COMPLEX WEB-API NETWORK CONSTRUCTION BASED ON BARABÁSI-ALBERT MODEL AND POPULARITY-SIMILARITY OPTIMIZATION MODEL

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF COMPUTER INFORMATION SCIENCE

Supervisors

Dr. Jian Yu

Dr. Sira Youngchareon

July 2019

By

Hengbin Wang

School of Engineering, Computer and Mathematical Sciences

Copyright

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the qualification of any other degree or diploma of a university or other institution of higher learning.

Signature of candidate

Acknowledgements

I would like to thank my thesis advisor Dr. Jian Yu, who gave me help when I needed. Besides, he always gave me enlightening advices for my dissertation. Meanwhile, although he respected me great freedom about how I wanted to compose this dissertation, he also corrected me when he found anything inappropriate.

Abstract

Today, Web services are applied in a variety of industries, and compose the building blocks of many Web-based and mobile applications. They are essential for the cross-organizational functional integration and data sharing across the network. On the one hand, how to construct a network in the Web service ecosystem to better organize them is the current research focus. On the other hand, Web service discovery is also a fundamental for integrating the right services into the business scenario. In this work, we used a mathematical method for the evaluation of the social web application programming interface (API) network on the basis of data collected from $ProgrammableWeb$ from the perspective of network science. We constructed two Web-API network models. One scale-free network is composed based on Barabasi-Albert model, the other used popularity-similarity optimization model which considers the similarity between nodes to enhance the performance of service discovery. We discussed the theoretical approach that we used to construct network models and also present the develop procedures. After the two Web-API network models were constructed, we evaluated the two network models, including nodes degree distribution, power-law, exponent, lower bound and preferential attachment. We discovered that Web-API network can suit the power-law distribution, and the performance of service discovery with them are better than typical means.

Keywords: Web services ecosystem, power-law, scale-free, node similarity

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Aim

With of the rapid revolution of Internet technology and also the emergence of network services, Web services on the basis of Internet technology have been applied in a variety of industries around the world. "Network Service" refers to some service-oriented, distributed program-based software modules that run on the network. Web services use common Internet standards like hyper text transfer protocol (HTTP) and extensible markup language (XML), which are the subset of the standard universal markup language, enabling people to access the data on the Web through different terminals, for instance, online booking, checking the status of the reservation. Network services have been used extensively in e-commerce, e-government, and electronic business process applications, and are assumed to be the next trend of the Internet by industry insiders (Chesbrough & Spohrer, 2006). More and more enterprises are serving as service providers to offer various services over the Web, while others as service consumers use existing Web services to develop their own businesses. With the in-depth study of web services, many materials and open source code can be easily searched on the Internet. Improving the API is not a quite complicated task for now while the quality

of these products is not uniform. As the number of available Web services continues to increase, it seems to be natural to reuse existing Web services. It is designed to create composite Web services (Gronmo, Skogan, Solheim & Oldevik, 2004). Many Web services such as Google Maps, Twitter, and YouTube have been applied extensively by thousands of service consumers. Besides, there are also a great amount of APIs hardly ever being applied ever since publication (W. Chen, Paik & Hung, 2013). To better acknowledge and explore the correlation between APIs, one of the methods is to build them into a complex network.

A complex network composed by nodes and edges. The graph is the research content in the field of mathematics. The algorithm of "graph theory" is universal and focuses on theory. The complex network focuses on engineering, which brings the theory into the real-life production and combines the theory of graph algorithm, application scenarios, science and technology to help observe and understand the real objective world (Barabási, Albert & Jeong, 2000). When it comes to the field of network science,, the scale-free network with node-degree power law distribution is a popular topic in recent years. It has been found that scale-free features are exhibited by the scientific citation networks in many real-world systems, such as the World Wide Web (WWW), the Internet, and scientific citation networks (Barabási & Albert, 1999). At the same time, this has been proved mathematically (Barabási et al., 2016). Clearly, network science helps people gain insight into complex networks.

On the other hand, in the process of constructing a network model for analysis, the usual complex network tends to ignore the similarity between nodes because it considers preferential attachment.

1.2 Background

In this section, we introduced the background of related work, including the history and current research status of network science, social network and Web services.

1.2.1 Network Science

Scale Network

Tracing the footprint of network science, the development of network theory is firstly benefited from the development of applied mathematics, such as graph theory and topology. Graph theory originated from the famous Seven Bridges in Königsberg (West et al., 1996). In the 18th century, seven bridges were built on the river, connecting the two islands in the middle of the river with the river bank. Some people have suggested that you can only walk once for each bridge and finally return to the original position. In 1736, Euler, the great mathematician at the time first simplified the problem. He regarded the two small islands and the two sides of the river as four points and meanwhile regarded the seven bridges as the connection between the four points. Euler created a map in the research, which is a new branch of mathematics. Therefore, it is the pioneering contribution of the first generation of scientists to the network. The problem is to start from any of the four lands, pass each bridge exactly once, and then return to the starting point. Euler solved this problem with used abstract analysis to turn this problem into the first graph theory problem: to replace each land with one point and then connect each bridge with two point (Yurke & Denker, 1984). This is done instead of getting a graph or creating a network.

Stochastic network

The Hungarian mathematicians Erdos and Renyi proposed stochastic network theory in 1959. They use a probability to construct a random network to determine whether or not to connect edges between two nodes. Network systems in communications and life sciences can be effectively simulated by randomly connecting edges between network nodes. According to the theory, it can be seen that although the connection is random, the resulting network is highly democratic, which means the number of nodes is roughly the same, and the number of node connections is a bell-shaped Poisson distribution. The random network proposed by Erdos and Renyi is the most studied network model. It generates a network by randomly connecting nodes with other nodes. The network provides an important network modeling reference framework. Although there is no real network conforming to this model, there are many of its features that can be accurately calculated. It is an important theoretical tool for studying real networks and can present a series of assumptions for people to test.

Bollobasla published "Random graphs" in 1985, which is a representative work of random networks. They use a relatively simple random graph to describe the network, which is referred to ER stochastic theory. Their most important finding is that the growth of the network scale suddenly emerged under certain conditions. In the nearly 40 years after 1960, this stochastic network theory was recognized as the theory to correctly understand the real network, which promoted the renaissance of graph theory and promoted the development of network theory.

Scale-free Network

Watts, a Ph.D. student in the Department of Theoretical and Applied Mechanics at Cornell University in the United States and his mentor Strogatz published a paper entitled Collective dynamics of small-world networks"in 1998. They proposed a small

world network model, which is a network model between a regular network and a random network. In 1999, Professor Barabási and his doctoral student Albert published in the journal Science entitled "The emergence of scales in random networks" which proposed a scale-free network model. The basic principle of their model is "growth" and "preferential attachment". It has discovered the scale-free nature of the network, which has attracted the attention of the whole world and promoted the vigorous development of the network (Strogatz, 2001).

The introduction of the small world network model and the scale-free network model symbolizes a new era for the research of network (Barabási, 2003; Watts, 2004; Barrat, Barabasi et al., 2004). The two major discoveries of small world networks and scale-free networks, and subsequent empirical studies of many real networks demonstrate that real-world networks are neither regular networks nor random networks. Instead, they have both small world and scale-free characteristics. Regular network and random graphs have completely different statistical characteristics. In the dynamic development of the Internet and the World Wide Web, as well as various other social, biological, and physical networks, some scientists have found that it is impossible to explain some new problems of their structure and evolution using the two network theories of regular network theory and stochastic network theory. They roughly call such networks a complex network.

The reviews and monographs on complex networks continue to emerge, from physics to biology, from social science to technology networks, from engineering technology to economic management. . Meanwhile, there are many other fields that have received unprecedented attention (Ben-Naim, Frauenfelder & Toroczkai, 2004; Dorogovtsev & Mendes, 2013). In recent years, new models based on the principles of imitation, optimization, multi-agent, and layering have emerged, which also play an important role in the study of complex network theory. People began to study complex networks from new heights.

1.2.2 Social Network

A social network is a form of social organization based on interconnections between nodes rather than clear boundaries and order. This is an analytical perspective that western sociology emerged from the 1960s. With the development of industrialization, urbanization and the rise of new communication technologies, society has become more and more networked.

Social network analysis was proposed in the attention of the famous British anthropologist Brown (1960). However, the focus of the network concept explored by Brown is about how culture regulates the behavior of members of bounded groups (such as tribes, villages, etc.), and actual interpersonal behavior is much more complicated. Therefore, in order to understand Brown's concept of "social structure", from the 1930s to the 1960s, more and more scholars began to construct social networks in the fields of psychology, sociology, anthropology, mathematics, statistics, and probability theory. Meanwhile, they about the network structure of society. Various network concepts including centrality, density, structural balance, blocks (Scott, 1988), etc have emerged continuously. Subsequently, the theory, methods and techniques of social network analysis have become much more in-depth and become an important research paradigm of social structure.

1.2.3 Web Service

With the continuous development of network services, various network service models have come up, such as traditional Web Services Description Language (WSDL) and Simple Object Access Protocol (SOAP)-based Web services, Semantic Web Services (SWS) and Representation State Transfer (REST) web API. Since it is only necessary to define the interfaces of the Web services to achieve interoperability with each other, regardless of their specific implementation language and internal data structure (Alonso,

Casati, Kuno & Machiraju, 2004), each API is relatively independent and has no connection or interaction with each other. Therefore, it is difficult for service consumers to easily find proper Web service, especially when the customer has multiple needs. On the other hand, due to the cooperation of Web service providers and Web service consumers, the large and small Web service ecosystems are formed and developed unconsciously (Barros & Dumas, 2006). A web service ecosystem can contain a set of web services and their relationships. In addition, rich interactions can lead to complex community structures. Then, with the development of the Web services ecosystem, the structure of these ecosystems will change dynamically, which makes the network service discovery even more difficult to achieve.

1.3 Motivation

With the rapid development of network technology, network service research has become a popular direction because of its features of loosely coupled and cross-platform. Web services are applied in vary aspect of people's lives. Social networks such as Facebook, Twitter and Linkedin are widely used, having a user base of millions of people nowadays, making it easier to get useful information for different specific purposes. The vast users and the huge number of network services make Web service network organization and Web service discovery a hot research topic.

From the perspective of graph theory, service consumers and service providers can be combined to form a binary network, and projections of bipartite graphs can be used to construct relationships between service providers. Feng, Lan, Zhang and Chen (2015) proposed a model to organize SWS into a Web service network. On the other hand, Cherifi and Santucci (2012) provided a comparative evaluation of Web services composition networks models from the view of topology.

As for Web service discovery, according to the SOA architecture design specification,

the entire architecture consists of three parts, namely service provider, service consumer, and service registry, among which, the service provider mainly uses a language such as WSDL to describe the specific service, and is responsible for publishing the Web service to the service registry for the service consumer to call. The service registry uses universal description, discovery, and integration (UDDI) to manage and register the available Web service description information. Meanwhile, it is also responsible for receiving the service of consumer's query request. After searching for the appropriate candidate service, the service registry establishes the service. The relationship between supply of the service provider and demand of requester is explored. The service consumer, as the user service registry of the web service, makes an application request. After returning the search result in the service registry, communicates with the service provider complete the final service call through SOAP. Although there are many existing service discovery methods, they all have some defects. Although UDDI provides a higher level of functionality, the service provider must be registered with the UDDI registry in order to ensure that customers can retrieve the service. This greatly reduces the search scope and efficiency (Tsalgatidou & Pilioura, 2002).

