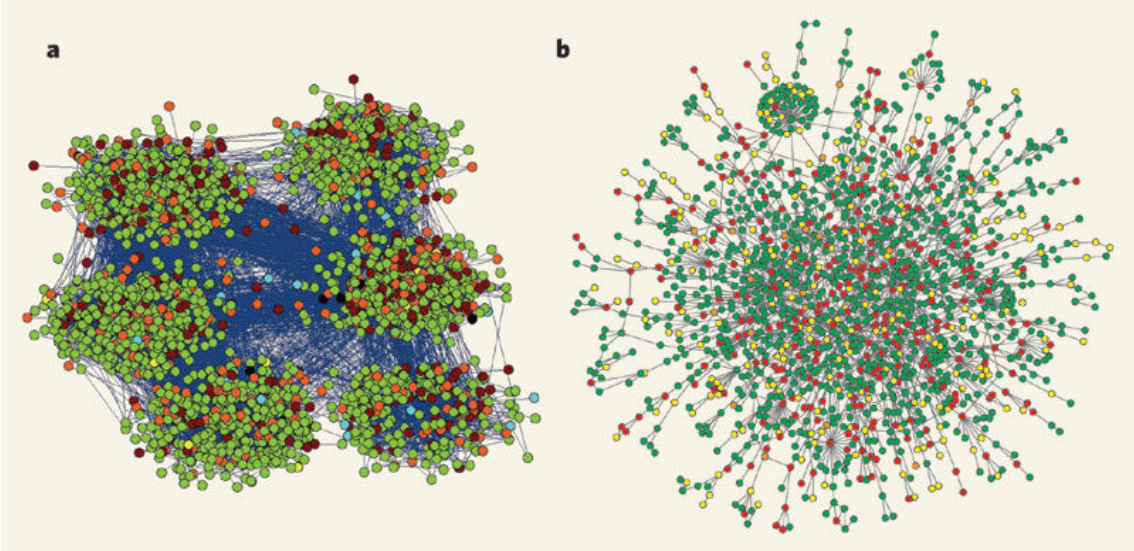


Fig. 3-11 Networks with positive and negative assortativity [64]

(a) nodes connect to nodes with similar degree (r>0)

(b) nodes connect to nodes with different degree (r<0)

The value of assortativity r lies between −1 and 1 ($-1 \leq r \leq 1$). When r falls into a positive value, nodes of similar degrees prefer to attach to each other (Fig. 3-11 a). Nevertheless, when r becomes negative, nodes of different degrees prefer to attach to each other (Fig. 3-11 b). That is, high-degree nodes tend to link to low-degree nodes.

According to Newman [62], [63], assortative mixing (r>0) is a characteristic of social networks. On the other hand, disassortative mixing (r<0) patterns are often found in other networks like technological and biological networks.

3.4.3 Modularity Q

Modularity defines how well a network is divided into communities. The detailed expression in mathematics has been discussed in Chapter 2. The network in Fig. 3-12 indicates a high-modularity property. Dense links exist within each community, whereas

fewer links go across two communities.

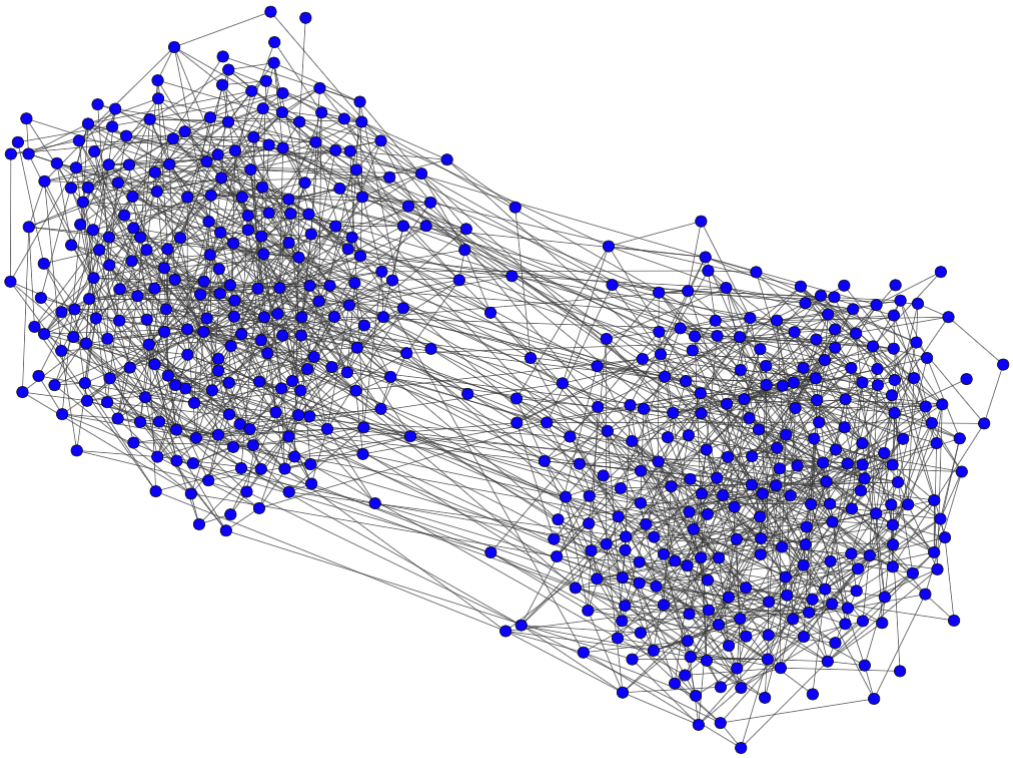


Fig. 3-12 A high-modularity network [65]

Modularity is an important characteristic of networks in real systems, reflecting whether a network exhibits the property of a community. In my study, since we produce networks with community structure by simulations, it is a significant index for evaluating networks.

3.4.4 Power index of degree distribution γ

Degree distribution $p(k)$ is the probability that one node connects to k other nodes. Even though it only captures a small amount of information about a network, degree distribution can distinguish different types of networks, especially a scale-free network.

If a network follows a power-law distribution, its degree distribution is determined by its power index γ [17]. The relation between degree k and distribution $p(k)$ can be written as the following formula:

$$p(k) \propto k^{-\gamma} \tag{3.6}$$

For networks with a scale-free property, constant parameter γ represents the power index of degree distribution. The γ value of most networks typically falls between 2 and 3, except for rare exceptions.

3.4.5 Coefficient of determination of degree distribution R^2

The coefficient of determination measures importance in statistical analysis [63]. When analyzing networks, this coefficient describes how well the degree distribution obeys a power-law distribution and explains the variability in the dataset. The following is a general definition:

$$R^2 = \frac{\sum_i (f_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3.7}$$

where y_i is the observed data with a value of log $p(k)$, $\bar{y}$ is the mean of y_i, and f_i is the predicted data value associated with y_i. The coefficient of determination R^2 ranges from 0 to 1. The closer the R^2 value is to 1, the more the network exhibits a typical scale-free property.

3.5 Evaluation Function

Even though other network features may exist, this study utilize the above five

measurements, because they can capture three vital properties of real network: community structure, small-world and scale-free. Furthermore, the distance of two networks is quantitatively evaluated by D, which is a normalized Euclidean distance between produced network G_i and target network G_0. Distance D is defined through the following equation [66], [67].

$$D(G_i, G_0) = \left(\frac{C_i - C_0}{\sigma_C}\right)^2 + \left(\frac{r_i - r_0}{\sigma_r}\right)^2 + \left(\frac{Q_i - Q_0}{\sigma_Q}\right)^2 + \left(\frac{\gamma_i - \gamma_0}{\sigma_\gamma}\right)^2 + \left(\frac{{R^2}_i - {R^2}_0}{\sigma_{R^2}}\right)^2 \quad \textbf{(3.8)}$$

where C, r, Q, γ, and R^2 are the network features of a network. σ is the standard deviation of each feature calculated from a real-world network dataset and is used as normalization. With this evaluation function, we measured how well our proposed model reproduced four real-world datasets: A Facebook network, a Renren network, a collaboration network, and an air traffic control network.

Why were a network index and a normalized Euclidean distance used in the evaluation? It is widely admitted that generating networks with identical target topologies is impossible. There is no perfect solution for representing the similarity of two networks. As a result, employing a proper network index is an effective way to partially capture topologies among networks. Modularity measures the strength of the division of a network into communities, thus is able to reflect community structure. A clustering coefficient captures the richness of tie weights, which is an important indicator of small world network. Scale-free topology can be described by power index of degree distribution and coefficient of determination of degree distribution. Other index of assortativity represent the inclination of connection among nodes. Finally, extra indexes can be easily added to this function if it needs to evaluate new parts of the network topologies. As for the weight of each item in evaluation function, emphasis on individual

network feature is not placed, because this research treats different structure properties equally. Consequently, distance D is calculated with the sum of five items.

3.6 Summary

This section proposed an effective high-modularity network generation model by layer aggregation based on a multilayer network to represent real networks. Section 3.1 gave an overview of this model. Four integral parameters were introduced in section 3.2, which is presented as follow:

(1) Community size distribution;

(2) Number of layers;

(3) Inner community network model;

(4) Inter-layer degree correlation.

In section 3.3, the procedure of algorithm is elaborated through many steps with figures. First, the power-law distribution is employed to generate community size distributions. The number of layers was set from 2 to 10. Later on, many basic models (ER, BA, WS, CNN model, a complete network) and their combinations were chosen to build links within nodes inside a community. Finally, the inter-layer degree correlation was selected by random or positive correlation.

Section 3.4 discussed the specific network features exploited in evaluation equation, including clustering coefficient C, assortativity r, modularity Q, power index of degree distribution γ, coefficient of determination of degree distribution R^2.

Section 3.5 use the five network features to measure the similarity between the targeted and produced networks. The similarity can be calculated quantitatively by a

normalized Euclidean Distance D with all of these features.

Chapter 4. Simulation of network generation

Simulation models are constructed as approximate imitations of real life. In this study, simulations of network generation were enforced to imitate real networks. As every newly designed simulation model needs to be verified, this chapter brought forth 4 real network datasets of different types as ground truth. Their network features were calculated in advance. The Euclidean distance *D* discussed in Chapter 3 was then adopted for the goal of analyzing the accuracy of generation algorithm.

4.1 Network Datasets

Four datasets of real networks are introduced in our evaluation experiments: two social networks (Facebook and Renren network), a collaboration network, and an air traffic control network.

4.1.1 Facebook network

Facebook network dataset were acquired from the open data repository in networkrepository.com[*1]. The network used in our experiments contained 8,578 nodes and 405,450 links and reflects the social relationship within Yale University. Only the information of the link relationships can be seen, and there is no personal profile information about each node.

[*1] http://networkrepository.com/index.php

4.1.2 Renren network

Renren network is a Facebook-like SNS network in China. This study collected the link relationships of users from Peking University (PKU) using the Renren API [47]. The network data acquired from Renren have 2,309 nodes and 60,532 links. These users were all 2009 PKU graduates.

4.1.3 Collaboration network

The data of collaboration networks were obtained from the Stanford Large Network Dataset Collection (SNAP[*1]). This research utilized a scientific co-authorship network called "ca-GrQc" (Arxiv General Relativity and Quantum Cosmology), which is an undirected network with 5,242 nodes and 14,496 links. In this dataset, nodes represent scientists and links represent collaborations of paper. The data cover papers from January 1993 to April 2003.

4.1.4 Air traffic control network

The air traffic control network dataset is part of the Koblenz Network Collection[*2]. This network was constructed from the USA's Federal Aviation Administration (FAA) National Flight Data Center (NFDC), Preferred Routes Database. The nodes represent airports or service centers, and links are created from strings of preferred routes recommended by the NFDC. The network having 1,226 nodes and 2,408 links were

[*1] http://snap.stanford.edu/data/index.html

[*2] http://konect.uni-koblenz.de/networks/

identified in 2016. Rather than a social network as Facebook or Renren, this dataset is a traffic real-world network. Employing this dataset to verify whether our model can capture the important features of a non-social network.

4.2 Simulation and experimental results

Simulation was carried out ten times to secure networks for each group of parameters to synthesize a network similar to the target network. This section compared proposed model with the baseline models by calculating distance *D*.

4.2.1 Simulation process

Simulations were conducted with the baseline and proposed multilayer-based models to mimic four real networks. Since many parameters are real numbers, traversing all the continuous values of the parameters is impossible. Discrete values were selected with intervals. Table. 4-1 summarizes the parameters set for the inner network generation model.