In summary, the prospect for the development of a Web service network model, which is constructed mathematically and rooms to be further improved and enhanced for service discovery motivated us to construct a social Web service network adopting the appropriate network models.

1.4 Research Scope and Methodology

Due to the great amount and variety of web services, it is unlikely to cover all forms of web services, such as traditional web services, SWS and REST-style web APIs. In

recent years, researches on APIs has been the most popular among all types of network services. In addition, Web APIs currently dominate the Web services sector compared to traditional Web services (Maleshkova, Pedrinaci & Domingue, 2010). Therefore, all the data collected in this thesis are related to the API. $ProgrammableWeb$ is a platform that provides information and source of APIs. It contains around 17,800 APIs and 6000 web service composite applications, which are called mashups. Therefore, the data of $ProgrammableWeb$ is representative. In this paper, we focus on the characteristics of the Web-API network. Besides, based on the network model, we constructed to study some characteristics of the network to obtain results.

The scientific research methods used in this study mainly include objectives, research, experiments, analysis and conclusions. The research process of the Web-API network is stated below

1. Stated the situation of Web services ecosystem and proposed questions.

2. Based on the raised research questions, an intensive literature review was conducted to understand the current research achievements in terms of network science, and Web service networks.

3. Under the guidance of various network theories, designed experiment of constructing Web-APIs network model

4. According to the result of experiments, measured appropriate characteristics of networks and analyzed with mathematical methods.

5. Concluded the previous work and results, a future work was given as well.

In summary, the purpose of this thesis is mainly to discover the following research questions:

1. Construct Web-APIs social network models: one is based on Barabási–Albert (BA) network model while the other is based on popularity x similarity (PS) between web services.
2. Evaluate the network characteristics of BA model and PS model.

1.5 Contributions

As mentioned above, the current popular research topic of Web service network is about organization and service discovery. Therefore, we mainly focus on these two aspects and the contributions of this research are stated below:

Firstly, in the field of network science, the aim of this paper is to construct a social Web service network in an appropriate method from the perspective of graph theory and measure them using the key characteristics of networks.

Secondly, the construction of social Web service networks builds connections between APIs, which helps with the service recovery.

1.6 Thesis Structure

- Chapter 2 presents a literature review in terms of complex network, network topology modelling, network properties, web service networks and network model construction strategy, which discussed the previous work and current achievement in the field of network science and web service.
- Chapter 3 describes about how to construct Web service network in details, including the tools, graph theory and mathematical method we used. A model fitting method is also included.
- Chapter 4 illustrates the result of network characteristic results.

- Chapter 5 discusses the understanding and meaning of the experiment results in the use of web service network and web service discovery.

Chapter 2

Literature Review

This chapter introduces and reviews the fundamental of complex networks, network topology modelling, network properties, Web services networks, Web service discovery and Web-API network model construction strategy.

We review the related typical network models, including random networks, small-world networks and scale-free net- works. The illustrations demonstrate how these network models are suitable for a real network. In addition, we also introduced the diverse properties of complex networks, including degree distribution, clustering coefficient, preferential attachment, betweenness and average path length.

Then, apart from the perspective of network science, we also discuss the significance of constructing a web service network for service discovery and recommendation and introduce the current methods of web service discovery.

Based on the motivation of this paper, the analysis of existing web service generation is necessary. It helps to choose a better approach for the network model. Since this thesis mainly focuses on two perspectives to construct the network model: popularity and similarity, this literature review will discuss previous related work based on these two points.

2.1 Complex Networks

Complex network has been explored and improved during the last twenty years. Networks have penetrated everywhere. The correlation between individuals compose the social network and every creature is the outcome a biochemical response network (Boccaletti, Latora, Moreno, Chavez & Hwang, 2006).

The mechanism of the system or the network is strongly related with the function, which explains the reason why there are many scholars focusing on this complex network. When it comes to the road traffic jam of the city, we need to have a comprehensive and clear understanding regarding the structure of the current urban traffic roads. Otherwise, it will be difficult to figure out a appropriate solution to deal with the issue of road congestion. Thus, the reason why we explore the network mechanism is due to that it can dramatically impact the results and functions.

A classical network is composed by a variety of edges and nodes, in which nodes are applied to stand for different individuals in reality while edges are applied to stand for the correlations between every individual. If there is a specific correlation between two nodes, they can be connected by an edge. Otherwise, the nodes will have no edge. Nodes being connected by edges are assumed to be adjacent of the mechanism. For instance, the nervous system can be assumed to be the system of cells connected through nerve fibers. The computer network can be assumed to be a network composed through interconnecting computers, which work independently via the communication media, for example coaxial cables, twisted pairs and optical cables, etc citewatts1998collective , which is same with the power network (Faloutsos, Faloutsos & Faloutsos, 1999), social network (Hofman, Sharma & Watts, 2017), traffic network, scheduling network and so on.

The complexity can be shown from the aspects in below:

- Complex structure: there is a huge number of nodes and the structure has a variety

of features.

- Network evolution: shown in the disappearance and generation of the connections or nodes. For instance, there might be disconnection or appearance of the pages or links at any time in the world wide network, leading to constant changes of the network structure
- Connection diversity: There are differences in terms of the weights of connection between nodes. Meanwhile, sometimes, there are directional edges.
- Dynamics complexity: Node sets are classified into nonlinear dynamic mechanisms, for example changes of the node states over time.
- Node diversity: Nodes of the complex network can stand for anything. For instance, the complex network node consisted by human relationships stands for a single individual. The complex network node consisted by World Wide Web can stand for various web pages.
- Multiple Complexity Fusion: The above multiple complexities have interaction with one another, resulting in more unpredictable outcomes.

2.1.1 Random Network

Erds and Rényi (1960) proposed the notion of random network model, which is also named as the Erdos-Renyi (ER) model. The random network is also called as the random graph, which is a complex network generated through the random process. It is on the basis of a "natural" construction approach. Assuming that there are altogether n modes, it is assumed that the chance of connection between every pair of nodes is constant $0 < p < 1$. The ER model network is therefore constructed (1960). Originally, this model is used by scientists to illustrate the real-life networks.

Barabási(2016) gives two definitions of random network model:

- G(N,L) model: There are N nodes being connected through the placed links L randomly.
- G(N,p) model: The probability p connects every pair of N labeled nodes (Gilbert, 1959).

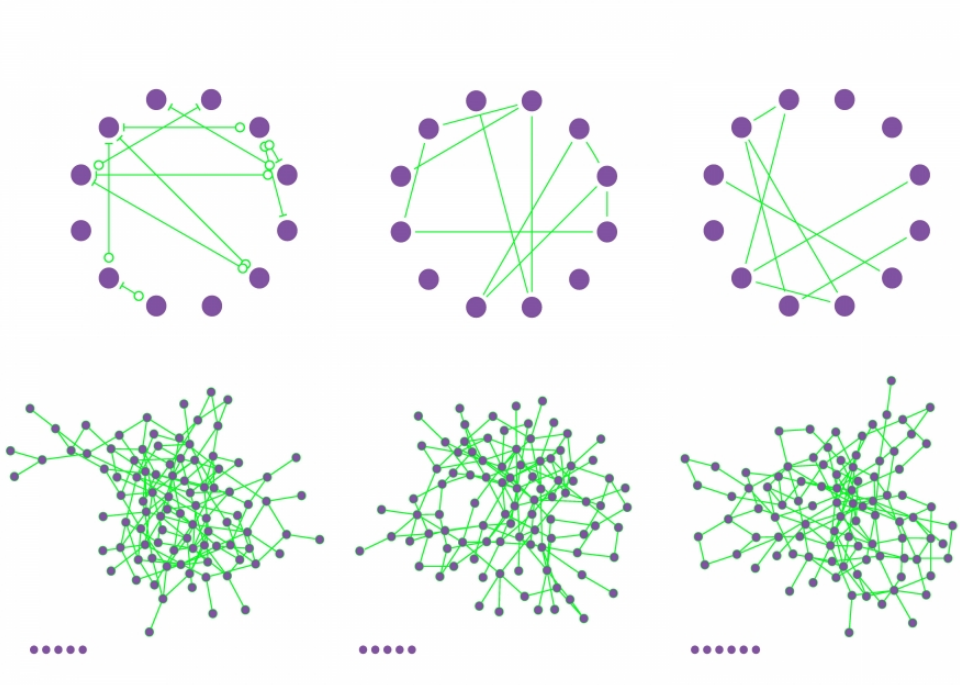


Figure 2.1: Random Network Model
Source, Barabási et al., 2016

Given n nodes, the probability p connects every pair of nodes. Maximally, there are $N(N-1)/2$ edges. Therefore, there is a random variable of all the connections. The probability p selects the relative nodes randomly. Afterwards, through connecting them, it can generate random network. The [Figure 2.1] demonstrates the random network model. Even though the connections are set randomly, the volume of the connections for the majority of nodes will be generally the same according to the random network model. A bell-shaped Poisson distribution is followed by the distribution of nodes, with the feature of "average". The average volume of nodes is way lower or higher compared with the number of connections. Corresponding with the increase of connections, there

is an exponential decrease of the probability. Thus, the level of the random network distribution is same with the Possion distribution, which is denoted as equation 2.1 where k means sparse real networks, as showed as [Figure 2.2]

$$p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!} \tag{2.1}$$

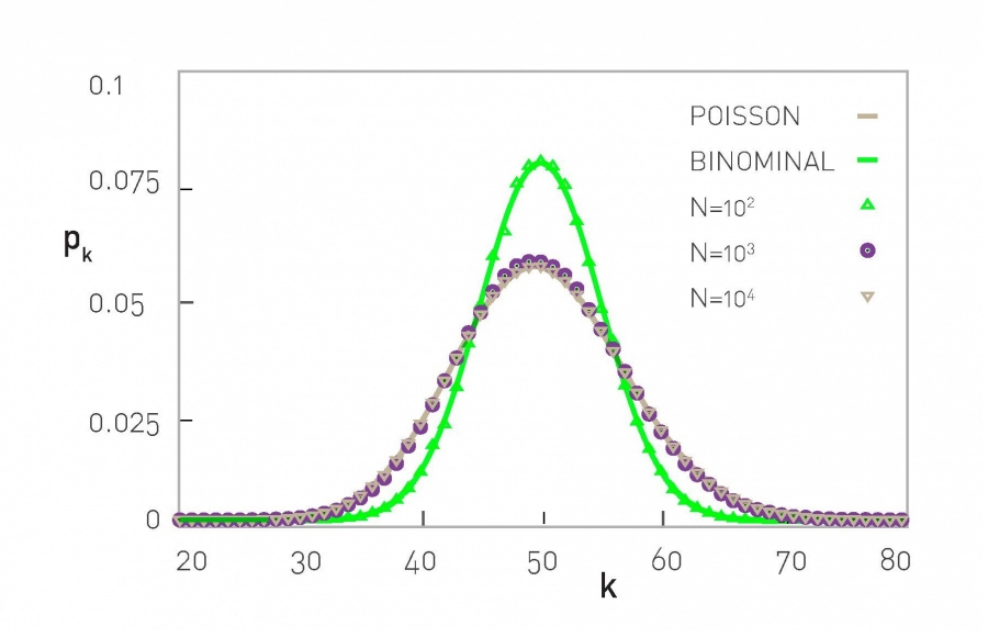


Figure 2.2: Random Network Distribution
Source, Barabási et al., 2016

2.1.2 Small-World Network

There are a variety of real networks belonging to small-world, which means they possess a smaller shortest path and relatively large clustering coefficients (Albert & Barabási, 2002). In the early network research, people only discussed the random network model and the regular network. A regular network is a network with a simple connection structure. For instance, the connection forms of each node in the network are the same. The regular network is featured through a very small clustering coefficient and a large

average path length. Until the end of the 20th century, a group of mathematics, physics and computer experts creatively constructed a mathematical model for this sociological concept, and cited the two structural parameters of network mean path length and network clustering coefficient to describe the small world nature of network systems. A typical representation of the small world network model is the Watts-Strogatz (WS) small world network model(Newman, 2000) and the Newman-Watz (NW) small world network model (Newman, 2000) [Figure 2.3].

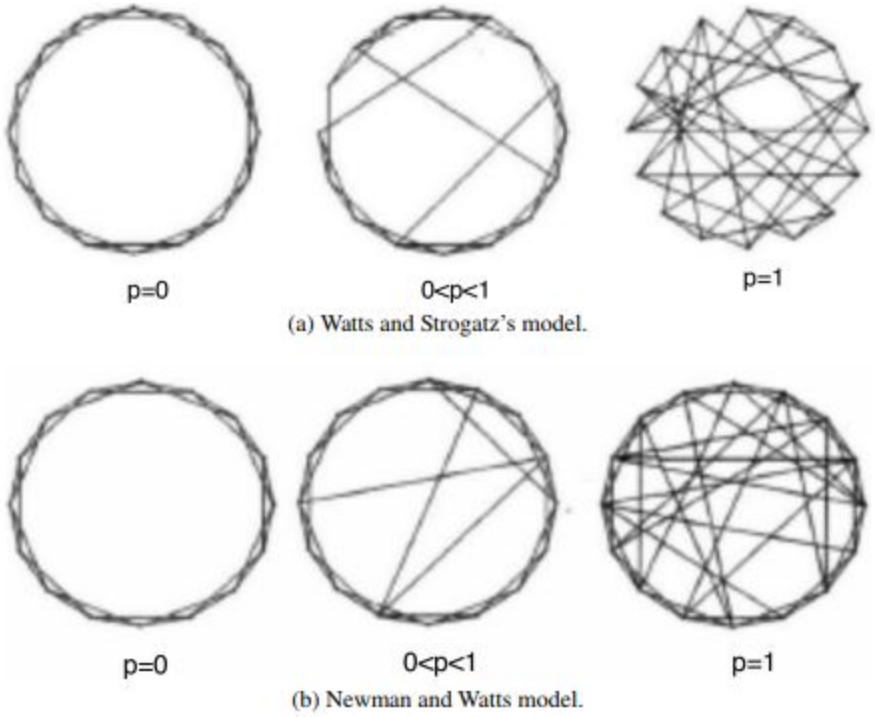


Figure 2.3: Small World Models
Source, Ziqian, 2016

Watts-Strogatz Model

Watts and Strogatz found that the regular network is more clustered. The average path length of the network is also larger while the average distance of random networks is shorter, the clustering is low. Neither the real-work network is complexly random nor

regular. Instead, it is somewhere in between. Therefore, they are based on a regular network and then randomly reconnect each edge in the rule network with probability p without considering the node's self-join and heavy-edge connections. These randomly reconnected edges introduce both long-range and randomness connections of regular network. These connections greatly reduce the average path length between network nodes, while maintain the high clustering of the original network, making the network have a small world characteristic.

Newman-Watts Model

Newman and Watts improved the WS small-world network model. With the prerequisite of not changing the edge between nodes of the original regular network, there are two nodes being chosen. Whether an edge is added among these two nodes is determined by the probability p. This method also introduces both randomness and long-range connections in the regular network, which makes the network have small world characteristics. The NW model differs from the WS model given that it does not cut off the original edges of regular network, but reconnects a pair of nodes with a probability p. The advantage of the NW model is that it simplifies the theoretical analysis because isolated nodes do exist in the WS model, but not in NW model. In fact, when p is small and N is large, the theoretical analysis of the two models is the same.

2.1.3 Scale-Free Network

However, most of the real networks are not random networks. A few nodes usually have a great number of connections, while the rest have few edges. Generally speaking, they conform to the zipf law (Powers, 1998). According to its nature, people have a special name for this kind of network - scale-free networks. The scale-free here means that the network lacks a characteristic value (or average value), that is, the fluctuation

range of the node degree value is quite large.

The characteristic of a scale-free network is that its degree distribution does not have a specific average indicator, which means the degree of most nodes is close. In the study of scale-free network degree distribution, Barabási, Albert and Jeong (1999) found that it follows the power law (also known as the Pareto distribution), meaning that when a node is extracted randomly, its degree d is proportional to a certain power of the natural number k (generally negative, denoted by $-\gamma$). Hence, the larger the k is , the lower the probability of d = k will be. Nevertheless, with the increase of k ,the probability is decrease slower: In a general random network, the rate of decline is exponential, while in a scale-free network its rate drops as polynomial.

$$p_k \sim k^{-\gamma} \tag{2.2}$$

Then $\log(k)$ linearly decide the value of $\log(p_k)$ with slope of this graph is γ [Figure 2.4].

$$\log p_k \sim -\gamma \log k \tag{2.3}$$

Barabási-Albert Model

In 1999, physicists Barabási and Albert proposed a scale-free network model, referred to as BA model, which is also referred as BA model. They found key characteristics of scale-free networks:

- Growth: Due to the new nodes comes, the size of network continuous increase (WWW generates many new web pages every day).
- Preferential Attachment: New nodes are more likely to connect with large nodes (such as consumers are more inclined to buy those items that sell well on eBay).

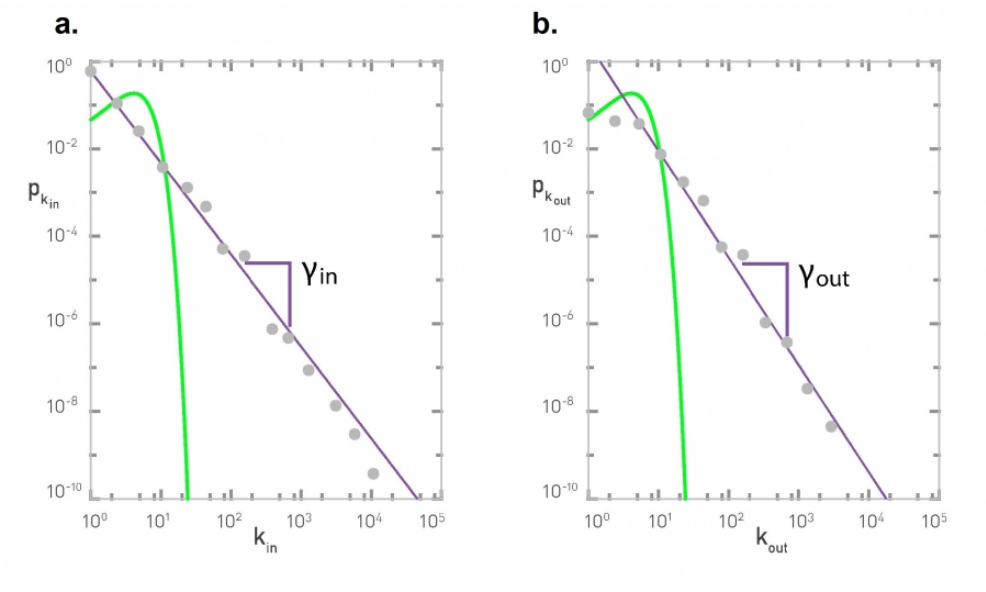


Figure 2.4: Power law
Source, Barabási et al., 2016

These two characteristics make the network evolve and present a self-organizing process.

The construction of a BA model is as follows: Given an initial node m_0, each new node connects to the existing node with the probability Π(k).

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{2.4}$$

The fundamental difference between the BA model and the previous model is that the nodes in the network are no longer fixed but continue to grow preferentially. For real networks, preferential attachment is indispensable. In order to approve their opinion, they construct two models. Model A changes the preferential attachment to random connection. There is only a random relation between the growing nodes. The random network degree distribution is subject to the exponential distribution. Meanwhile, the number of nodes in model B is fixed given that the number of nodes in the network is kept constant. With the continuous increase of the connection, all nodes in the network

will get interconnected at last.

Apart from classic random network models and BA scale-free network models, scientists have also proposed other models describing the structure of real network, such as local-World Evolving Network (Li & Chen, 2003), weighted evolving networks (Barrat, Barthélemy & Vespignani, 2004) and deterministic network model (Wood, 1993). These models can help people to understand the real network and solve existing problems better.

2.2 Network Topology Modelling

Internet topology models can be divided into two categories: One is to describe Internet topology features, including Waxman model (Waxman, 1988), Tiers (Doar, 1996), Transit-Stub, and power law while the other one describes the mechanism of topological feature formation, including BA and ESF (Extended Scale-Free Model) (Albert & Barabási, 2000) and an improved generalized linear preference (GLP) model (Bu & Towsley, 2002).

For the first type of model, the discovery of Internet topology features is actually the discovery of metrics. A topology model belonging to the first category is composed by several existing or newly discovered metrics. Then, the values of these metrics are obtained from the actual Internet topology data. Therefore, the evaluation of such models needs to start from two aspects. On one hand, it requires to evaluate the topology data used by it. On the other hand, it needs to evaluate its metrics. Among all the Internet topology metrics that have been discovered, the most basic metric is the node frequency distribution f_d. Its distribution is the most important basis for judging whether or not a topology map is similar to the Internet topology. According to the distribution of node frequency, the Internet topology model can be divided into the following categories.

Random Topology

The random type means that the Internet topology map is in a completely disordered state and is uniform on a large scale. The Waxman model is a stochastic model similar to the ER model while the out-degree frequency is Poisson distribution. There are two versions of this model:

1. The nodes are randomly arranged in a Cartesian grid. The distance between the nodes is their Euclidean distance.

2. According to (0, L) uniform random distribution for the specified distance of the node pair.

In both versions, the probability $P(u, v)$ of the connection between nodes is related to its distance and follows the Poisson distribution. The closer the distance, the greater the probability.

$$P(u, v) = \beta \exp \frac{-d(u, v)}{L\alpha} \tag{2.5}$$

where $d(u, v)$ represents the distance between nodes u and v, L is the longest distance between nodes, and the range of α and β is (0,1).