Table. 4-1 Discrete values of parameters in inner model

Inner model	Discrete values of each parameter
ER (*p*)	0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
BA (*m*)	1, 2, 3, 4, 8, 12, 16, 20
WS (*m*)	4, 8, 12, 16, 20
CNN (*p*)	0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

Table. 4-1 lists the discrete values of the parameters in the inner community model

of our proposed generation model. The letter in parenthesis under the column of inner model represents the main parameter of each model. The ER model has a random graph where each edge has fixed probability p of being present or absent. Algorithm traverse p from 0.05, 0.1, 0.2 to 0.9. The BA model starts with an initial network. At each time, a new node is added, and m links are created between the new node and the nodes selected from the graph. Since a generated network is constrained by the number of nodes and links, parameter m whose value is too big will lead to a network with too many links. Therefore, the biggest value of m is 20. Parameter m of the WS model is the number of edges connected to each vertex and represents the local neighborhood size. Since it must be an even number, m traverses between 4 and 20 by four steps, which is also constrained by the number of nodes and links like the BA model. Finally, the CNN model has a single parameter p called the conversion probability. p is decided by the number of nodes and links, too. The discrete values of p are selected from 0.05, 0.1, 0.2 to 0.9.

The algorithm selected all the possible discrete values of the parameters to produce networks with N nodes and M links. Furthermore, four network features and the distances from the real-world data were calculated for both the baseline and proposed models. The minimization of distance D is executed in order to find the best parameters for every generation model.

Table. 4-2 Network topology of Facebook network (Yale) and simulation results with existing and proposed models

	Facebook (N:8578 M: 405450)	ER	BA	CNN	Kronecker [k=13]	Pasta [c=10; Pt=0.5 Pc=1.0]	Multilayer [ba(m=32)+ cnn(p=0.1), corr, 10 layers]
Clustering coefficient C	0.2424	0.0110	0.0445	0.4956	0.1577	0.3779	0.2803
Assortativity r	0.0186	0.0025	-0.0231	-0.1308	-0.2395	0.0079	0.0102
Modularity Q	0.1636	0.0218	0.0051	0.0540	0.0449	0.3303	0.1453
Power index of degree distribution γ	1.1979	-1.6882	1.9661	1.0418	0.6503	1.4099	1.1489
Coefficient of determination R^2	0.7027	0.0366	0.8196	0.8574	0.3852	0.8219	0.6075
Distance D		33.8028	4.2414	5.4465	3.1861	1.6794	0.1133

Table. 4-3 Network topology of Renren network (PKU) and simulation results with existing and proposed models

	Renren (N:2309 M: 60532)	ER	BA	CNN	LFR [β=2; μ=0.2]	Kronecker [k=11]	Pasta [c=20; Pt=0.5 Pc=1.0]	Multilayer [er(p=0.5)+ cnn(p=0.05), corr, 3 layers]
Clustering coefficient C	0.4566	0.0223	0.0738	0.5353	0.5431	0.2084	0.3864	0.4152
Assortativity r	0.2421	0.0017	-0.0331	-0.1001	0.3320	-0.2725	0.0765	0.1339
Modularity Q	0.5716	0.0364	0.0239	0.0844	0.5219	0.0671	0.4548	0.5833
Power index of degree distribution γ	0.8658	0.2942	1.8628	0.9679	0.9133	0.6104	1.4327	0.8220
Coefficient of determination R^2	0.5462	0.0022	0.7887	0.8220	0.7794	0.3815	0.8206	0.4462
Distance D		18.3797	18.7636	9.3651	0.7250	13.8717	2.4723	0.2899

Table. 4-4 Network topology of collaboration network (ca-GrQc) and simulation results with existing and proposed models

	Collaboration network (N:5241 M: 14484)	ER	BA	CNN	LFR [β=1; μ=0.1]	Kronecker [k=12]	Pasta [c=10; Pt=1.0; Pc=0.5]	Multilayer [cpn+cnn(p=0.4), rand, 6 layers]
Clustering coefficient ***C***	0.6865	0.0011	0.3266	0.4685	0.5340	0.0011	0.7257	0.4982
Assortativity ***r***	0.6593	-0.0122	-0.1629	0.2182	-0.2092	-0.4442	0.0748	0.6476
Modularity ***Q***	0.7584	0.1389	0.2077	0.5965	0.7030	0.1991	0.5969	0.7175
Power index of degree distribution γ	2.0412	*NaN*	1.4833	1.7998	2.0079	2.0123	1.7020	1.9082
Coefficient of determination R^2	0.8792	*NaN*	0.6565	0.9070	0.9431	0.6156	0.8521	0.8973
Distance ***D***		56.6383	24.6079	5.4465	11.9193	43.4824	6.2182	1.3256

Table. 4-5 Network topology of air traffic control network and simulation results with existing and proposed models

	Air traffic control (N: 1226 M: 2408)	ER	BA	CNN	LFR [β=2; μ=0.6]	Kronecker [k=10]	Pasta [c=10; Pt=0.5; Pc=0.5]	Multilayer [er(p=0.2)+ba(m=1), rand, 3 layers]
Clustering coefficient ***C***	0.0733	0.0035	0.7942	0.4685	0.0407	0.0095	0.3042	0.0538
Assortativity ***r***	-0.0152	-0.0381	-0.3295	0.2182	0.1028	-0.3719	0.0631	0.0129
Modularity ***Q***	0.5340	0.2517	0.3496	0.5965	0.2687	0.2629	0.5345	0.5506
Power index of degree distribution γ	1.9217	*NaN*	1.0279	1.7998	1.2839	1.6752	1.8408	1.9683
Coefficient of determination R^2	0.8572	*NaN*	0.4706	0.9070	0.9356	0.6048	0.8976	0.9151
Distance ***D***		8.4759	23.7358	5.4465	3.7798	4.7467	1.9402	0.0568

4.2.2 Comparison between Baseline and Proposed Models

This section makes a comparison between proposed models and baseline modes, such as LFR benchmark, Kronecker graph model and Pasta's model. The distances between the generated network and real networks with different models are straightforwardly displayed in Fig. 4-1.

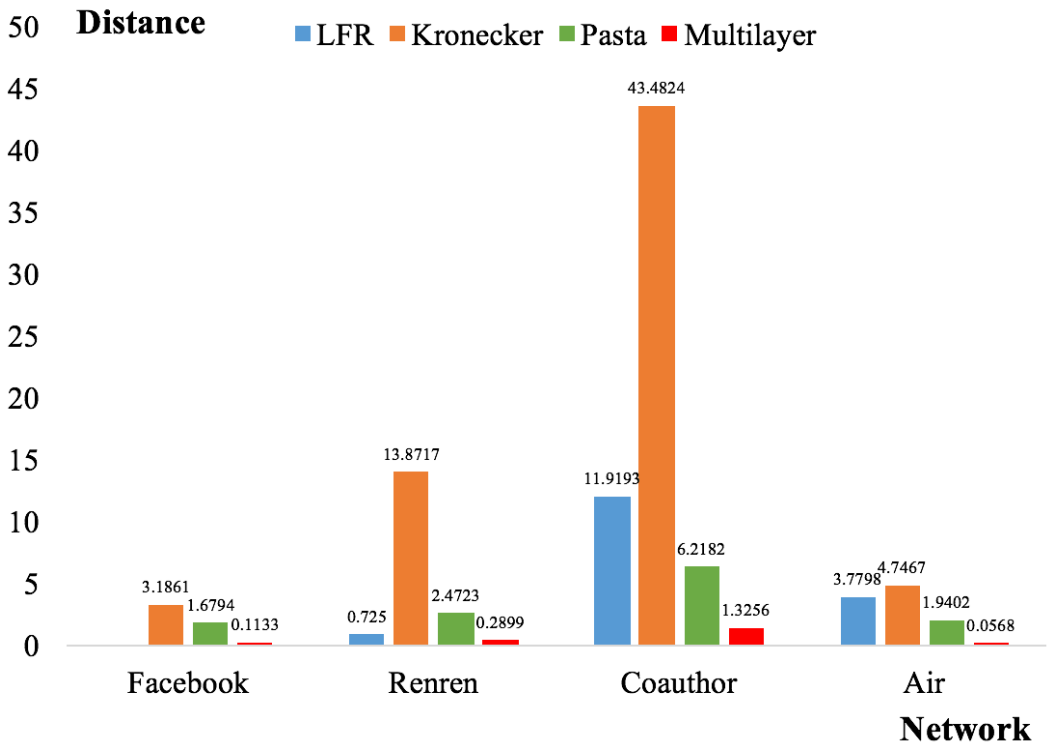


Fig. 4-1 Comparison of distance *D* for different models

Table. 4-2 to Table. 4-5 show the details of our experimental results for four datasets: Facebook, Renren, the collaboration network, and the air traffic control network. The features of the four networks and their distances from the generation models are precisely recorded. The first row exhibits the real-world network, three basic models (ER, BA, and

CNN), three baseline models (LFR, Pasta, and Kronecker, discussed in related work), and our multilayer-based model. The main parameters of the models are listed in square brackets. In proposed model, three parameters represent the inner community model, the inter-layer correlation, and the layer number. The parameter selection of the inner community model is listed in Table. 4-1, and the inter-layer correlation is given by two kinds of situations: correlation or random. The number of layers is an integer that ranges between 2 and 10 (See Section 3.2.2). The last column shows the best outcomes. Minimum distance D is selected from different parameter combinations by our multilayer model. Since the ER model lacks a scale-free property, it is difficult to estimate reasonable constants γ and R^2 to represent the degree distribution. Thus, γ and R^2 in Table. 4-4 and Table. 4-5 are marked by *NaN*. They are omitted when calculating distance D. Moreover, modularity Q of the ER random network in Table. 4-4 (0.1389) and Table. 4-5 (0.2517) is high. According to a previous work of Guimera et al. [68], ER graphs have high modularity due to the fluctuations in the establishment of links, which are magnified by the large number of ways in which a network can be partitioned into modules. Since these parameters of network indexes do not affect the final evaluation of the distance compared to other networks, special adjustments are not conducted for modularity.

According to Table. 4-3, Table. 4-4, and Table. 4-5, modularity Q of a synthesized network is high for LFR, Pasta, and multilayer-based model: $Q>0.4$ for Table. 4-3 and $Q>0.5$ for Table Table. 4-4, where a network has an obvious community structure if $Q>0.3$ [45]. Hence, multilayer-based model along with the LFR and Pasta models can produce high-modularity networks or networks with strong community structure. However, in Table. 4-2, the modularity is so low that most models could never produce networks with high modularity. In addition, since the LFR model cannot reach a

convergence to reproduce a Facebook network (Yale), the result is not listed in Table. 4-2 and Fig. 4-1. Other basic models, like the CNN model, cannot generate networks with community structure as real-world networks, so they barely acquired networks with high modularity.

For assortativity *r*, multilayer-based model can replicate this value more precisely than the other models in all four datasets because it incorporated the inter-layer degree correlation of the nodes in the algorithm. Since the inter-layer degree correlation in our model controls the neighbor number of the nodes, proposed model produces a network with a bigger range of assortativity values and finds the best one that resembles the real-world data. From Table. 4-2 to Table. 4-5, multilayer-based model reproduces this property better than the LFR and Pasta models.

Finally, proposed model outperforms other existing models considering Distance *D*. A multilayer model is capable of reproducing two social networks (Facebook and Renren) and the air traffic control network with distance $D < 1$, which represents a high similarity to real networks. Other generation models could never achieve this result, except LFR in Table. 4-1 (with 0.7250, which remains worse than our model's 0.2899). Although *D* is over 1 for the collaboration network, our model still outperformed the other baseline models (Table. 4-4). Considering Fig. 4-1, the red column exhibits that multilayer-based model has the shortest distance for four datasets.

To conclude, proposed multilayer model and the LFR and Pasta models are high-modularity generation schemes, but multilayer-based model outperforms the others on average.

4.2.3 Reason for effectiveness of proposed model

The LFR, Pasta, and proposed models are capable of generating networks with high modularity. Nonetheless, multilayer-based generation model captured the network features in the real world better than the other models. The reason why my proposed method imitated the real-world networks much better than the other models lies in the following two reasons.

One one hand, my model can produce not only multilayer construction but also the settings of the number of layers. Multilayer structure provides a rich representation of a real network. For instance, such previous models as ER, BA, CNN, and LFR from Table. 4-2 to Table. 4-5 are single layer networks, while generated networks of my model can be a multilayer ER, multilayer BA, or a multilayer LFR based on the selected parameters. In fact, mLFR benchmark proposed by Bródka [54] is a multilayer version of LFR, which can be a subset of our model by setting each layer with LFR model. My proposed multilayer model handles such cross-layer information and uncovers the hidden topology features of network in a real system, which are ignored by other existing models. Furthermore, algorithm incorporated a parameter called the number of layers to precisely control the reproduction.