Layer Topology

Layer topology generated from the understanding of the hierarchical characteristics of the Internet structure, the nodes on the same layer are close to each other, and the degree of nodes between different layers are different. The Waxman model method is used for node layout on the same layer. The Tiers model divides the Internet into three levels: local area network (LAN), metropolitan area network (MAN) and wide area network (WAN). In this model, there is only one WAN while the topology map is constructed by specifying the number of LANs and MANs. The Transit-Stub model (Zegura, Calvert Donahoo, 1997) divides the AS domain into Transit and Stub classes. In this model,

Transit nodes are interconnected to form a node group. At least one Transit node group forms the core of the topology graph, and the Stub nodes are distributed in the Transit node group, which is connected to the Transit node. Transit-Stub is part of the Georgia tech Internetwork topology (GT-ITM) model package. Sometimes, GT-ITM refers to the Transit-Stub model.

Power-Law Topology

The node degree distribution of the Internet satisfies the power law distribution, which means that the node degrees are actually very different. In the double logarithm (log-log) graph, it should appear as a straight line with a negative slope. Such linear relationship is to evaluate the random variable in the given instance whether or not can satisfy the basis of power law distribution. The power-law distribution characteristics currently known altogether include three expressions (Faloutsos et al., 1999).

- The node out-degrees is proportional to the R power of the node ranking:

$$d_v \propto {r_v}^R \tag{2.6}$$

  where d_v refers the degree of the node v, and r_v represents the rank of nodes v in descending order of degree in the network topology

- The percentage of nodes with degrees greater than d in the network topology is proportional to the D power of node out-degrees:

$$f_d \propto d^D \tag{2.7}$$

  where f_d represents the percentage and D is the reciprocal of R.

- The eigenvalue λ_i is proportional to the ϵ power of its order i:

$$\lambda_i \propto i^{\epsilon} \tag{2.8}$$

Where λ_i is the eigenvalue of the network corresponding connection matrix, and i is the sequence number when the eigenvalues are arranged in descending order.

2.3 Network Properties

2.3.1 Degree Distribution

Degree distribution is a geometric property that reveals the basic statistical properties of complex networks. Generally, the degree distribution of nodes in a network is represented by the function P_k, which means that for a node, the probability of its degree (the number of connections of this node to other nodes) is value k. From the current point of view, there are mainly two methods that can be used for the calculation of the degree distribution. One is the physical kinetics (Pitaevskii, Lifshitz & Sykes, 2017), such as the mean-field theory. The other one is the probability theory, which mainly includes the master equation approach and Marcov chain method (Gilks, Richardson & Spiegelhalter, 1995).

There are two major types of distribution. One is the Poisson distribution, in which the P_k is exponentially decreasing away from the peak. The other type is the power law distribution, which is also called the scale-free distribution. Since the power law distribution has some scale-free properties. When a probability distribution function $f(x)$ is considered, an arbitrary constant d is given, there is a constant g that meets the following condition:

$$f(dx) = gf(x) \tag{2.9}$$

According to studies in recent years, it has shown that the Poisson distribution is no longer consistent with the distribution of most real-world networks. The distribution of most networks differs from the random network in which their curves are obviously tilted to the right, and the power-law distribution describes the degree distribution of real networks better [Figure 2.5]. The degree exponent is between 2 and 3, such as the WWW , the Internet, bio-networks and social networks.

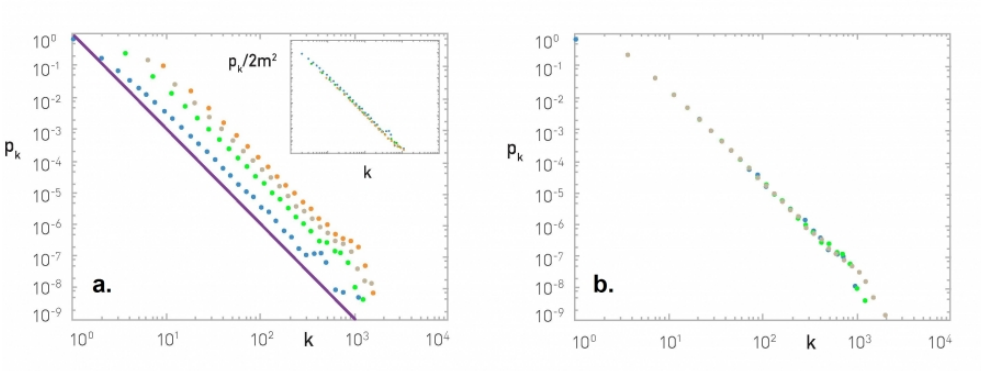


Figure 2.5: Degree Distribution
Source, Barabási et al., 2016

The prediction of BA model degree distribution follows a power law with degree exponent γ=3, denoted as:

$$p(k) \approx 2m^{1/\beta}k^{-\gamma} \tag{2.10}$$

with

$$\gamma = \frac{1}{\beta} + 1 = 3 \tag{2.11}$$

2.3.2 Clustering Coefficient

The clustering coefficient is a local feature quantity in the network, which describes the aggregation of these nodes in the network. For example, researchers have found that some friends of a person may also know each other in social networks.

The calculation of the clustering coefficient is listed in below: if a node j has the number of edges k_j and is connected with k_j nodes, it can be assumed that under this condition all nodes are interconnected to each other while the maximum number of edges will be denoted as: $k_j(k_j - 1)/2$. The actual exist number of edges is defined as e_j, the ratio of $k_j(k_j - 1)/2$ and e_j is the clustering coefficient C_j (Travels, 1967).

$$C_j = \frac{2e_j}{k_j(k_j - 1)} \tag{2.12}$$

The clustering coefficient of a whole network is the average of clustering coefficient of each nodes in this network. Obviously, if all nodes are isolated, the clustering coefficient is definitely zero. Meanwhile, if all nodes are fully connected, the C_j is one. However, C_j is always smaller than 1 in real network. Nevertheless, according to studies, nodes in real networks tend to gather together. Although the cluster coefficient C_j does not exceed 1, they are much larger than N^{-1}.

The clustering coefficient of the Barabási-Albert model follows (Klemm & Eguiluz, 2002; Bollobás & Riordan, 2003):

$$\langle C \rangle \sim \frac{(\ln N)^2}{N} \tag{2.13}$$

Since the dependence of $1/N$ is quite different between the random network and scale-free network, the clustering coefficient increase when N is large. Consequently the Barabási-Albert network, compared with the random network, is more clustered locally [Figure 2.6].

2.3.3 Preferential Attachment

The concept of preferential attachment is similar to cumulative advantages (Price, 1976) or Matthew effect (Merton, 1968).It is observed that when a new node is added to

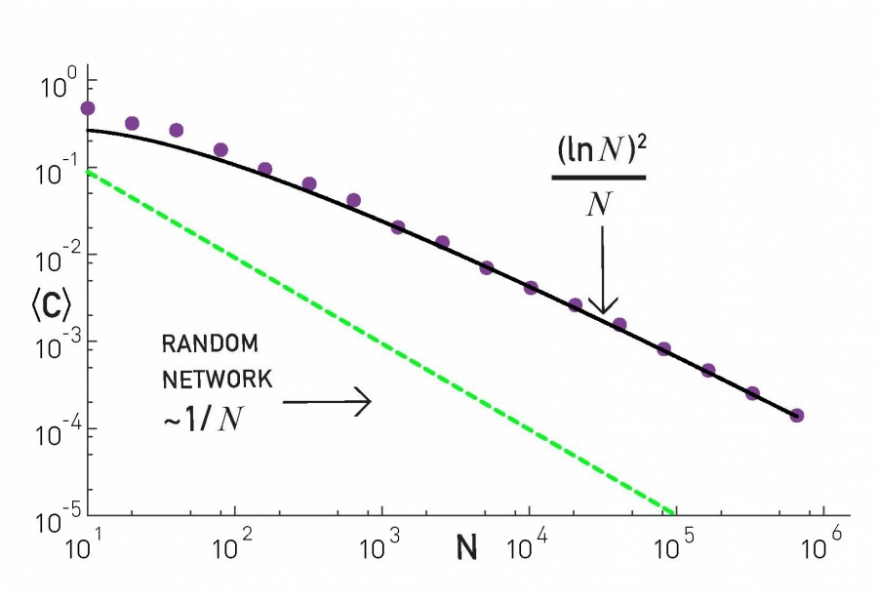


Figure 2.6: Clustering Coefficient
Source, Barabási et al., 2016

an existing network, it prefers to connect to nodes with higher degree distribution (Newman, 2001).

$$\frac{\Delta k_i}{\Delta t} \sim \Pi(k_i) \tag{2.14}$$

As equation 2.14 shows, the probability of checking a node is related to its degree Δk_i and changed time Δt, where $\Delta k_i = k_i(t + \Delta t) - k_i(t)$.

Barabasi uses two parameter (growing character of the network and eliminate preferential attachment) of network growth to prove that PA is actually present and significantly important in the real networks. PA has been measured in various real systems, such as citation network, Internet, neuroscience network and actor network (Jeong, Néda & Barabási, 2003). It is found that there is a linear increase of PA, so $\Pi(k_i)$ can be approximated with:

$$\Pi(k) \sim k^{\alpha} \tag{2.15}$$

2.3.4 Betweenness

The betweenness (Holme, Kim, Yoon & Han, 2002) is also an important metric for Internet network topology, which can be divided into node betweenness and edge betweenness. Among them, the number of nodes computes the shortest paths via the node of the network, demonstrating the pivotality of the nodes in the network. The larger the betweenness of nodes is, the stronger the pivoting of this node will be. Deleting such a node will cause the shortest path between a large number of nodes grow. The edge betweenness is defined as the number of shortest paths through a certain edge in all the shortest paths of the network. Similarly, the edge betweenness also reflects the pivotality of the edge in the network.

If σ_{ij} is used to represent the number of shortest paths between nodes i and j, l is a node or edge, and $\sigma_{ij}(l)$ represents the number of shortest paths passing through l between nodes i and j, and the mathematical expression of the number is obtained as:

$$B(l) = \sum_{n=1}^{j} \sigma_{ij} \frac{l}{\sigma_{ij}} \tag{2.16}$$

Goh et al.(2001) found that the distribution of node betweenness in a network follows a power-law. The betweenness is a performance metric that reflects the traffic characteristics and is related to the link bandwidth and the utilization of router nodes, especially in networks generated based on the shortest path.

2.3.5 Average Path Length

The distance d_{ij} between two nodes i and j is assumed to represent the number of the sides over the shortest path linking the two nodes. The maximum distance is defined as

the network's diameter, which can be denoted as:

$$D = \max_{i,j} d_{ij} \tag{2.17}$$

The average path length ℓ of the network is defined as the mean of the distance between any two nodes:

$$\ell = \frac{1}{N(N-1)} \sum_{i \geqslant j} d_{ij}, \tag{2.18}$$

where N is the number of network nodes. The average path length of the network is also referred to as the characteristic path length of the network. In order to facilitate the mathematical processing, the distance from the node to itself is included in the formula.

2.4 Web Service Network

The composition of the web service and social network has drawn great attention, which basically involves three approaches (Maamar, Faci, Wives, Yahyaoui & Hacid, 2011; Maamar, Faci, Wives, Badr et al., 2011; Maamar, Hacid & Huhns, 2011; Jiang, Yang, Yin, Zhang & Cristoforo, 2011; Tan et al., 2011).

- Collaboration: In a complex network, social web services evaluate and combine their functionalities and request collaboration when they are faced with various user needs. Based on previous experience, social web service decides whether they need to recommend and collaborate with other peers.

- Competition: Some web services offer similar functionalities, which can be distinguished by their nonfunctional properties. In other words, there is a competitive relationship between them. They need to know and learn from other competitors in the network to satisfy users. When web services meets users' requirement, they has the ability to improve their own nonfunctional properties according to

other peers (Alrifai, Skoutas & Risse, 2010).

- Substitution: If a service fail, other peers offering similar functionalities will also help even though they compete (Maamar, Wives et al., 2011). When there is a potential failure, the service then identifies its own best substitutes in response to meet users' nonfunctional requirements.

However, the above approaches have limitations during the construction of a large-scale global web service network. In order to enhance the capability to a global scale, W. Chen et al. (2013) to connect web services, which are previously isolated to global social service network. According to linked social service-specific principles they propose, all services were published on the world wide web. After that, a new complex network model was constructed which following principles. All data items are linked with others. The linked data technology improves the data discover (Bizer, Heath & Berners-Lee, 2008). Generally. the linked data suggest:

- using URIs for things
- using HTTP URIs when people search for data
- using standards (RDF, SPARQL) for looking up a URI
- including links to other URIs for rich discovery

To recommend social services for referral links, an algorithm was also used, which considers the comprehensive analysis of dependency satisfaction rate, quality of service (QoS) preference, sociability preference, and preferential service connectivity. The results of experiment demonstrate the improvement of success rate and discovery time. The author mainly focuses on the preferential attachment property of complex network while there are others, such as transitivity, resilience and centrality (Ghoshal, 2009).

2.4.1 Network Characteristics

Transitivity

The transitivity of the network is calculated by weighted averaging the vertex-gathering coefficients of all the vertices, the degree of which is at least two. It is similar to clustering coefficient. In the previous sections, the generalized random graph models strongly pointed out the major shortcomings of early network models (such as Poisson random graphs), that is, the distribution of degrees are inconsistent with reality. However, most of these theoretical network models fail to grasp the common phenomenon of transitivity (Stephenson & Zelen, 1989; Burda, Jurkiewicz & Krzywicki, 2004; Park & Newman, 2004).

At present, the random graph model considering transitive includes bipartite graph model, community structure model and specific dual-graph model (Zha, He, Ding, Simon & Gu, 2001; Girvan & Newman, 2002). Because of the existence of closed loops and the important nature of adjacent vertices, independence is destroyed and this issue has not been solved. Currently, there is no effective way to combine the transitive and random graph models for the general networks.

Resilience

Cyber resilience, which is also called operational resilience, is a network attribution, which is related with the degree distribution when vertices are removed. The functionality of most of the networks depends on its relevance, which means that it depends on the existence of paths between pairs of vertices. If the network vertices have been removed, the marked length of these paths will have an increase, eventually making the vertex pairs non-associative. Meanwhile, the communication between them via the network will be interrupted (Newman, 2003).

During recent years, the work of Albert, Jeong and Barabási(2000) has stimulated

research interest in network resilience. They used two networks as examples to explore the effects of vertices being deleted. One is the autonomous system level of 6000-vertices depicting topological network while the other is the World Wide Web with a subset of 326,000 pages. It can be observed that the vertex distribution of the these two models basically follows the power law (Albert, Jeong & Barabási, 1999; Q. Chen, Chang, Govindan & Jamin, 2002; Faloutsos et al., 1999; Vázquez & Weigt, 2003). They found that for both networks [figure 2.5], random vertex removal has almost no effect on distance, which means networks are highly resilient to random vertex removal. Given that degrees for most vertices in these networks are low, there are only a few paths between them. Therefore, removing these vertices will hardly destroy the communication. On the other hand, when the removal is performed on the vertex with the largest degree, it will be found to have a destructive effect. The average distance increases from the vertex to the vertex as the proportion of removed vertex, and it is destroyed through the network. The original communication is only removed as a part of the vertices.

Centrality

As for centrality, it is used to quantify the importance of a vertex in the graph. Similarly, the centrality can also be used to quantify the importance of a node in the network. Freeman (1978) proposed to measure the centrality of social networks with degrees, closeness, and betweenness indicators. All of these three indicators can also be used to measure the centrality of the network.

Degree Centrality is the most common and easiest way. In an undirected network, network centrality can be measured by the degree of a node. High-degree nodes, such as base stations, access points in wireless networks are more likely to act as information exchanges information exchange site (Everett & Borgatti, 1999), denoted as equation

below.

$$C_D(v) = deg(v) \tag{2.19}$$

Closeness centrality is featured by the distance of nodes (geodesic distance, the number of edges included in the shortest path between two vertices). For instance, when the shortest path between one node to the other is short, then the node's closeness is high. Such metric can be applied to evaluate the volume of information being transmitted between one to another. The shortest distance between nodes is added to be reciprocal, which belongs to a type of closeness centrality (Borgatti, 1995).

For the two nodes A and B in the network, there are unique shortest paths between them. Through the calculation of all the shortest of any two nodes, when some of these shortest paths pass through a node, then the node's betweenness centrality is considered high. If a node often appears in the shortest distance path between other nodes, the path with the shortest distance often contains this node. Then, the node is more capable of facilitating communication between other nodes. The formula is as follows. The denominator represents the sum of the shortest paths between all nodes, and the numerator represents the cumulative number of the shortest paths between all nodes.

$$C_B(P_i) = \sum_{j=1}^{N} \sum_{k=1}^{j-1} \frac{g_{jk}(P_i)}{gjk} \tag{2.20}$$

2.4.2 Web Service Discovery

Service matching is the key to service discovery. The performance of the service matching algorithm determines the performance of the service discovery. The objects and parameters of the service matching will also affect the final results. These two aspects are the two major entry points and the research focus of Web services discovery.

Rule-Based Reasoning Method

Since the service registry can only support keyword-based Web service retrieval, the conditions of which are too simple. In most cases, the user's needs cannot be accurately expressed. Therefore, the retrieval efficiency is generally low. Meanwhile, the retrieval result accuracy is relatively poor. To solve such problems, many methods describe the service request by constructing rules or logic, so that it can more clearly and accurately describe the specific content of the service request. Then, based on these rules or logic, the automatic reasoning of service discovery process can be realized.

Based on ontology and computational logic, an inference engine is proposed to solve the problem of service discovery (Alberti et al., 2011). The inference engine is primarily directed to behavioral interfaces between service publications and service requests to establish inference rules and plans. In order to reduce the complexity of service discovery. García, Ruiz and Ruiz-Cortés (2012) added a pre-process before matching the service. In this part, through programming rules, the unrelated service is first filtered out in the candidate service set, thereby narrowing the scope of service discovery and improving the accuracy of service discovery.

The performance of rule-based reasoning method relies intensively on their rules and reasoning logic. Ideally, such methods can effectively automate the service discovery process and achieve satisfactory service discovery accuracy. However, how to design effective rules and reasonable logical reasoning is often difficult. These rules and logic are generally more complicated. Execution in the service discovery process requires additional time and computational overhead, resulting in that the performance sometimes does not meet the expectations.

Semantics-Based Method

Compared with rule-based reasoning service discovery, the service discovery method using semantic technology can improve the automatic reasoning ability in the service discovery process. Meanwhile, the accuracy of the result is generally very high. Therefore, there are many methods enhancing semantic information in Web services and service requests to improve the performance of service discovery. Through the typical means of semantic annotation and semantic extension, the semantic information of service providers and service consumers is increased to highlight the functional attributes. Thus, the expression is more targeted and the semantic features are more obvious. The accuracy of service matching can thus improved.

Paliwal, Shafiq, Vaidya, Xiong and Adam(2011) proposed a service discovery method based on semantic automatic annotation. The core idea is to comprehensively utilize the multi-layer ontology concept and an improved service vector model to label and cluster services, which produce a semantic classification catalog of Web services similar to UDDI. This method mainly used the latent semantic index defined by it to enhance the semantics of the service request. Firstly, according to the service function parameters, the Web service in the semantic classification directory is initially filtered. Afterwards, the corresponding semantic similarity is calculated to match and finally obtain a sorted Web Service candidate list

Such methods generally focus on the description of services and requests. The method is relatively simple and has strong operability. However, the performance of such methods depends on the ontology while the building standards of the ontology are not uniform. Even if modelling the same domain, the performance between different ontologies will be different. Furthermore, the cost of building and maintaining the ontology library is high.

Quality-Based Reasoning Method

The matching objects considered by the foregoing two types of methods in their service matching are basically related to parameters, which are essentially functional variables of the candidate service, for example the input, output, preconditions and service effects of the service. During the process of service discovery, if only the function attribute of the service is used as the matching object, a batch of candidate services having same functions will be demonstrated in the result. It is difficult to select the best service from the lack of the parameter basis other than the function attribute. Thus, it is an inevitable choice to take the non-functional variables into account in addition to functional attributes to realize Web service discovery. Among them, service quality (QoS)-based service discovery is the focus of the research, which becomes a candidate for discovery and selection to better meet the needs of users.

Farzi, Akbari and Bushehrian (2017) designed and implemented a QoS method to support non-functional web services. The method is based on the OWL-S extension and builds a QoS metric model by adding the information needed to obtain non-functional parameters. The experimental results demonstrate that the proposed method can improve the accuracy of service discovery. To find the best web service, Rangarajan and Chandar (2017) proposed a formal client request message structure and service proxy architecture. First, the proxy architecture obtains the Web service requirements information with QoS requirements from the client. Then, it retrieves the Web services with similar functionality. Based on the QoS attributes confirmed by the agent, the agent will rank the candidate services according to an algorithmic mechanism. Similarly, Samir, Sarhan and Algergawy (2017) proposed a two-stage Web services discovery framework. The first phase is a feature matching method, which evaluates the similarities between a given set of Web services and provides related services based on the user's functional requirements. After that, based on the user context and non-functional requirements of

the related services generated from the first phase, the similarity between the context information of the two is calculated. According to the calculated similarity, a set of services satisfying the functional and non-functional requirements of the user is returned at the same time.

Compared with the above two methods, the service quality-based service discovery method focuses on the service sequencing or service selection phase after service matching. The algorithm principle of the service matching phase does not have great differences with the former two. Therefore, the service discovery accuracy of such methods is often higher than the former two. However, its computational overhead and computational complexity are also larger than the former two.