On the other hand, since the inner community model is a mixture of different existing models, the generated networks are hybrid multilayers with such different models as ER, WS, and so forth. Since a multilayer structure simplifies putting various models on disparate layers, I combined diverse models and exploited their specialties to establish my hybrid model. Such a hybrid feature is beneficial for modeling different types of real-world networks and generating the most similar networks in the real world. For example,

a social network (Facebook or Renren) is a mixture of the CNN and other models. The CNN model is useful for capturing some common features of social communication, while the other model controls the discrepancy between the Facebook and Renren data. Considering a co-authorship network, communities in different layers are built up with the CNN model and the complete network. The complete network describes the co-author relationships within the tiny communities of a technical paper, whereas the CNN model summarizes co-authorship among different papers. In addition, the air traffic control network is reproduced involving the BA and ER models rather than the CNN model as inner community models, for the reason that it is a non-social network and has a scale-free topology.

4.3 Estimation of hidden structure in real network

This section makes en effort to employ multilayer-based model to estimate hidden structure of real system. By looking into what features of this model contribute to synthesizing networks, it helps to better understanding of the hidden structure in real network.

The generation algorithm combines different inner community network models to reproduce four networks. All of the parameters were traversed for 20 different inner models and the best results were chosen. Table. 4-6 to Table. 4-9 illustrate the distances between the real-world networks and the simulation results with our models. The horizontal and vertical axes represent two inner community network models in my algorithm. The horizontal axis encompasses the inner models to create layers with large-sized communities, while layers with small-sized communities are created by models in the vertical axis. The numbers in the parentheses are the layers for constructing a

multilayer network. In this section, the hidden structures are discussed for the four networks according experimental results.

4.3.1 Facebook network

According to Table 6, models with 8 to 10 layers generate the shortest distances for most occasions. Hence, a tendency exists where Yale's social network has a relatively high layer structure. The best one is a 10-layer network with BA+CNN as an inner model (the red figure in Table. 4-6). Furthermore, the proposed models (with the BA model, the CNN model, or their combination as an inner model) tended to outperform other combinations. However, users from Yale did not exhibit an apparent community structure because Facebook dataset's modularity is too low (0.1636) (cf. Table. 4-2).

Table. 4-6 Distance from Facebook network (Yale) with different inner model combinations (Horizontal axis: models in large-sized communities)

Distance *D*	ER	BA	WS	CNN
ER	0.6114 (8)	0.4527 (8)	0.4758 (10)	0.3965 (8)
BA	0.4873 (8)	0.2592 (9)	0.4287 (10)	**0.1133** (10)
WS	1.0578 (9)	0.4930 (9)	1.7686 (8)	1.2752 (10)
CNN	0.4650 (7)	0.2821 (5)	0.3697 (8)	0.2060 (9)
CPN	1.9499 (8)	1.6361 (4)	2.1306 (5)	3.4626 (5)

4.3.2 Renren network

The distance between the Renren and the produced networks is depicted in Table. 4-7 with different combinations of inner community network models. Unlike the Facebook data, the Renren network enjoys a lower layer (mostly between 3 and 4) in the multilayer structure. A 3-layer network with the ER+CNN model and inter-layer correlation produced the best Renren data result.

Table. 4-7 Distance from Renren network (PKU) with different inner model combinations (Horizontal axis: models in large-sized communities)

Distance *D*	ER	BA	WS	CNN
ER	2.4501 (4)	1.7929 (3)	1.9555 (9)	**0.2899** (3)
BA	2.8298 (3)	3.2489 (3)	2.2715 (3)	0.2969 (3)
WS	2.3805 (4)	1.3842 (4)	1.5864 (3)	0.5400 (4)
CNN	2.5782 (4)	2.7413 (5)	1.4147 (3)	0.4979 (5)
CPN	1.8828 (3)	0.7104 (5)	1.5945 (2)	0.8020 (3)

Following part makes a comparison of two similar online social networks. From the viewpoint of the number of layers, Facebook tends to have more layers than the Renren network. This expresses that Yale (Facebook) has more kinds of relationships among users, who might come from the following communities: fraternities, sororities, academic majors, grades, ethnic groups, etc. On the other hand, the Renren network just has a low number of layers, which is partly due to the small size of the dataset. 2,309 nodes have

students who graduated from Peking University in 2009. Such an ivy league university like Yale might have more kinds of relationships that go across their students' social communication than a Chinese university due to the diversity of students. Thus, a model with 8 to 10 layers can better express the Facebook data, while a model with 3 to 5 layers represents the Renren network.

Furthermore, let us consider the inner community network model. The Facebook network with BA+CNN and the Renren network with ER+CNN outperformed the other models (Table. 4-6 and Table. 4-7). Both networks utilized the CNN model to connect the links of inner communities with large sizes, because this approach is more suitable for large networks and captures the features of the large clustering coefficient as well as the small-world property, which could never be well exhibited in a small dataset. However, the BA model suggests that the Facebook network has a scale-free property within the nodes, which means that Yale students might be friends with others who have many friends. In contrast, the ER model implies that the Renren network has a free connection between different students in small communities at Peking University.

4.3.3 Collaboration network

The data shown in Table. 4-8 give the distance from the collaboration network for different inner community network models. A combination of the CPN and CNN models as an inner community model performed the best. As for the number of layers, it can be identified that no apparent range for this co-authorship network, like in the Facebook and Renren networks.

Table. 4-8 Distance from co-authorship network (ca-GrQc) with different inner model combinations (Horizontal axis: models in large-sized communities)

Distance *D*	ER	BA	WS	CNN
ER	14.1916 (3)	16.6345 (2)	6.1614 (3)	2.3690 (7)
BA	14.6491 (2)	16.5003 (2)	9.9046 (9)	4.2168 (2)
WS	13.9773 (3)	16.5852 (2)	9.9789 (4)	4.1124 (2)
CNN	13.7575 (2)	16.6140 (8)	8.3799 (2)	4.1100 (2)
CPN	6.9651 (2)	4.7510 (2)	6.2434 (10)	**1.3256** (6)

The CPN+CNN network performed best because of the mechanism behind the co-authorship network. The data used in my model are from a scientific co-authorship network of Arxiv General Relativity and Quantum Cosmology. When constructing a network, all of the author nodes who co-author a paper are connected. Complete networks fit fairly well with such small communities in co-authorship networks. One example of co-authorship network from Information Systems Research Seminar in Scandinavia can be observed in Fig. 4-2. There are many small communities of complete network and one large connected component with several local clusters.

Additionally, the CNN model creates networks with a high clustering coefficient as well as small-world and scale-free properties and captures the basic features of larger communities in collaboration networks. Compared with baseline models, multilayer-based model produced a network with high modularity better than other models. In summary, proposed multilayer-based generation model with a CPN+CNN network as an inner community model produced a network that most closely resembles a co-authorship network in real life.

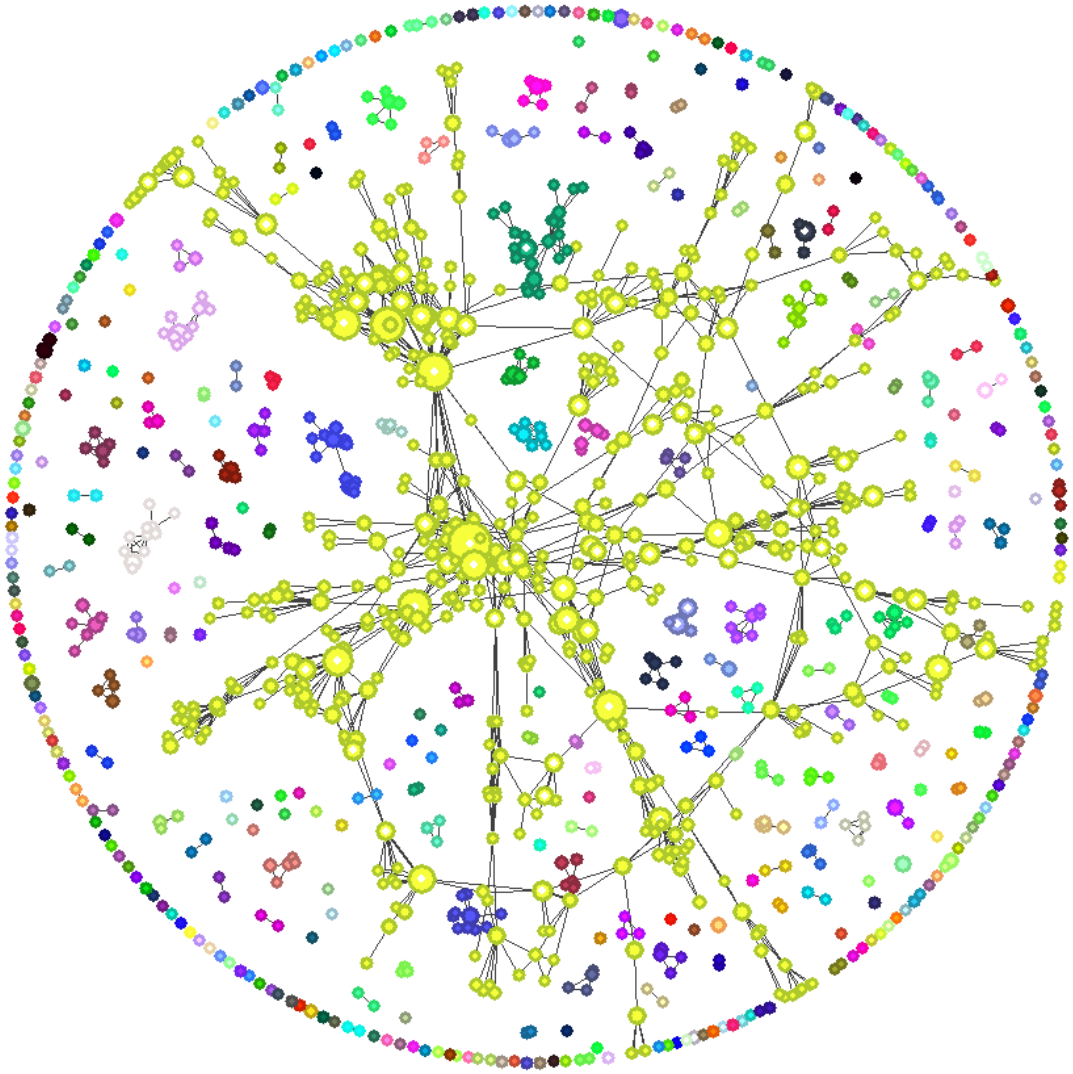


Fig. 4-2 IRIS Conference co-author network from 1978 to 2006

(Source: Molka-Danielsen, IRIS2007, 2007 [69])

4.3.4 Air traffic control network

Unlike other social networks, the CNN model does not play a vital role in the air traffic control network. The ER and BA models capture the best features of the inner community network construction. Also, no obvious tendency of layer's number can be found in Table. 4-9.

Table. 4-9 Distance from air traffic control network with different inner model combinations (Horizontal axis: models in large-sized communities)

Distance *D*	ER	BA	WS	CNN
ER	1.5621 (2)	**0.0568** (3)	3.4606 (2)	0.4696 (3)
BA	0.8587 (2)	0.0993 (6)	3.7990 (5)	0.0818 (5)
WS	1.7069 (3)	0.1999 (4)	3.9288 (3)	0.8238 (10)
CNN	1.7149 (3)	0.0690 (10)	4.1295 (5)	0.1766 (2)
CPN	1.3896 (2)	1.0423 (2)	9.3750 (3)	6.6970 (6)

The BA model denotes that large inner communities in the air traffic control network are built with a scale-free property. This conforms to our image that air transportation networks often include hub airports that are linked to smaller airports, such as the Narita and Frankfurt airports. In addition to connecting hub airports, experimental result suggests that small airports are randomly connected to other small airports because small-sized inner communities are created with the ER model. The multilayers in proposed model can also reflect different airlines in the air traffic control network. Therefore, multilayer-based model is not limited to social networks; it is also applicable to other types of real-world networks, such as air traffic control networks, etc.

4.4 Summary

This chapter discussed simulation experiments and investigate the experimental results. Simulation of network generation was conducted by proposed multilayer-based model, some basic models and three baseline model: LFR benchmark, Kronecker graph

model and Pasta's model. Three real network were immitated by simulation, which are shown as follows:

(1) Facebook network (Yale): 8,578 nodes and 405,450 links;

(2) Renren network (PKU): 2,309 nodes and 60,532 links;

(3) Scientific co-authorship network (ca-GrQc): 5,242 nodes and 14,496 links;

(4) Air traffic control network: 1,226 nodes and 2,408 links.