Graph-Based Method

Although adding semantic information in the service discovery process will improve the autonomy of the service discovery process and the accuracy of the results, the difficulty and complexity of the application are high since most semantic models and methods can only be applied to specific fields. Moreover, these methods usually require strict conditions in the matching process, so that the recall rate of the results tends to be low. In order to solve this problem, methods focusing on service interface logic relationships or dependencies between service parameters have attracted attention gradually. Such methods usually use bipartite graphs as the major tool for the analysis of the problems

J. Zhou et al. (2008) introduced a clustering and bipartite graph matching approach in service discovery using the space vector model to represent services. A special clustering algorithm is designed for cluster services. In the service discovery process, the optimal matching idea of the weighted bipartite graph is used to match the functional attributes of the service. The weights of the bipartite graphs are calculated based on the semantic similarity between concepts. This paper also presents a discussion regarding how to construct a weighted bipartite graph problem that satisfies the optimal matching

condition.

The graph-based service discovery method requires loose matching conditions in the matching process. The complexity of the method is relatively low. It is simple to apply compared with the previous methods. This method achieves an ideal recall rate and accuracy rate in service discovery efficiency. However, such methods still have shortcomings in terms of service quality.

2.5 Network Model Construction Strategies

In this section, we introduce the main characteristics of network nodes for BA model and PSO model which are popularity and similarity. The previous work has guidance for our work.

2.5.1 Popularity-Based Model

Traditionally, Web-API networks construction is based on the popularity of its nodes. As mentioned in previous section, nodes with high degree distribution are more likely attracted when a new node is joined in the network. There have been many scholars studying and investigating service ecosystems for model construction. Most of these research based on $ProgrammableWeb$ (http://www.programmableweb.com), which list the API information and details of mashup. According to data source, M. Weiss and Gangadharan (2010) examined the structure of the mashup ecosystem and how it grows as time goes by. They extract the API and mashup data from the site and define relationships between them. An affiliation network was created and the edges of the network indicate which APIs are used in which mashups (Uzzi, Amaral & Reed-Tsochas, 2007). They found that the degree distribution of mashups follows the power-law. However, when a new node is added and connected with the network, its growth is linear, which means not all APIs are used in application niches and only a

few nodes provide the basis. Their observation shows that there is a complementary relationship between APIs, depending on their location in the entire ecosystem. For one API, the more existing number of mashups it currently contributes, the higher propensity when compared to another API in the same mashup.

Moreover,K. Huang, Fan and Tan (2012) also studied service ecosystems based on $ProgrammableWeb$ mashup data. They collected information details from Programmable for each service and composition, including compositions, providers and the creation time of them. By sorting out the creation times, authors can establish a specific relation between compositions. Two network models were constructed: one is a composition- service network, which is based on relations between nodes and the other one is service-provider network. As they formalizing the network with a matrix (Tan, Zhang & Foster, 2010), it is found that services with higher popularity are more concentrated while the results is consistent with the ones obtained from the above research.

Fallatah, Bentahar and Asl (2014) used a three-step engineering method to build web service social networks. Besides web services, they also focus on users. After constructing the network model based on functionality and QoS of nodes, they also established between users and services. Popularity is one of the important metrics they identified for analyzing the network.

$$P_{ui} = \frac{|u_i \rightarrow u|}{|U|} \tag{2.21}$$

As [Equation 2.21], the popularity of a user is calculated by how many users have been connected to that user where numerator is the set that other users have been connected to the target user and U is the set of all users. Similar to users' nodes, the popularity of service nodes is also defined by the number of other nodes which has relation to that

service [Equation 2.22].

$$P_{wsi} = \frac{|ws_i \rightarrow ws|}{|W|} \tag{2.22}$$

Authors built various links, including user to user, user to service and service to service which helped to simulate the nodes behavior. Initially, nodes were not fully connected within the network. When user nodes begin to request service, the social network grows fast, both between users and services. Their approach is to improve the advertising and discovery of Web services through the merge of the users and the web services into active components of the global social network. The benefit of such an approach is that it helps to solve many problems of web services. According to their simulation result, the exposure for web services is much wider. It is more efficient and easier when users' nodes try to find and request web services. However, the social network analysis of characteristics was not discussed. Meanwhile, one of the limitations of the literature is that the simulated data is not collected from the real world.

Wang, Feng, Chen, Xu and Sui (2010) built a service network according to service classification and annotation. APIs are classified into specific domains. An annotation method is proposed to construct abstract service layer and to introduce semantic information into service network. During the process of network construction, authors firstly used tf-idf algorithms to extract service features and calculate vector distance (Ramos et al., 2003) to classify each nodes into different domains. Then, Web service data is explained using domain ontologies. A model was constructed from domain knowledge, which is a new aspect for building a service network. In this paper, the number of services was fixed while the network increase over time in real world.

2.5.2 Similarity-Based Model

In many real-world networks, its growth follows a preferred mechanism. Meanwhile, it can become more attractive when a node is more popular. However, popularity is only one aspect of scale-free networks while the other dimension is similarity (McPherson, Smith-Lovin & Cook, 2001; Şimşek & Jensen, 2008). For example, when an individual uses Facebook or Twitter in the web, in addition to those popular large websites and applications, others try to connect according to his or her preferences and interests as well, even if they are not so popular (Menczer, 2004, 2002). These circumstances indicate that there is a balance between popularity and similarity of nodes in complex network.

Pan, Li, Liu and Liang (2010) proposed an approach to detect the community structure with the node similarity through the iterative combination of the containing nodes that have the highest similarity for the discovery of the community structure, therefore establishing new community. Generally, it requires to measure the intensity between every pair of nodes on the basis of different methods to find the community structure, for instance edge betweenness (Newman & Girvan, 2004), edge clustering coefficient (Radicchi, Castellano, Cecconi, Loreto & Parisi, 2004), dissimilarity index (H. Zhou, 2003), information centrality (Fortunato, Latora & Marchiori, 2004), similarity based on random walks (Pons & Latapy, 2005) and clustering centrality (Yang & Liu, 2008) while the approach they proposed has relatively low computation complexity given that only the network's local information is required and it does not need any previous knowledge regarding the community. It is demonstrated from the simulation results that the approach can be used for the more effective detection of the community structure within the complex networks in comparison with traditional algorithms.

Lü, Jin and Zhou (2009) proposed a network model with controllable density and noise intensity when generating links. Through data collection from six real networks,

the similarity index based on local paths is used to estimate the possibility of links between two nodes. By comparing the simulated network with the real network, the local path index has high efficiency, which solves some problems of link mining and miss link prediction (Ifrim, Bakir & Weikum, 2008).

Papadopoulos, Kitsak, Serrano, Boguná and Krioukov (2012) established a framework to illustrate the correlation between popularity and similarity with geometric interpretation. The way they build the model is to control the other variables to be invariant and randomly place the nodes in a circle. Some measurement methods, such as cosine similarity (Crandall, Cosley, Huttenlocher, Kleinberg & Suri, 2008), calculate the angular distance between nodes to represent the similarity between the two nodes.

Chapter 3

Research Method

As mentioned in Chapter 1, the main research of this paper is to construct complex Web-API network based on BA method and PSO method. This chapter list and discuss the procedure of our method to construct a Web-API social network, including data acquisition, model construction and model fitting.

We show that how we collect the empirical data and list the result of data acquisition. Also, we detailed the tools (networks, numpy, NTLK), algorithms (PA, RWR, TF-IDF, etc), theoretical knowledge, and models we constructed to achieve the goals.

For BA model, the main procedures including growth and preferential attachment, while procedures for PSO model construction includes popularity extraction, similarity estimation, matrix normalization and dimensionality reduction. As the result of network construction, we mapped the data for a better visualization.

3.1 Data Acquisition

$ProgrammableWeb$ is a service directory website collecting and managing detailed information about web APIs and mashups. It has documented over 17000 open web APIs and 6000 of mashups. This platform helps people to invoke and create APIs even though they do not require professional web service development knowledge. The $ProgrammableWeb$ data set has been used in many previous researches about API-networks and mashup development (Elmeleegy, Ivan, Akkiraju & Goodwin, 2008; Jhingran, 2006; Lyu et al., 2014; W. Chen et al., 2013).

Since the database of $ProgrammableWeb$ is not open to the public, we crawled data from its webpages. A web crawler is a web robot used to systematically browse the World Wide Web, the purpose of which is generally to compile a network index (Chau, Pandit, Wang & Faloutsos, 2007). Sites such as online search engines update their own website content or their index to other websites through crawler software. Web crawlers can save the pages they visit so that search engines can generate indexes for users to search afterwards (Cho, 2001). With the help of crawling program , we automatically grabbed valuable information from the Internet which eliminated the need for cumbersome manual steps.

Table 3.1: ProgrammableWeb Dataset Overview (as of July, 2018)

Data Type	**Amount**
API	17952
Mashup	6252

As the result of data crawling, we used the number of mashups to represent degree distribution for each API. A mashup is a new network phenomenon on the Internet

today. It combines two or more web applications using public or private databases to form an integrated application. Through the secondary development of the developed API, Mashup can quickly and easily develop new applications. Developers can show these mashups or these new APIs can be released to the web at any time, which forms an ecosystem. If an API is included in many mashups, its degree is considered as high.

According records of $ProgrammableWeb$, [Table. 3.2] lists most used Web-APIs. The output of data scraping is structured into two data files respectively as described in [Table. 3.3].

Table 3.2: Most popular Web-APIs

API Name	**Number of Links**
FaceBook	3523
Google Maps	2952
Twitter	1939
Youtube	1515
AccuWeather	1490

3.2 Tools

In this section, we mainly introduce the tools in this work for data analysis and Wen-API networks construction, including NetworkX, Numpy and Natural Language Toolkit (NTLK).

Table 3.3: Web Service Data Structures

Data Description	Data Items
Web APIs (api.csv)	Web API Name
	Description
	Publishing Date
	Category
	Degrees
Mashups (mashups.csv)	Mashup Name
	Description
	Publishing Date
	Category

3.2.1 NetworkX

Based on our method, the NetworkX module has bene used for the construction of the complex network models. NetworkX is a Python software package for the creation, manipulation of complex networks and also for the learning of the dynamics, structure and functions of complex networks (Hagberg, Swart & S Chult, 2008). The NetworkX developers will be able to load or keep the networks in standard or non-standard data formats. It can produce many types of classic or random networks, explore the network mechanisms, establish the network models, design new network algorithms, draw networks, etc.

3.2.2 Numpy

Numerical Python (NumPy) is used as a type of extension library for the Python language, which supports a variety of dimensional arrays and also the matrix operations (Van Der Walt, Colbert & Varoquaux, 2011). It also offers a considerable library of mathematical purposes for the operations of array

Ascher et al. (2001)first developed NumPy's predecessor, Numeric. In 2005, Oliphant did the combination of Numarray with another library with the same nature in Numeric and developed NumPy with other extensions. NumPy serves as an open source and is kept by many collaborators.

NumPy is a very fast math library, mainly used for array calculations, including:

- Powerful N-dimensional array object
- Broadcast function
- Tools to integrate C/C++/Fortran code
- Linear algebra, Fourier transform, random number generation, etc.