By comparing proposed model with baseline, it illustrated that multilayer-based model can perform better than others in reproducing four real networks from the viewpoint of network features. Further, this study analyzed what features of this model contribute to synthesizing networks. A number of hidden structure can be found by multilayer-based model. For instance, co-authorship networks may be generated from a combination of CNN model and complete network.

Chapter 5. Reproduction of realistic information diffusion

Diffusions of information which can be observed occur in many single networks, however, it may not be a real case because real networks often involve multilayer structures. Hence, more realistic information spreading over multilayer networks could be realized if we introduce multiple layers and enforce multilayer information propagation.

In previous chapters, a network generation model by layer aggregation based on multilayer network was developed to imitate networks in real data. In this process, a multilayer model was constructed to represent underlying multilayer structure. This chapter utilized such multilayer network to simulate information diffusion in a more realistic way.

A measurement of information diffusion, average influence degree (AID), is calculated. Proposed model succeeded in reproducing information propagation over multilayer network by comparing AID of built multilayer structure with real data. Before stepping in simulation and evaluation parts, some related work about information diffusion, especially diffusion in multilayer network, will be introduced at first.

5.1 Information diffusion model

Information diffusion appears to be one of most important phenomena in real networks. Studies on diffusions or spreading processes have been conducted for many years in various research fields such as informatics, sociology, biology, epidemiology, and so forth. According to Zafarani et al. [70], information diffusion is defined as the proceeding that information or knowledge is disseminated so as to reach many people or

entities through interactions. It contains three essential parts: senders, receivers and medium. Besides social communication, this definition is also applicable to other propagation fields.

Information diffusion model is thus a tool for modeling such information diffusion process. There are two major methodologies for modeling: one is the epidemic model (perspective of senders); the other is decision-based model (perspective of receivers). When talking about information diffusion model, epidemic model comes to most people's mind. Epidemic model reckons information as infectious disease so that the information spreading can be described by infection process. For example, SIR (Susceptible-Infected-Recovered) model [71] and SIS (Susceptible-Infected-Susceptible) model [79] are well-known epidemic model used in analyzing social phenomena. On the other hand, decision-based model like LT (Linear Threshold) model [86] also appears in some studies.

Until now, a great number of information diffusion models in both single and multilayer networks have been proposed in solving different issues. In this section, several classical information diffusion models and other models in multilayer network will be discussed in detail.

5.1.1 Epidemic model

In epidemic model, information is diffused from the perspective of senders. That is, information spreads through the interaction between nodes of network, and these nodes as senders determine whether to distribute information to their adjacent nodes. In most case, senders only handle information around themselves in a local level. A couple of epidemic models have modeled spreading processes including SIR model, SIS model, IC model, SI1I2R model, etc.

5.1.1.1 SIR model

In Susceptible-Infected-Recovered (SIR) model [71], a node performs three types of states: Susceptible (S), Infected (I), and Recovered (R). All nodes are susceptible and not immune to the disease at first. With infection rate α $(0 \leq \alpha \leq 1)$, susceptible nodes become infected and are capable of spreading the disease to their neighbors. After time span of t, the infected nodes will recover from the disease with a recovery rate β $(0 \leq \beta \leq 1)$, having immunity from further infection. The transformation of node's state in SIR model can be illustrated through Fig. 5-1. Three colors represent three different states of nodes respectively. One node transfers from one state to the other with defined probability.

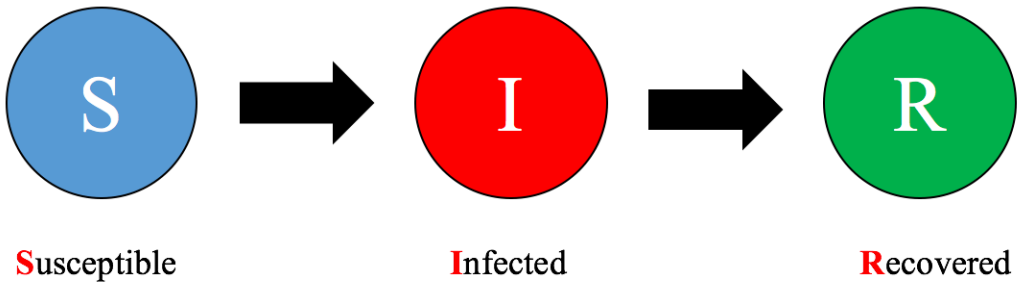


Fig. 5-1 The state of node in SIR model

An application of SIR model was proposed in order to analyze the spread of rumors happening in Great Eastern Japan Earthquake in Japan 2011 (3.11 earthquake) [101]. Once the receiver of a message becomes aware that it is a rumor, he or she will not believe the rumor again. More application case of SIR model can be also found in other references

[72], [73], [74].

5.1.1.2 SIS model

Rather than two states of nodes in SIR model, Susceptible-Infected-Susceptible (SIS) model [79] has only two types of states: Susceptible (S) and Infected (I). In other words, the recovery of nodes does not occur and infected nodes will be changed back to susceptible states with probability γ $(0 \leq \gamma \leq 1)$ after infection. Fig. 5-2 presents the transforming process of a node's state. Only two colors show two states of Susceptible and Infected. More research concerning SIS model and details can be referred in some literatures [79], [80], [81].

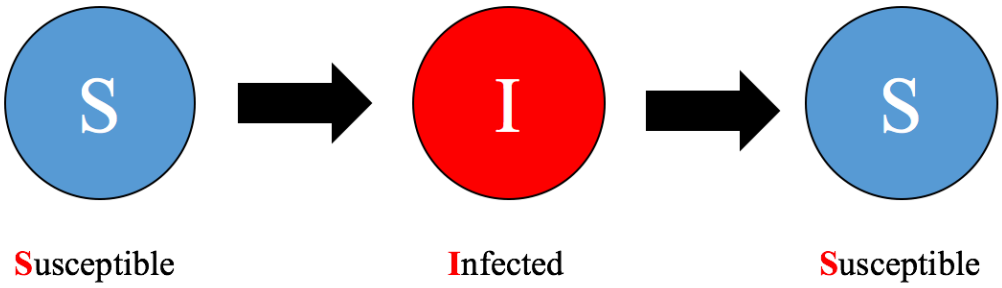


Fig. 5-2 The state of node in SIS model

5.1.1.3 Independent cascade model

SIR and SIS models are two important dynamic models in classical information diffusion research. Later on, a number of variants or extensions like SI1I2R [82], [83] have been brought forth by different researchers. The independent cascade model (IC

model), developed by Goldenberg et al.[84], is one of the most significant versions of extension.

IC model is a time-discrete variant of SIR model, where the node can exhibit two states: Active and inactive. Parameter p_{ij} $(0 \leq p_{ij} \leq 1)$ is the propagation rate of information. At the beginning, some nodes (or one node) called seeds are tagged as active nodes in advance. At time *t*, the active node *i* attempts to transmit information to one of its adjacent inactive nodes *j* with probability p_{ij}, and *i* has only once chance to activate *j*. If neighbor *j* is successfully activated at time *t*+1, it becomes active. Otherwise, node *j* remains inactive. When the active node no longer increases, spreading process terminates. Independent cascade is such a stochastic process that information is transmitted through cascade propagation. The start and termination state of IC model is shown in Fig. 5-3. The yellow nodes display the inactive nodes, while the red ones represent the active nodes.

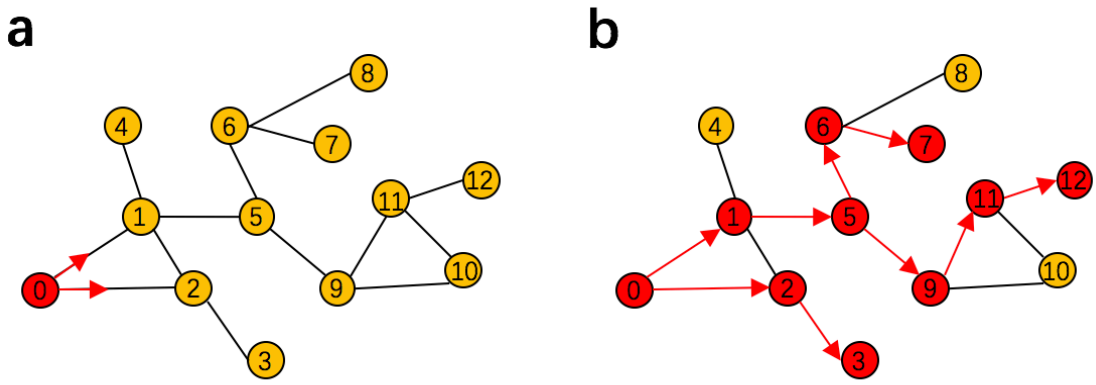


Fig. 5-3 The start and termination state of IC model

IC model is of importance in this study, because an extension of IC model based on multilayer network will be employed in latter experiments. In this section, a measure is introduced to characterize the ability of information diffusion. That is average influence

degree (AID) [97], which can be obtained from the following function:

$$\sigma = \frac{\sum_{v=0}^{N} \sigma(v)}{N} \quad \textbf{(5.1)}$$

The influence degree $\sigma(v)$ of a node *v* is defined as the expected number of active nodes when information diffusion process terminates, in which node *v* is a seed node activated at the beginning. For instance, Fig. 5-3 (a) gives information diffusion that starts from an active node 0. As the information diffusion proceeds and finally stops in a state like Fig. 5-3 (b), 10 of 13 nodes become active in the network. Therefore, $\sigma(0)$, the influence degree of node 0, is 10/13. By calculating the average of influence degree $\sigma(v)$, it leads to the result of AID.

5.1.2 Decision-based model

Decision-based model [90], also known as threshold models, is an spreading model from the perspective of receivers. Nodes decide whether to receive information depending on their neighbor [86], [87], [88], [89]. Two types of method can be found in decision-based studies including: informational and direct-benefit effects [100]. Most studies focus on the informational effects. Two of them will be discussed in this section: Linear threshold model and Watts threshold model.

5.1.2.1 Linear threshold model

Linear threshold model (LT model) [86], [91], proposed by Granovetter, is originally utilized to model the spread of influence. If node *u* and *v* are linked in LT model, they have an influence p_{uv} on each other. Let *Adj[u]* be the set of node *u*'s neighbors, then there

is a formula below to make sure the sum of influence from node u's neighbors does not exceed 1.

$$\sum_{v \in Adj[u]} p_{uv} \leq 1 \tag{5.2}$$

Further, there is a threshold θ for each node and an equation as follows:

$$\sum_{v \in AdjAct[u]} p_{uv} \geq \theta_u \tag{5.3}$$

where *AdjAct[u]* is the set of active nodes that are adjacent to node u. At each time, the inactive node u is activated when the sum of influence from its active neighbors is larger than the threshold θ. Spreading process terminates when there are no newly activated nodes. As shown in Fig. 5-4, it presents the activation process of node u in information diffusion.

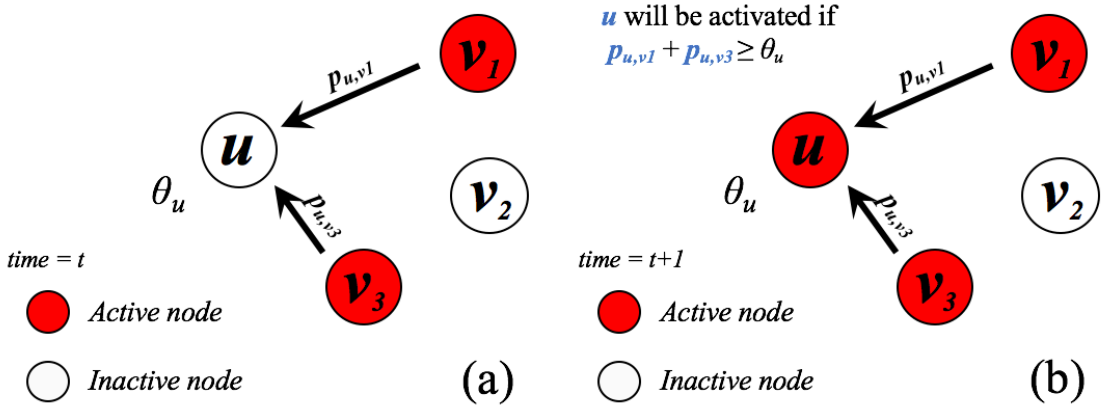


Fig. 5-4 Activation process in LT model

5.1.2.2 Watts threshold model

Watts improved linear threshold model in order to analyze global cascades on a random network [87], which is latter called Watts threshold model. He investigated the role of threshold and the relationship between cascade propagation and connectivity of the network. If the threshold fraction of neighbors who have adopted a decision satisfies a global cascades condition, a node will decide to accept it. A generating function for susceptible nodes can be employed in order to obtain such particular condition, which is described by following:

$$g_0(x) = \sum_i \rho_i p_i x^i \tag{5.4}$$

where p_i represents the probability that a node possesses degree i. Following function will lead to different values as condition of k changes, in which f gives the distribution of the threshold fraction of nodes.

$$\rho_i = \begin{cases} 1 & i = 0 \\ \int_0^{1/i} f(x)dx & i > 0 \end{cases} \tag{5.5}$$

When information diffusion happens in a sparse, random network, the cascade sizes follow a power law distribution. On the other hand, local stability of nodes plays a significant role in a dense network as global cascade propagation starts. Further, more and more heterogeneous thresholds increase the vulnerability of network to the global cascades.

5.1.3 Spreading model in multilayer networks

With the development of multilayer network analysis technology, the spreading

models discussed in previous section have been extended to multilayer version for both epidemic models and decision-based models.

For epidemic models, for instance, many studies modeled information propagation over multilayer networks such as SIR model [75], [76], [77], [78], SIS model [80], [81]. IC model was also generalized to simulate cascade propagation over a multilayer network [85]. The information diffusion across layers performs a different infection rates compared with diffusion within layers. In the literature [94], [95], [96], infection rates differ depending on the types of links.

As for Watts threshold model, some extensions have modeled spreading processes in multilayer networks [92], [93]. According to their conclusions, multilayer structure can facilitate the global cascades compared with single layer structure.

In this chapter, an extended independent cascade model considering cross-layers delay [102] is employed to carry out information spread simulation over multilayer network. The state of a node can be active or inactive. If an inactive node is successfully activated by its neighbors, it becomes active and has the ability of disseminating information to both intra-layer and across-layers neighbor nodes. The intra-layer probability is given by p_{uv} $(0 < p_{uv} \leq 1)$, whereas the cross-layers probability is defined by q_{uv} $(0 < q_{uv} \leq 1)$. Fig. 5-5 gives an illustration of information propagation process of this model where cross-layers delay is considered at each time step. As depicted in the Fig. 5-5, an across-layers propagation happens at time t_1 and t_2. the counterpart of active node in the lower layer at t_0 will be activated with probability q_{uv} at t_1. Information propagation terminates when no active nodes increase in multilayer network.

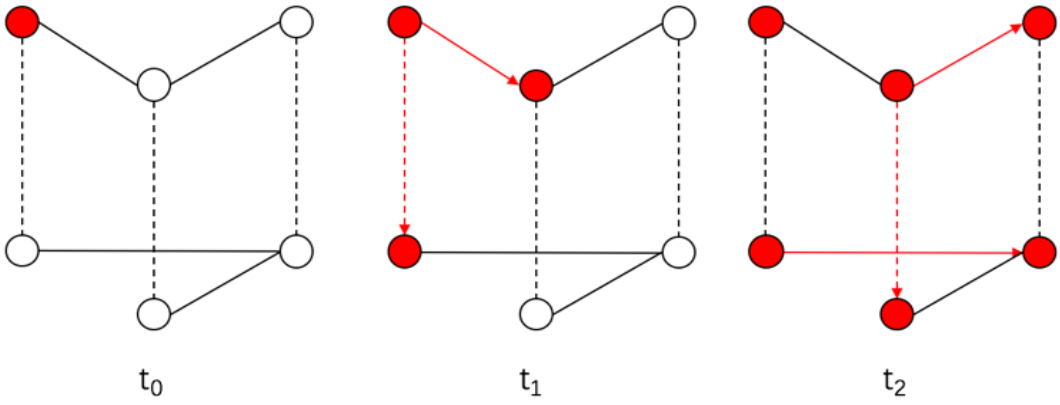


Fig. 5-5 Information diffusion with cross-layers delay

The average influence degree (AID) over multilayer network, which is called "multilayer AID" in this research, can be also calculated by formula (5.1). All nodes representing the same entity in different layers are treated as one entity node. If one of nodes describing the same entity is activated at the end of simulation, the entire entity node will be tagged as active. Hence, the AID for multilayer network is the average of all entity nodes' infection degree.

5.2 Simulation of information diffusion and results

In this section, the difference of AID between multilayer network and aggregated single network is investigated. The former AID calculated from information spreading over multilayer network is called "multilayer AID", whereas the latter AID for single network is called "singular AID". Singular AIDs are often the cases which can be seen from real-world data of multilayer network. Nevertheless, they cannot describe realistic information diffusion over different types of links in real interactions, because most real networks have multilayer structures. As a matter of fact, spreading model over multilayer

network introduced in previous section and multilayer AID is capable of capturing more accurate information propagation in real-world systems. Therefore, the objective of this section is to synthesize multilayer networks according to network features of aggregated real multilayer networks and calculate their multilayer AIDs. By investigating the difference of multilayer AID between original and produced multilayer networks, we can see whether proposed multilayer model can reproduce more realistic information diffusion occurring on data in real life, rather than a simple information propagating process without a consideration of multilayer structure.

This section introduces two real multilayer networks used in information diffusion experiments at first. Then a "OAR" process is presented to explain the procedure of evaluation. Finally, some experimental results will be shown to denote the accuracy and effectiveness of proposed multilayer network model.

5.2.1 Real multilayer network datasets

Two real multilayer networks are included with the purpose of evaluating realistic information diffusion. One is a social network with multilayer structure and the other is a multiplex transport network. Both of them can be downloaded freely from web dataset[*1].

5.2.1.1 CKM physicians' innovation network

This CKM physicians' innovation network (Physicians) is a social multilayer network collected by Coleman, Katz and Menzel on medical innovation [103], involving physicians from four towns in Illinois, Peoria, Bloomington, Quincy and Galesburg. As

[*1] http://deim.urv.cat/~manlio.dedomenico/data.php

displayed in Table. 5-1, this dataset[*1] contains 246 nodes and 1,551 links, as well as 3 layers. Layers are built through three socio-metric matrices. The nodes were linked with the influence of network ties on the physicians' adoption of a new drug, tetracycline. When carrying out information propagating simulation, influence of medical innovation flows through different layers of the multilayer network, resulting in an influence spreading network.

Table. 5-1 Basic features of Physicians' innovation network

Network	Nodes' number	Links' number	Layers' number
Physicians	246	1,551	3

5.2.1.2 London multiplex transport network

The data of London multiplex transport network was acquired in 2013 from the official website of Transport for London[*1] [104]. In this multilayer network, nodes and links denote train stations and existing routes between stations in London. Basic information of this network includes 369 nodes, 441 links, and 3 layers (cf. Table. 5-2). Three layers correspond underground line, Over-ground line, and DLR stations' line. If an epidemic disease breaks out in this metropolitan city, the disease spreading through different traffic tools can be represented by the information diffusion over the London multiplex transport networks.

[*1] https://www.tfl.gov.uk/

Table. 5-2 Basic features of London multiplex transport network

Network	Nodes' number	Links' number	Layers' number
London transport	369	441	3

5.2.2 The "OAR" process

In order to evaluate the accuracy of information diffusion over multilayer network, an "Original-Aggregated-Reconstructed (OAR)" process is proposed. It can be represented through three steps:

(1) O step: An original multilayer network is constructed from real data or artificial data from simulation (cf. Fig. 5-6);

(2) A step: Aggregating all layers into a monolayer network (cf. Fig. 5-7);

(3) R step: Reconstructing multilayer structure from the monolayer network. (cf. Fig. 5-8)

5.2.2.1 O step

The first "O" step expresses the construction of an original multilayer network, which is a multilayer network from real dataset in this research. As shown in Fig. 5-6, it is an instance of original multilayer network that will be used for further information diffusion. Moreover, simulation experiments of information spread over this multilayer network are carried out and their average influence degrees (AIDs) are calculated as baseline.

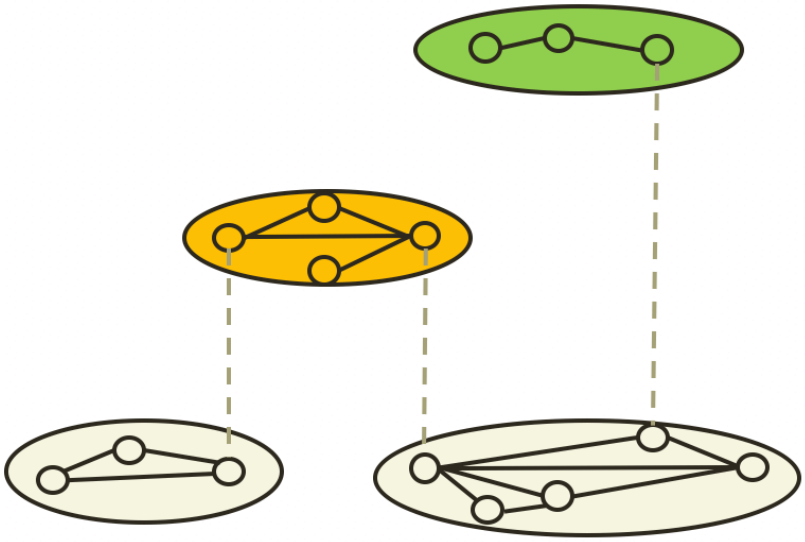


Fig. 5-6 Original multilayer network from real data

5.2.2.2 A step

The second “A” step represents the aggregation process of original multilayer network. All layers of a multilayer network are aggregated into one layer and the AID of information diffusion over monolayer network is calculated later. This process of aggregation can be illustrated in Fig. 5-7.

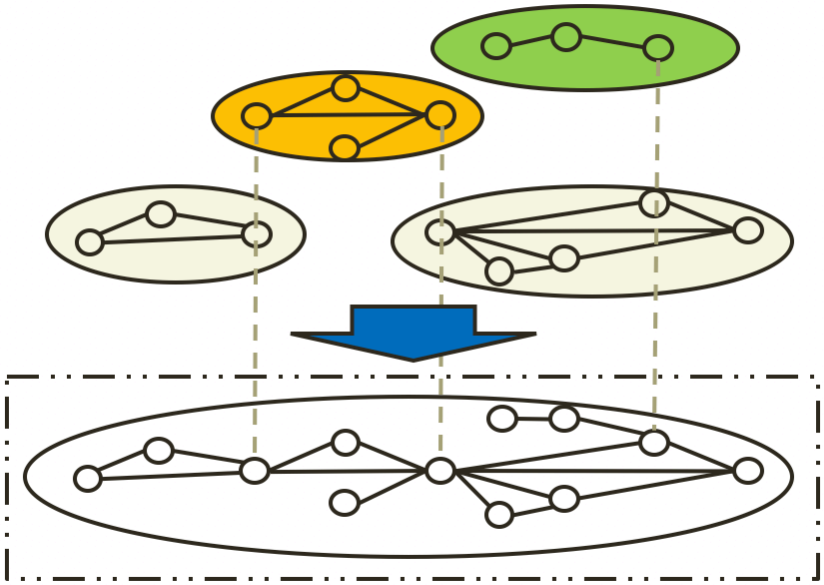


Fig. 5-7 Aggregation into a monolayer network

5.2.2.3 R step

The final “R” means the reconstruction of a multilayer network with multilayer model. As depicted in Fig. 5-8, the multilayer structure is rebuilt by proposed multilayer model taking fully use of the network features of aggregated single layer in “A” step. Another experiments of information propagations are conducted over the reconstructed multilayer network as well as the computation of multilayer AIDs. By comparing the AID of rebuilt multilayer network with baseline AID of the original multilayer network, it helps to evaluate the accuracy of information diffusion through multilayer network. If AID in the third “R” step resembles the baseline AID in the first “O” step, it can be concluded that proposed multilayer method can reproduce more realistic information diffusion than a singular layer method.