3.2.3 Natural Language Toolkit

NLTK serves as an effective Python building platform to process the human natural language statistics. It offers a very easy-to-use interface providing access to more than 50 corpora and vocabulary resources (such as WordNet), as well as a set of text processing libraries for classification, tokenization, stemming, parsing, and semantic reasoning, as well as industrial grade (Loper & Bird, 2002). NLTK is made ideal for the industry users, researchers, educators, students, engineers and linguists through comprehensive API documentation and the manual programming guides. NLTK can be used on Windows, Mac OS X and Linux systems. Best of all, NLTK is a free, open source, community-driven project.

3.3 Framework

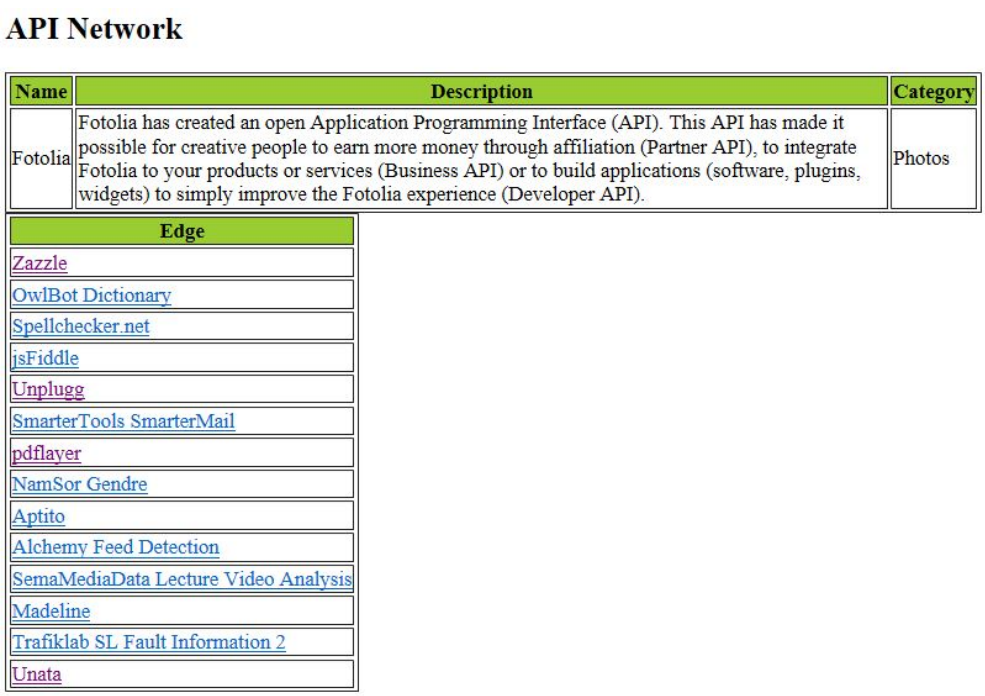
To make it easier to measure and compare network models in different construction methods in the future, we built a web page with data information. According to the results of data crawling, each API and mashup also has an isolate XML format. After transforming these files into HTML format with a better visualization, we invoked the Google custom search service to search the content of this website

3.3.1 Extensible Stylesheet Language Transformations

Since the original data type is XML, which does not fit the website, we used Extensible Stylesheet Language Transformations (XSLT) method to transform XML files to HTML type. As long as a XML and XSLT interpretation engine exist, XSLT is used to convert XML into HTML or other documents in any language. Meanwhile, different languages have bene used without affecting the result which is language-independent (Kay, 2001). The web browser will automatically convert an XML file if an XSLT file is imported. According to the XSLT template, the results of the transformation include two tables in web pages. One is the basic information of APIs or mashups and the other shows and links to other page which service is connected to it [Figure 3.1].

3.3.2 Custom Search Engine

Google Custom Search Engine (CSE) is applied for the searching of at least one website. After the installation of the module, the Drupal module is configured by using CSE by entering the CSE unique ID of Google. Once one or more roles are authorized for the searching by using Google CSE, users can search the current Google CSE index from Search/Google or Search Blocks. Google CSE uses Google's website index. Content changes are not demonstrated immediately through the score search outcomes of Drupal.



API Network

Name	Description	Category
Fotolia	Fotolia has created an open Application Programming Interface (API). This API has made it possible for creative people to earn more money through affiliation (Partner API), to integrate Fotolia to your products or services (Business API) or to build applications (software, plugins, widgets) to simply improve the Fotolia experience (Developer API).	Photos

Edge
Zazzle
OwlBot Dictionary
Spellchecker.net
jsFiddle
Unplugg
SmarterTools SmarterMail
pdflayer
NamSor Gendre
Aptito
Alchemy Feed Detection
SemaMediaData Lecture Video Analysis
Madeline
Trafiklab SL Fault Information 2
Unata

Figure 3.1: Illustration of XSLT result

Google provides a free registration page and CSE is created according to the following steps:

1. Sign up for a Google account
2. Go to the page of the custom search engine, $https://cse.google.com/cse$
3. Fill in the basic configuration of the search engine, including the domain of website, language and CSE name.
4. Check the code of CSE, it also expose the URL. Get this code to embed the CSE box into personal defined web page.

There are two approach to called the CSE service:

- Customize an html page and copy the following code into the div
- Request CSE service through the url

3.4 API-Mashup Affiliation Network

An affiliation network is an important form of network representation in complex networks. Many networks in the real world present a natural dichotomy, such as audiences and music groups, football teams and players, scientists and essays, and so on. An affiliation network is composed by two kinds of nodes, and edges exist only between different types of nodes. A series of cooperative networks in nature and society can be described as an affiliation network composed of cooperation subjects and cooperative affairs. The affiliation network is universal and has become an important target for complex network research. In the existing research work on the binary network, the usual practice is applied for the projection of the affiliation network into a single vertex network and then perform network analysis.

A bipartite graph (West et al., 1996) 1996) is assumed to be a very unique model in the graph theory. Its vertices can be classified into two disjoint subsets. There are two vertices associated with every edge in the graph belong to these different set of vertices (Fern & Brodley, 2004). Bipartite graphs have many applications in complex network analysis, such as urban social networks (Eubank et al., 2004), ecological networks (Dormann, Gruber & Fründ, 2008) and the human disease network (Goh et al., 2007).

Given $G = (M, A)$, where M is the set of mashups and A is the set of APIs. The sufficient and necessary situation for the undirected graph G to be a bipartite graph is that G has more than two vertices and all of the loops have even length. For any edge (m,a) in affiliation network, $m \in M$ and $a \in A$ [Figure 3.2].

3.5 Popularity-Based Model

As mentioned in Section 2.5, growth and PA are key characteristics in the BA model. In this paper, the construction strategy of popularity-based API network model is based on

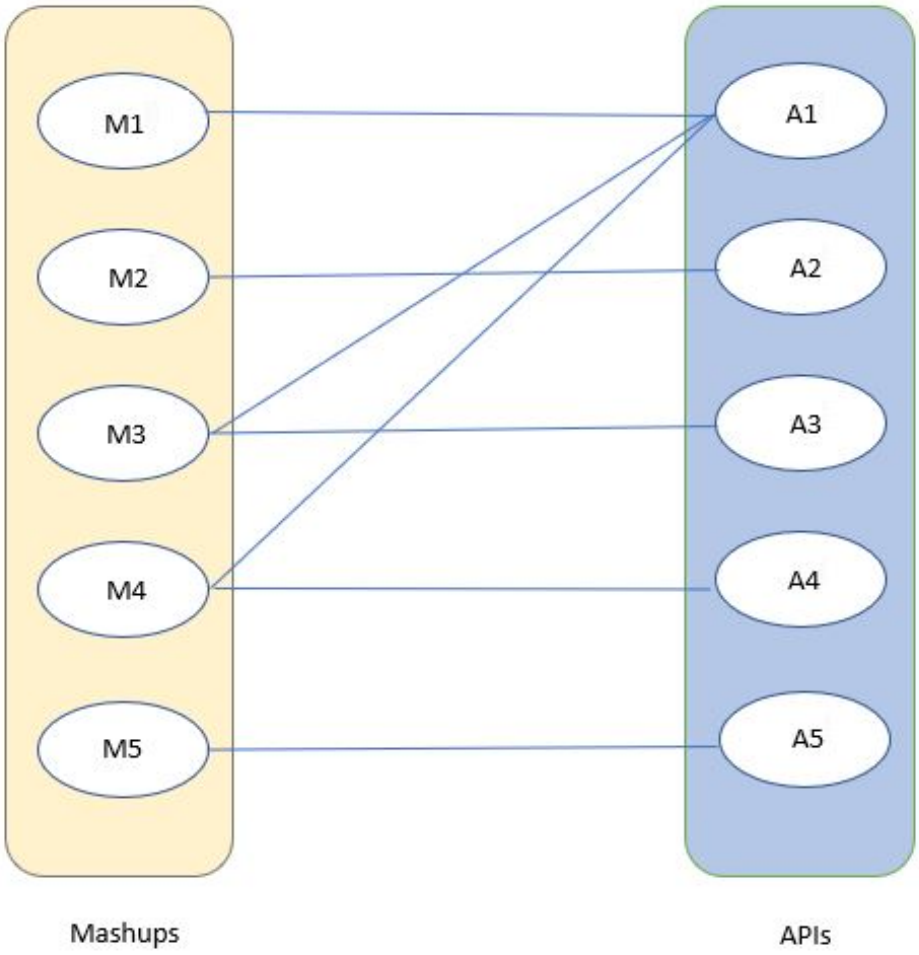


Figure 3.2: Illustration of affiliation network bipartite graph

the BA model.

3.5.1 Growth

In the initial network, it starts with fully connected node $m_0 = 4$. In order to express growth feature in model, new nodes will continue to join with m links at every step. At each time point, a new node with m links joined the network and connected to existing nodes, where $m_0 \geqslant m$.