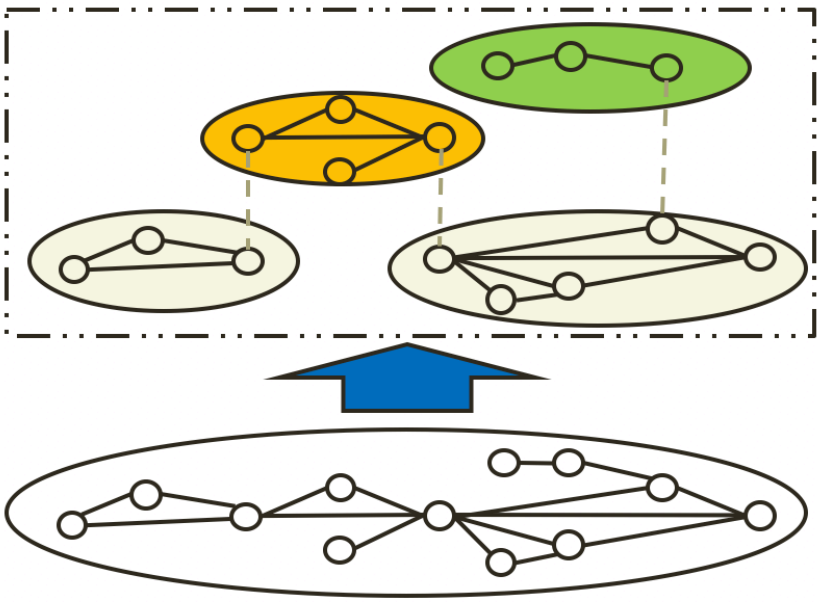


Fig. 5-8 Reconstruction of multilayer from monolayer network

5.2.3 Experimental results of information diffusion

Two real multilayer networks, CKM physicians' innovation network and London multiplex transport network, appear to be employed in "O" step. Later on, aggregation of multilayer networks was done in "A" step. Here, IC model and its extended version in section 5.1.3 were implemented to simulate information spreading process. AID of both situations, multilayer AID and singular AID, were calculated and shown in Table. 5-3. In this research, information propagation over multilayer network of real system is reckoned as realistic spreading process, thus the value of multilayer AID is regarded as ground truth for evaluating information diffusion of rebuilt multilayer network. As described in Table.

5-3, AID of multilayer network has a smaller value than single network due to the cross-layers delay for both cases.

Table. 5-3 AIDs of two real networks

	Physicians' innovation	London transport
Multilayer AID	0.3005	0.0201
Singular AID	0.3113	0.0234

The reconstruction of a multilayer network was called "*R*" step. In order to reproduce information diffusion of original real multilayer network, network features of aggregated networks were utilized by proposed multilayer models. This multilayer networks are created with same nodes and links as real multilayer networks, having the number of layers between 2 and 10. Each multilayer network is created with distinct inner model, which is a combination of five different models. This model synthesizes an artificial multilayer network by minimizing the distance *D* in Function (3.8) between aggregated original network and produced network. Further, information spreading is carried out over reconstructed multilayer network and the multilayer AID is obtained by calculation. The results are shown in Table. 5-4 as a comparison with multilayer and singular AID for both set of real networks.

Table. 5-4 Comparison of AIDs between real data and reconstructed network

	Physicians' innovation	London transport
Multilayer AID	0.3005	0.0201
Singular AID	0.3113	0.0234
Multilayer AID of rebuilt network	**0.8615**	**0.0095**

Table. 5-5 Distances for physicians' innovation network

(Horizontal axis: models in large-sized communities)

Distance	ER	BA	WS	CNN
ER	3.4802	0.2461	0.4570	1.6260
BA	2.1196	**0.0873**	0.4684	0.6240
WS	2.5909	0.1838	0.6819	0.8037
CNN	1.9235	0.1091	1.8230	0.3732
CPN	3.1885	0.2655	2.7543	5.5455

Table. 5-6 Multilayer AIDs for physicians' innovation network

(Horizontal axis: models in large-sized communities)

AID	ER	BA	WS	CNN
ER	0.7565	0.2484	0.4227	0.6371
BA	0.8981	**0.8615**	0.3337	0.7408
WS	0.8353	0.9211	0.4519	0.7470
CNN	0.7981	0.8835	0.7679	0.7552
CPN	0.7353	0.7105	0.3705	0.3828

As presented in Table. 5-4, reconstructed multilayer networks cannot reproduce accurate information diffusions for two real networks, since the value of 0.8615 and 0.0095 deviate from original multilayer AIDs and singular AIDs.

Table. 5-7 Distances for London transport network

(Horizontal axis: models in large-sized communities)

Distance	ER	BA	WS	CNN
ER	0.6769	0.9971	0.1027	2.4530
BA	0.7486	1.0376	0.9303	3.0459
WS	0.7398	1.4120	**0.0488**	3.8120
CNN	0.4249	1.1534	0.5936	2.3154
CPN	4.7062	1.2113	3.2626	5.4341

Table. 5-8 Multilayer AIDs for London transport network

(Horizontal axis: models in large-sized communities)

AID	ER	BA	WS	CNN
ER	0.0484	0.0625	0.0096	0.0468
BA	0.1529	0.0626	0.0081	0.0407
WS	0.0480	0.0667	**0.0095**	0.0457
CNN	0.0627	0.0674	0.0091	0.0321
CPN	0.0468	0.0500	0.0106	0.0365

Table. 5-5 and

Table. 5-7 give minimum distances for all possible inner model combinations in detail, whereas Table. 5-6 and Table. 5-8 depict corresponding multilayer AIDs. Both two set of multilayer AIDs have a serious deviation from baseline multilayer AIDs. The

reason is that minimizing distance of network features, such as clustering coefficient, modularity, etc., is not enough to characterize all topological structure features of original multilayer network, especially information propagation. Therefore, other properties are necessary to be introduced in order to improve the method of optimization.

In this research, singular AID of aggregated network is utilized to obtain the best rebuilt multilayer network. The detail process is illustrated through the following steps:

First, distance *D* in Function (3.8) is calculated. Instead of minimizing distance *D*, 10 best rebuilt multilayer networks are selected for each inner model combination.

Second, the differences of singular AID between aggregated real networks and aggregated constructed multilayer networks are calculated.

Finally, best multilayer network with the minimum difference of singular AID is chosen from 10 rebuilt networks.

Table. 5-9 Comparison of AIDs between real data and reconstructed network with new method

	Physicians' innovation	London transport
Multilayer AID	0.3005	0.0201
Singular AID	0.3113	0.0234
Multilayer AID of rebuilt network	**0.3020**	**0.0224**

After selection of best multilayer network, simulation of information diffusion is enforced and the multilayer AIDs of reconstructed network is acquired by using formula. This result can be presented in Table. 5-9.

Table. 5-10 Distances of singular AIDs for physicians' innovation network

(Horizontal axis: models in large-sized communities)

Distance	ER	BA	WS	CNN
ER	0.0048	0.0045	**0.0001**	0.2781
BA	0.4099	0.0416	0.0020	0.2687
WS	0.2652	0.0607	0.2575	0.0214
CNN	0.0506	0.0964	0.0831	0.2262
CPN	0.3417	0.0662	0.1022	0.1784

Table. 5-11 AIDs of rebuilt multilayer physicians' innovation network

(Horizontal axis: models in large-sized communities)

AID	ER	BA	WS	CNN
ER	0.3067	0.2584	**0.3020**	0.5565
BA	0.6429	0.2196	0.3121	0.5264
WS	0.4534	0.3189	0.4583	0.2320
CNN	0.2031	0.3396	0.3171	0.4933
CPN	0.5921	0.3069	0.3259	0.4862

Furthermore, the Table. 5-10 and Table. 5-12 displayed the smallest distance for every inner model combination. The red figures give the smallest distance and denote best combinations. Table. 5-11 and Table. 5-13 described the corresponding AID of rebuilt multilayer network.

For both physicians' innovation network and London multiplex transport network, AIDs of reconstructed multilayer networks in "R" step not only have a smaller value than single layer in "A" step, but are closer to original real multilayer network in "O" step.

Table. 5-12 Distances of singular AIDs for London transport network

(Horizontal axis: models in large-sized communities)

Distance	ER	BA	WS	CNN
ER	0.0495	0.0232	0.0077	0.0228
BA	0.0072	0.0154	0.0042	0.0148
WS	0.0510	0.0202	0.0003	0.0045
CNN	0.0160	0.0254	0.0011	0.0230
CPN	0.0521	**0.0002**	0.0030	0.0106

Table. 5-13 AIDs of rebuilt London multiplex transport network

(Horizontal axis: models in large-sized communities)

AID	ER	BA	WS	CNN
ER	0.0594	0.0314	0.0114	0.0410
BA	0.0305	0.0283	0.0129	0.0314
WS	0.0655	0.0371	0.0132	0.0241
CNN	0.0155	0.0322	0.0146	0.0343
CPN	0.0519	**0.0224**	0.0191	0.0250

According to Table. 5-10 and

Table. 5-5, it can be observed that WS+ER as inner model produced smallest

distance of 0.0001. Moreover, the AID of rebuilt multilayer network with WS+ER combination is 0.3020, which is less than 0.3113 (the singular AID of aggregated real network) and closest to 0.3005 (the multilayer AID of original real network). Accordingly, more realistic information diffusion can be inferred by applying proposed method and minimizing distance of singular AIDs for CKM physicians' innovation network.

As shown in Table. 5-12 and Table. 5-13, the smallest distance between singular AIDs of aggregated real networks and aggregated constructed networks is 0.0002, and BA+CPN as inner model can produce an AID of multilayer with the value 0.0224, which is smaller than singular AID value of real data (0.0234) and closer to original multilayer AID (0.0201). As the multilayer AID of produced network resembles original multilayer AID of real data compared with singular AID, it can be concluded that proposed method can reproduce realistic information propagation for London transport network.

Therefore, experimental results show that the proposed multilayer model is capable of reconstructing a multilayer network which can perform a relatively realistic information diffusion property for different types of real networks.

5.3 Conclusion

This chapter utilized the proposed multilayer model to carry out spreading simulation over multilayer network and calculated the average influence degree (AID). It succeeded in verifying the effectiveness of multilayer model in reproducing realistic information diffusion.

Section 5.1 briefly introduced several classical information diffusion models, such as epidemic model and decision-based model. In epidemic model, agents of the network spread information through their neighbor to other agents. Three classical model, SIR,

SIS, and Independent Cascade (IC) model, are well-known epidemic models. On the other hand, decision-based model is derived from the idea that agents decide whether to receive information depending on their local neighbor. Two of the most important models, linear threshold model and Watts threshold model, were described in detail in this parts.

Section 5.2 discussed the experiments of information propagation simulation in detail. Real multilayer networks used for simulation were talked in 5.2.1. Next part of 5.2.2 expressed the “OAR” process and calculated AID of information diffusion for two multilayer networks and one aggregated network. In 5.2.3, real multilayer networks are employed to evaluate proposed multilayer model and corresponding experimental results were shown. According to the results of comparison, it comes to a conclusion that the proposed multilayer model successfully reproduced realistic information spreading over multilayer network.

Chapter 6. Conclusion

6.1 Discussion and conclusion

This research attempts to explore the mechanism of network generation based on multilayer network and clarify the relationship between phenomena occurring on the network and the network structure. Many networks observed in real life have community structure, or high-modularity property. Existing studies focusing on the single network, however, cannot produce high-modularity networks with high precision. Considering the multilayer property of many real systems, multilayer network is introduced to model network generation.

Presently, this research has made the following contributions:

(1) Proposed a high-modularity network generation model by layer aggregation based on the multilayer network to produce artificial networks that resemble real networks;

(2) Estimated the hidden structures in real networks by investigating how artificial networks are constructed from generation model;

(3) Reproduced realistic information diffusion over multilayer network by enforcing spreading simulation.

To begin with, the developed multilayer model is a generation model that can reproduce networks in real systems. The core property of created networks is high-modularity, or community structure. In order to construct a high-modularity network, proposed model encompasses four essential parts: community size distribution, number

of layers, inner community network model, and inter-layer degree correlation. They control the topological structure of produced networks.

Specifically, the community size follows a power-law distribution with exponent β. For number of layers, its value ranges between 2 and 10. In inner communities, ER, BA, WS, CNN, a complete network (CPN), and their combinations were chosen to build links within nodes inside a community. Moreover, the inter-layer degree correlation was selected by random or positive correlation.

As for measurements, a couple of network features are chosen to measure the similarity or distance between the produced networks and target real networks, such as clustering coefficient C, assortativity r, modularity Q, power index of degree distribution γ, and coefficient of determination of degree distribution R^2. Employing five representative network features, this research quantitatively evaluates the distance of two networks by a normalized Euclidean distance D.

For the purpose of simulating network generation of real world, four real networks of different types were given in this research, including the Facebook network (Yale University), the Renren network (Peking University), a collaboration network (a scientific co-authorship network of Arxiv general relativity), and an air traffic control network (preferred routes database).

To clarify whether the proposed model can properly reproduce real networks, three existing models for constructing real networks are implemented as baseline: Lancichinetti-Fortunato-Radicchi (LFR) benchmark, Kronecker graph model, and Pasta's model. LFR benchmark model synthesizes networks with planted community structures with some strong constraints. Kronecker graph employs a self-similar nesting process to produce a hierarchical structure. Pasta's model is a tunable and growing network

generation model that ensures 3 properties of communities: internal structure, power-law degree distribution, and high clustering coefficients. These baseline models can produce networks with community structures.

In simulations, multilayer-based model combined all of the parameter values to mimic four real networks. A comparison of proposed model and baseline model was drawn so as to evaluate the accuracy of proposed model. The experimental results showed that proposed model outperformed other baseline in reproducing real networks with community structure.

Furthermore, this research successfully estimated hidden structures of networks in real system, which cannot be completed by other existing models. For instance, the comparison results show that the dataset in Facebook (Yale University) is estimated to have a high number of layers (about 8 to 10 layers), whereas the Renren dataset (Peking University) has a relatively low value (mostly between 3 and 4). That is to say, Yale has more kinds of relationships among students than Peking University due to the diversity of students' source. Other two real network data do not exhibit an obvious tendency of number of layers. In the scientific co-authorship network, researchers who co-author a paper tend to form a small community with a complete network. On the contrary, the features of larger communities can be captured by CNN model. Hence, the generation model with CPN plus the CNN model as an inner community model obtained networks that resemble the real-world data most. For the air traffic control network, a combination of the ER and BA models reflects the existence of many hub airports and other smaller airports are randomly connected.

Finally, this research employed the proposed multilayer model to simulate information propagation over multilayer networks and calculated the average influence

degree (AID) of information diffusion. Both singular AIDs and multilayer AIDs were calculated to make a comparison for two types of real multilayer networks: CKM physicians' innovation network and London multiplex transport network. By minimizing the distance based on network features and singular AIDs, this research can infer multilayer AIDs of original multilayer networks. Simulation results discovered that proposed model can reproduce more realistic information diffusion than the single network by enforcing spreading simulation over multilayer networks.

As discussed above, this research proposed a new high-modularity network generation model based on multilayer network. Not only did it synthesize high-modularity networks that resemble real world data, but it reproduced realistic information propagation as well. Moreover, mechanism of network generation can be uncovered by employing this multilayer-based model, thus it helps to the understanding of hidden structure in real networks

6.2 Future work

Even though the high-modularity network generation model proposed by layer aggregation based on multilayer reproduced the target real networks fairly well, several disadvantages still remain unsolved. Some prospective improvements or future studies are presented as follows.

Firstly, current algorithm of network generation model completely searches all possible values of parameters by minimizing the normalized Euclidean distance. As the full search is time-consuming, future work will consider optimization methods to accelerate this process.

Secondly, even though the number of nodes (links) used in this research can be

extended to a thousand (hundred thousand) level, it is worthy of analyzing real networks on a larger scale.

Thirdly, improvement of evaluation methodology will be considered. Since weights of the evaluation function are not effectively used, future studies will place emphasis on some specific features by adding weights. Also, other network features may also be easily introduced into the evaluation function.

Finally, the average influence degree (AID) was used to evaluate the ability of information diffusion. However, other measurements of information spreading process still exist, such as cascade velocity, transmissibility probability, etc. Analysis of such various measurements will also be research objective in the future.

Acknowledgements

I would like to express my deepest gratitude to my dissertation supervisor, Associate Professor Fujio Toriumi, who gave me an opportunity to pursue my Ph.D. dream after I had received the MS degree. During the last three and a half years, he has been always guiding me with patience and encouragement. From research meetings to private conversations, his continuous support and constructive criticisms help me face a number of research difficulties and make it through, which will be a precious treasure to me for a lifetime. Without his inspiration and constant supervision, I cannot finally complete this dissertation.

I would like to extend my greatest gratitude to Professor Hirotada Ohashi, who often offered me many enlightening suggestions in research meeting. His patient guidance and kindly encouragement will never be forgotten in my future life. Meanwhile, the study and research in Ohashi-Toriumi laboratory is an amazing experience for me to understand myself and think about future, so I really want to send my deep thanks to Professor Hirotada Ohashi.

I should also sincerely thank other reviewers of this dissertation, Professor Kiyoshi Izumi, Professor Kazuhiro Aoyama, and Associate Professor Yu Chen, who are taking charge of reviewing this research work and offered many insightful comments. I want to extend my grateful thanks to all of them.

My thanks should also be extended to all members and graduates of Ohashi-Toriumi laboratory. For graduates, I am deeply grateful to Dr. Kimitaka Asatani, who had always showed many concern on my research and helped me handle some complicated problems. Thank him very much for his help and encouragement. I express sincere appreciation to

the help from Dr. Shohei Usui, who often discussed research problems with me and gave me several valuable advices. I also want to extend my thanks to other graduates, such as Yu Matsuzawa, Motoka Motoka, Kazuki Komura, Hidenori Mabe, Kenta Koyama, Kenta Kaihoko, Seigo Baba, Kengo Kajiwara, Keigo Yamamoto, Kouichi Hattori, Kota Honda, Yuka Kamiko, Hirotaka Kawazu, Yating Wang, Tomoya Maruyama, Kei Murakami, etc. In particular, I thank Hidenori Mabe, Yu Matsuzawa, Keigo Yamamoto, Seigo Baba, Yuka Kamiko, and Yating Wang for their kind help and consideration in research field or laboratory life.

For current members of laboratory, I would like to thank Project Researcher Miki Hirabayashi, who is really kind and always shares souvenir of tasty food like “Okashi” with us. I am thankful to the Ph.D candidate Takashi Nicholas Maeda for his useful comments and feedbacks of research and Japanese manner. I should also thank other members, like Yuichi Hirano, Kazuki Uchida, Tomoki Fukuma, Takuro Yamazaki, Yuki Kawasaki, Yui Murayama, Jianwei Xie, Kota Kakiuchi, Atomu Sonoda, Naoto Kajimura, Ko Kudo, Nina Sugiyama, and Shota Chikaraishi. I also want to extend my thanks to Ms Yumiko Nagoh, the secretary of our laboratory, who often actively contact me and help me prepare many written materials. I am really thankful to all the members at Ohashi-Toriumi laboratory.

I also warmly appreciate the support of Japanese Government (MEXT) Scholarship. The financial supports from MEXT enabled me to continue my Ph.D. research and life in Japan for three and a half years.

Finally, and certainly not the least, I would like to express my gratitude to the family, particularly my parents. Without your unconditional love, support, encouragement, and belief in me for so many years, I would not have the courage to study abroad and pursue

a Ph.D. degree at the University of Tokyo. Because of you, I became the person I am today.

November, 2017

Chao FAN

References

[1] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. Why rumors spread so quickly in social networks. *Commun. ACM*, Vol. 55, No. 6, pp. 70-75, 2012.

[2] Jie Zhou, Zonghua Liu, and Baowen Li. Influence of network structure on rumor propagation. *Physics Letters A*, Vol. 368, No. 6, pp. 458-463, 2007.

[3] Damián H. Zanette. Critical behavior of propagation on small-world networks. *Physical Review E*, Vol. 64, p. 050901, 2001.

[4] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, Vol. 406, No. 6794, pp. 378-382, 2000.

[5] Paolo Crucittia, Vito Latorab, Massimo Marchiori, and Andrea Rapisarda. Error and attack tolerance of complex networks. *Physica A: Statistical Mechanics and its Applications*, Vol. 340, No. 1, pp. 388-394, 2004.

[6] Mahdi Jalili. Error and attack tolerance of small-worldness in complex networks. *Journal of Informetrics*, Vol. 5, No. 3, pp. 422-430, 2011.

[7] Yating Wang and Fujio Toriumi. Analysis of group behavior bias in Financial Markets using artificial market, *The Japanese Society for Artificial Intelligence Interest Group on Financial Infomatics*, Vol. 18. 2017.

[8] Hazem Krichene and Mhamed-Ali El-Aroui. Artificial stock markets with different maturity levels: simulation of information asymmetry and herd behavior using agent-based and network models. *Journal of Economic Interaction and Coordination*, pp. 1-25, 2017.

[9] Martin A. Nowak and Robert M. May. Evolutionary games and spatial chaos. *Nature*, Vol. 359, p. 826, 1992.

[10] Zhihai Rong, Xiang Li, and Xiaofan Wang. Roles of mixing patterns in cooperation on a scale-free networked game. *Physical Review E*, Vol. 76, p. 027101, 2007.

[11] Salvatore Assenza, Jesús Gómez-Gardeñes, and Vito Latora. Enhancement of cooperation in highly clustered scale-free network. *Physical Review E*, Vol. 78, p. 017101, 2008.

[12] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, Vol. 74, No. 1, 47, 2002.

[13] Paul Erdős and Alfréd Rényi. On Random Graphs I. *Publicationes Mathematicae*, *Debrecen*, 6, pp. 290-297, 1959.

[14] Béla Bollobás. Random Graphs (2nd ed.). *Cambridge University Press*, pp. 180-181, 2001.

[15] Edgar N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, Vol. 30, No. 4, pp. 1141-1144, 1959.

[16] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, Vol. 393, No. 6684, pp. 440-442, 1998.

[17] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, Vol. 286, No. 5439, pp. 509-512, 1999.

[18] Alexei Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, Vol. 67, No. 5, p. 056104, 2003.

[19] Mikko Kivelä, Alexandre Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, Vol. 2, No. 3, pp. 203-271, 2014.

[20] Alessio Cardillo, et al. Emergence of network features from multiplexity. *Scientific Reports*, Vol. 3, 2013.

[21] Michael Szell, Renaud Lambiotte, and Stefan Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, Vol. 107, No. 31, pp. 13636-13641, 2010.

[22] Vincenzo Nicosia, et al. Growing Multiplex Networks. *Physical Review Letters*, Vol. 111, No. 5, p. 058701, 2013.

[23] Ginestra Bianconi. Statistical mechanics of multiplex networks: Entropy and overlap. *Physical Review E*, Vol. 87, No. 6, p. 062806, 2013.

[24] Anna Saumell-Mendiola, M. Ángeles Serrano, and Marián Boguná. Epidemic spreading on interconnected networks. *Physical Review E*, Vol. 86, No. 2, p. 026106, 2012.

[25] Emanuele Cozzo, et al. Contact-based social contagion in multiplex networks. *Physical Review E*, Vol. 88, p. 050801, 2013.

[26] Byungjoon Min, et al. Layer-switching cost and optimality in information spreading on multiplex networks. *Scientific Reports*, Vol. 6, 2016.

[27] Jesús Gómez-Gardeñes, et al. Evolution of Cooperation in Multiplex Networks. *Scientific Reports*, Vol. 2, 2012.

[28] Vincenzo Nicosia, et al. Remote Synchronization Reveals Network Symmetries and Functional Modules, *Physical Review Letters*, Vol. 110, No. 17, p. 174102, 2013.

[29] Sergio Gómez, et al. Diffusion Dynamics on Multiplex Networks. *Physical Review Letters*, Vol. 110, No. 2, p. 028701, 2013.

[30] Manlio De Domenico, et al. Random Walks on Multiplex Networks. *PNAS*, Vol. 111, 8351, 2014.

[31] Haifeng Zhang, Jie Zhang, Changsong Zhou, Michael Small, and Binghong Wang. Hub nodes inhibit the outbreak of epidemic under voluntary vaccination. *New Journal of Physics*, Vol. 12, No. 2, p. 023015, 2010.

[32] Zike Zhang, Chuxu Zhang, Xiaopu Han, and Chuang Liu. Emergence of blind areas in information spreading. *PloS one*, Vol. 9, No. 4, pp. 1-7, 2014.

[33] Erlend Nier, Jing Yang, Tanju Yorulmazer, and Amadeo Alentorn. Network models and financial stability. *Journal of Economic Dynamics and Control*, Vol. 31, No. 6, pp. 2033-2060, 2007.

[34] Stefano Battiston, Domenico Delli Gatti, Mauro Gallegati, Bruce Greenwald, and Joseph E. Stiglitz. Default cascades: When does risk diversification increase stability? *Journal of Financial Stability*, Vol. 8, No. 3, pp. 138-149, 2012.

[35] Robert M. May and Nimalan Arinaminpathy. Systemic risk: the dynamics of model banking systems. *Journal of the Royal Society Interface*, Vol. 7, No. 46, pp. 823-838, 2010.

[36] Jeffrey Travers and Stanley Milgram. The small world problem. *Phychology Today*, Vol. 1, pp. 61-67, 1967.

[37] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, Vol. 65, No. 2, p. 026107, 2002.

[38] Konstantin Klemm and Victor M. Eguiluz. Growing scale-free networks with small-world behavior. *Physical Review E*, Vol. 65, No. 5, p. 057102, 2002.

[39] Mark EJ Newman, Duncan J. Watts and Steven H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, Vol. 99, suppl 1, pp. 2566-2572, 2002.

[40] Jianwei Wang, and Lili Rong. Evolving small-world networks based on the modified ba model. *In Computer Science and Information Technology, 2008 (ICCSIT'08), International Conference on. IEEE*, pp. 143-146, 2008.

[41] Michele Catanzaro, Guido Caldarelli and Luciano Pietronero. Assortative model for social networks. *Physical Review E*, Vol. 70, No. 3, p. 037101, 2004.

[42] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, Vol. 99, Vol. 12, pp. 7821-7826, 2002.

[43] Aaron Clauset, Mark EJ Newman and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, Vol. 70, No. 6, p. 066111, 2004.

[44] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, Vol. 69, No. 2, p. 026113, 2004.

[45] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, Vol. 103, No. 23, pp. 8577-8582, 2006.

[46] Jiyang Chen. Community Mining: Discovering Communities in Social Networks. University of Alberta, Edmonton, Alberta, 2010.

[47] Chao Fan and Jingsong Yu. Finding community structure in social network of renren. *ICIC Express Letters*, Vol. 7, No. 5, pp. 1693-1698, 2013.

[48] Andrea Lancichinetti, Santo Fortunato and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, Vol. 78, No. 4, p. 046110, 2008.

[49] Gergely Palla, Imre Derényi, Illés Farkas and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, Vol. 435, No. 7043, pp. 814-818, 2005.

[50] Roger Guimerà Manrique, Leon Danon, Albert Díaz-Guilera, Francesc Giralt and Alex Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, Vol. 68, No. 6, p. 065103, 2003.

[51] Leon Danon, Jordi Duch, Alex Arenas and Albert Díaz-Guilera. Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science. *World Scientific*, Singapore, pp. 93-114, 2007.

[52] Reuven Cohen, Keren Erez, Daniel Ben-Avraham and Shlomo Havlin. Resilience of the internet to random breakdowns. *Physical Review Letters*, Vol. 85, No. 21, p. 4626, 2000.

[53] Pastor-Satorras, Romualdo and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, Vol. 86, No. 14, p. 3200, 2001.

[54] Piotr Bródka. A method for group extraction and analysis in multi-layered social networks. *Ph.D. disertation*, Wrocław, pp. 88-94, 2012.

[55] Erzsébet Ravasz, Anna L Somera, D. A. Mongru, Zoltan N. Oltvai and Albert-László Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, Vol. 297, No. 5586, pp. 1551-1555, 2002.

[56] Erzsébet Ravasz and Albert-László Barabási. Hierarchical organization in complex networks. *Physical Review E*, Vol. 67, No. 2, p. 026112, 2003.

[57] Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, Vol. 11, Feb., pp. 985-1042, 2010.

[58] Muhammad Qasim Pasta, Zohaib Jan, Arnaud Sallaberry and Faraz Zaidi. Tunable and growing network generation model with community structures. *3rd International Conference on Cloud and Green Computing (CGC) IEEE*, pp. 233-240, 2013.

[59] Shuqin Zhang. Hierarchical modular structure identification with its applications in gene coexpression networks. *The Scientific World Journal 2012*, 2012.

[60] Steven L. Johnson, Samer Faraj and Srinivas Kudaravalli. Emergence of power laws

in online communities: The role of social mechanisms and preferential attachment. *Mis Quarterly*, Vol. 38, No. 3, pp. 795-808, 2014.

[61] Manlio De Domenico, Vincenzo Nicosia, Alexandre Arenas and Vito Latora. Structural reducibility of multilayer networks. *Nature Communications*, Vol. 6, p. 6864, 2015.

[62] Mark EJ Newman. Assortative mixing in networks. *Physical Review Letters*. Vol. 89, No. 20, p. 208701, 2002.

[63] Mark EJ Newman. Mixing patterns in networks, *Physical Review E*, Vol.67, No.2, p. 026126, 2003.

[64] Sid Redner. Networks: Teasing out the missing links. *Nature*, Vol. 453, No. 7191, pp. 47-48, 2008.

[65] Santo Fortunato and Darko Hric, Community detection in networks: A user guide. *Physics Reports*, Vol. 659, pp. 1-44, 2016.

[66] Fujio Toriumi, Ken Ishida and Kenichiro Ishii. Proposal for a network growth model of social network service. *The Institute of Electronics, Information and Communication Engineers*, Vol. J93-D, No.7, pp. 1135-1143, 2010.

[67] Shohei Usui, Fujio Toriumi, Masato Matsuo, Takatsugu Hirayama and Kenji Mase. Greedy network growth model of social network service. *Computational Intelligence and Intelligent Informatics*, Vol. 18, No. 4, pp. 590-597, 2014.

[68] Roger Guimera, Marta Sales-Pardo and Luís A. Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, Vol. 70, Vol. 2, p. 025101, 2004.

[69] Judith Molka-Danielsen, Matthias Trier, Vadim Shlyk, Annette Bobrik, Markku I. Nurminen. IRIS (1978-2006) Historical reflection through visual analysis. *Proceedings of the 30th Information Systems Research Seminar in Scandinavia IRIS*, p. 2, 2007.

[70] Reza Zafarani, Mohammad Ali Abbasi and Huan Liu. Social Media Mining: An Introduction. *Cambridge University Press*, New York, NY, USA, 2014.

[71] William Ogilvy Kermack and Anderson Gray McKendrick. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London A: Mathematical, *Physical and Engineering Sciences*, Vol. 115, No. 772, pp. 700-721, 1927.

[72] Mark Dickison, Shlomo Havlin, and H. Eugene Stanley. Epidemics on interconnected networks. *Physical Review E*, Vol. 85, No. 6, p. 066109, 2012.

[73] Dajun Qian, Osman Yağan, Lei Yang and Junshan Zhang. Diffusion of real-time information in social-physical networks. *IEEE Global Communications Conference*, pp. 2072–2077, 2012.

[74] Dawei Zhao, Lixiang Li, Haipeng Peng, Qun Luo and Yixian Yang. Multiple routes transmitted epidemics on multiplex networks. *Physics Letters A*, Vol. 378, No. 10, pp.

770–776, 2014.

[75] Vincent Marceau, Pierre-André Noël, Laurent Hébert-Dufresne, Antoine Allard and Louis J. Dubé. Modeling the dynamical interaction between epidemics on overlay networks. *Physical Review E*, Vol. 84, No. 2, p. 026105, 2011.

[76] Emanuele Cozzo, Raquel A. Banos, Sandro Meloni and Yamir Moreno. Contact-based social contagion in multiplex networks. *Physical Review E*, Vol. 88, No. 5, p. 050801, 2013.

[77] Osman Yağan, Dajun Qian, Junshan Zhang and Douglas Cochran. Conjoining speeds up information diffusion in overlaying social-physical networks. *IEEE Journal on Selected Areas in Communications (JSAC)*, Vol. 31, No. 6, pp. 1038-1048, 2013.

[78] Camila Buono, Lucila G. Alvarez-Zuzek, Pablo A. Macri and Lidia A. Braunstein. Epidemics in partially overlapped multiplex networks. *PloS one*, Vol. 9, No. 3, e92200, 2014.

[79] Anna Saumell-Mendiola, M. Ángeles Serrano and Marián Boguná. Epidemic spreading on interconnected networks. *Physical Review E*, Vol. 86, No. 2, p. 026106, 2012.

[80] Huijuan Wang, Qian Li, Gregorio D'Agostino, Shlomo Havlin, H. Eugene Stanley and Piet Van Mieghem. Effect of the interconnected network structure on the epidemic threshold. *Physical Review E*, Vol. 88, No. 2, p. 022801, 2013.

[81] Joaquín Sanz, Chengyi Xia, Sandro Meloni, Yamir Moreno. Dynamics of interacting diseases. *Physical Review X*, Vol. 4, No. 4, p. 041005, 2014.

[82] Xuetao Wei, Nicholas Valler, B. Aditya Prakash, Iulian Neamtiu, Michalis Faloutsos and Christos Faloutsos. Competing memes propagation on networks: a case study of composite networks. *ACM SIGCOMM Computer Communication Review*, Vol. 42, No. 5, pp. 5-12, 2012.

[83] Xuetao Wei, Nicholas Valler, B. Aditya Prakash, Iulian Neamtiu, Michalis Faloutsos and Christos Faloutsos. Competing memes propagation on networks: A network science perspective. *IEEE Journal on Selected Areas in Communications*, Vol. 31, No. 6, pp. 1049–1060, 2013.

[84] Jacob Goldenberg, Barak Libai and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, Vol. 12, No. 3, pp. 211-223, 2001.

[85] Ceren Budak, Divyakant Agrawal and Amr El Abbadi. Limiting the spread of misinformation in social networks. *Proceedings of the 20th international conference on World wide web*, ACM, pp. 665-674, 2011.

[86] Mark Granovetter. Threshold models of collective behavior. *The American journal of sociology*, Vol. 83, No. 6, pp. 1420-1443, 1978.

[87] Duncan J. Watts. A simple model of global cascades on random networks.

Proceedings of the National Academy of Sciences, Vol. 99, No. 9, pp. 5766-5771, 2002.

[88] Damon Centola, Víctor M. Eguíluz and Michael W. Macy. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, Vol. 374, No. 1, pp. 449-456, 2007.

[89] Stephen Morris. "Contagion". *The Review of Economic Studies*, Vol. 67, No. 1, pp. 57-78, 2000.

[90] David Easley and Jon Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world. *Cambridge University Press*, 2010.

[91] David Kempe, Jon M. Kleinberg and Éva Tardos. Maximizing the spread of influence through a social network. *Theory of Computing*, Vol. 11, No. 4, pp. 105-147, 2015.

[92] Charles D. Brummitt, Kyu-Min Lee and K-I. Goh. Multiplexity-facilitated cascades in networks. Physical Review E, Vol. 85, No. 4, p. 045102, 2012.

[93] Osman Yağan and Virgil Gligor. Analysis of complex contagions in random multiplex networks. Physical Review E, Vol. 86, No. 3, p. 036103, 2012.

[94] Yubo Wang and Gaoxi Xiao. Effects of interconnections on epidemics in network of networks. *Wireless Communications, Networking and Mobile Computing (WiCOM), 2011 7th International Conference on. IEEE*, 2011.

[95] Sebastian Funk and Vincent AA Jansen. Interacting epidemics on overlay networks. *Physical Review E*, Vol. 81, No. 3, p. 036118, 2010.

[96] Byungjoon Min and K. I. Goh. Layer-crossing overhead and information spreading in multiplex social networks. *seed*, Vol. 21, No. T22, T12, 2013.

[97] Shohei Usui, Fujio Toriumi, Takatsugu Hirayama, and Kenji Mase. Analysis of information diffusion on directed networks. *The 6th International Workshop on Emergent Intelligence on networked Agents* (WEIN'2014), Paris France, May 2014.

[98] Peng Jia, Anahita MirTabatabaei, Noah E. Friedkin and Francesco Bullo. Opinion dynamics and the evolution of social power in influence networks, *SIAM Review*, Vol. 57, No. 3, pp. 367-397, 2015.

[99] X. Flora Meng, Robert A. Van Gorder, Mason A. Porter. Opinion formation and distribution in a bounded confidence model on various networks. *arXiv:1701.02070*, 2017.

[100] Mostafa Salehi, Rajesh Sharma, Moreno Marzolla, Matteo Magnani, Payam Siyari and Danilo Montesi. Spreading processes in multilayer networks. *IEEE Transactions on Network Science and Engineering*, Vol. 2, No. 2, pp. 65-83, 2015.

[101] Yoshiyuki Okada, Keisuke Ikeda, Kousuke Shinoda, Fujio Toriumi, Takeshi Sakaki, Kazuhiro Kazama, Masayuki Numao, Itsuki Noda and Satoshi Kurihara. SIR-extended information diffusion model of false rumor and its prevention strategy for Twitter. *Journal*

of Advanced Computational Intelligence and Intelligent Informatics, Vol. 18, No. 4, pp. 598-607, 2014.

[102] Zhaofeng Li, Fuhan Yan, Yichuan Jiang. Modeling temporal propagation dynamics in multiplex networks, *Agents and Data Mining Interaction (ADMI)*, Lecture Notes in Computer Science, Vol. 9145, pp. 26-37, 2015.

[103] James Coleman, Elihu Katz, and Herbert Menzel. The diffusion of an innovation among physicians. *Sociometry*, Vol. 20, No. 4, pp. 253-270, 1957.

[104] Manlio De Domenico, Albert Solé-Ribalta, Sergio Gómez and Alex Arenas. Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences* (*PNAS*), Vol. 111, No. 23, pp. 8351-8356, 2014.