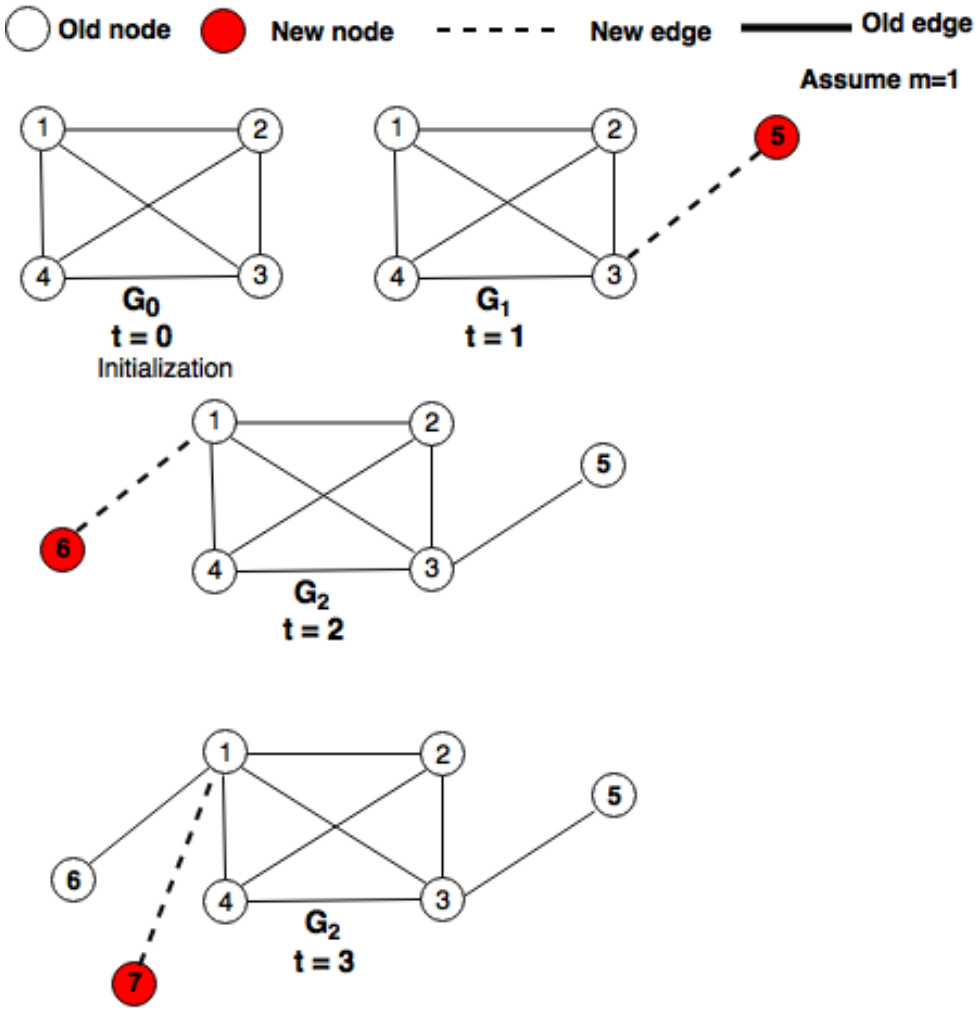


Figure 3.3: Web-API network growth procedure

3.5.2 Preferential Attachment

As for PA, we use the linear-PA equation to dynamically estimate the probability that a new node will connect to existing node i, depending on its degree k_i. According to the degree k_i of existing node, new nodes connected them with the probability Π_{ki}. For instance [Figure 3.3)], given that there are four fully connected nodes in the initial network $t = 0$. When a node is added ($t = 1$), the probability of new node connect to them is same since node 1-4 has the same degree, which is calculated by the degree of existing node over total degree of the network P. So when $t = 2$, the probabilities for

node 1-5 to attract the node are $3/P$, $3/P$, $4/P$, $3/P$ and $1/P$.

3.5.3 Data Mapping

After building the model, we need to map the real service data to the nodes since nodes are serialized without API name. Meanwhile, we add edges into its XML files for each API service to display links between them. To achieve these goals, the numbers of nodes in BA model and API services should be consistent. Afterwards, we got the result for a BA model, which is associated with real-world data.

The procedure is described below:

1. First, we sort APIs according to their degree and publishing date in a descending order. Nodes which have higher degree in the front to the list. For those APIs having the same degree, the earlier published more front.

2. Then, we associate the data of the list with nodes in the model in order.

3. Finally, traversing every node and every edge in the model and add edges information for each side to the API file.

An overview of Web-API network constructed by BA model is shown with popular nodes labeled [Figure 3.4].

3.6 Similarity-Based Model

As far as we know, in addition to the degree of the node, which represents the popularity, the similarity between nodes affects the structure of the network in the real network. For the next step of the API network construction, this element is added to build the model.

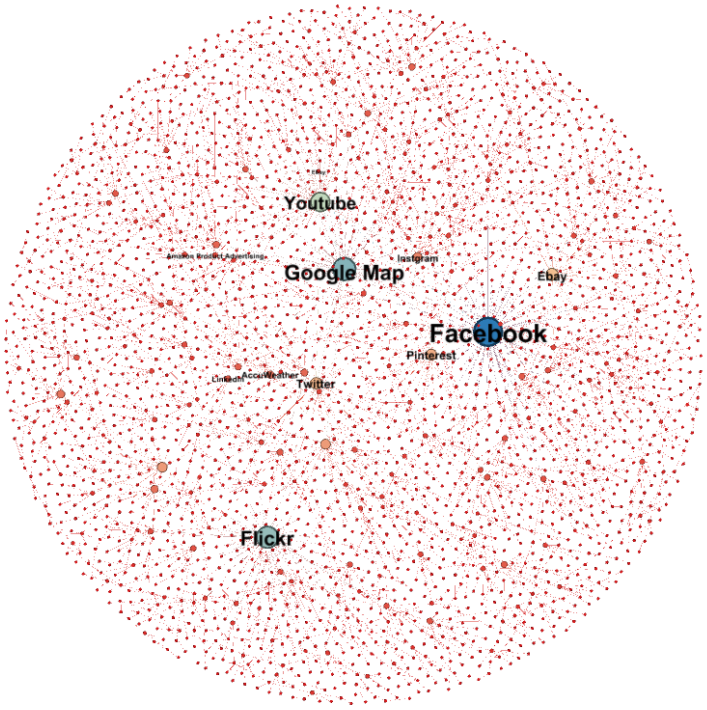


Figure 3.4: Overview of the Web-API network constructed by BB model

The approach taken in this paper is to initialize an empty network model first. As time goes on, new nodes appear and are placed in a circle, which represents the area of similarity with random angular position θ_t. The angular distance between the nodes represents the similarity between them. In this thesis, we choose Random Walk with Restart (RWR) algorithm and cosine similarity metric to estimate the similarity between nodes. Then we connect the new node t to the previous existing node s to generate a new network. The subset s consist of the m nodes with the m smallest values of product $s\theta_{st}$. The parameter m is used to control the average node degree $k = 2m$ and θst is the angular distance between nodes t and s.

For geographical interpretation, the distance of all nodes lies on a plane with polar coordinate (r_t, θ_t) and (r_s, θ_s) is hyperbolic, where

$$x_{st} = r_s + r_t + \ln(\theta_{st}/2) = \ln(st\theta_{st}/2) \tag{3.1}$$

This hyperbolic distance is used to represent the combination of two important parameters of complex networks, radial popularity and angular similarity.

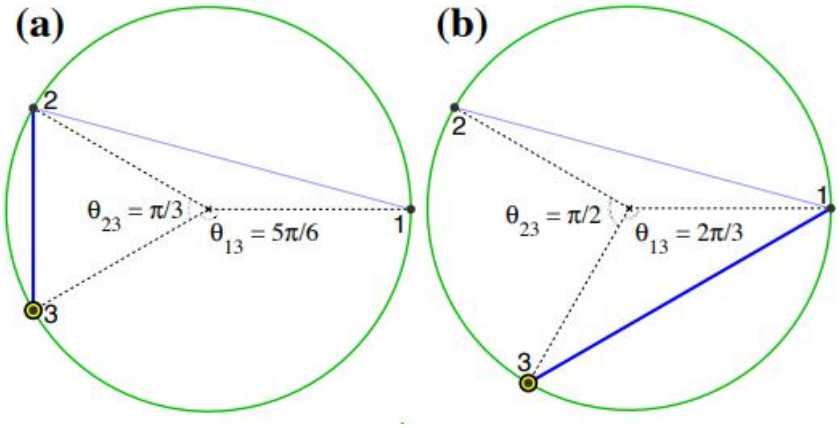


Figure 3.5: Geometric of popularity & similarity

In figure 3.4, t = 3 and m = 1, node3 connects to node2 in (a) since $2\theta_{23} = 2\pi/3 < 1\theta_{13} = 5\pi/6$ and node3 connects to node1 in (b) because $1\theta_{13} = 2\pi/3 < 2\theta_{23} = \pi$

3.6.1 Popularity Extraction

The APIs are ordered according to their degree. Nodes with higher degree are more popular. For these nodes who has same degree the earlier published date of time has a higher ranking.

3.6.2 Similarity Estimation

The methods we used in this thesis to estimate nodes similarity are random walk with restart (RWR) algorithm and cosine similarity.

RWR

The fundamental concept of random walk algorithm is to change the graph from one or a series of vertices. On any vertex, the traversal will walk to the neighbor vertices of this vertex with probability $1\ a$, and randomly jump to any vertex in the graph with probability p, which is called a jump probability of occurrence (G. H. Weiss & Rubin, 1982). After every walk, it can obtain a probability distribution. The probability is plotted that every vertex in the graph is accessed. Using such probability distribution as a type of input for the next walk and iterate through the whole process. This probability distribution tends to converge when certain preconditions are met (Newman, 2005). After convergence, a smooth probability distribution can be obtained. The random walk model has been applied extensively in data mining and Internet. The PageRank algorithm (Xing & Ghorbani, 2004) can be regarded as an example of a random walk model.

RWR algorithm is an improvement on the basis of the random walk algorithm. Starting from a node in the graph, there are two choices for each step, including randomly selecting neighboring nodes, or returning to the starting node. This algorithm includes a parameter a for the restart probability, and 1 a for the probability of moving to the adjacent node (Tong, Faloutsos & Pan, 2008). After the iteration reaches the stationary state, the probability distribution obtained after the smoothing can be regarded as the distribution affected by the start node (Cowles & Carlin, 1996). RWR can capture the multi-faceted relationship between the two nodes and capture the overall structural information of the graph. RWR can be defined as the following equation:

$$\overrightarrow{r_i} = c\widetilde{W}\overrightarrow{r_1} + (1-c)\overrightarrow{e_i} \tag{3.2}$$

where W = $[w_{i,j}]$ is a weighted graph and $\widetilde{W}$ is a normalized matrix of W. c is the restart probability. Correlation between two nodes obtained by RWR can capture global

structural information compared to standards that measure only two points. Compared to traditional methods of calculating the distance on the graph (such as shortest path, maximum traffic, etc.), it can capture multiple aspects of information on two nodes.

We use RWR algorithm to measure service correlation based on a Mashup-Web APIs interaction in the network and save the discrete similarity value of each Web-APIs with respect to another in an adjacency matrix.

Cosine Similarity

The cosine similarity applies the cosine of angles of the two vectors of the vector space as the measure of the difference between the two individuals (Tata & Patel, 2007). In comparison with distance metrics, cosine similarity attaches more attention over the difference of direction between two vectors, instead of distance or length.

$$sim(X,Y) = \cos\theta \frac{\overrightarrow{x} \cdot \overrightarrow{y}}{\| x \| \cdot \| y \|} \tag{3.3}$$

Being same with the Euclidean distance, the calculation approach on the basis of cosine similarity also takes the preference of users as a point in the n-dimensional coordinate (Amatriain, Jaimes, Oliver & Pujol, 2011). By connecting points to form a vector with the origin of the coordinate, the value of similarity between the two points is the cosine of the angle, which is smaller than the angle. The more similar these two vectors are and the larger the angle is, the less similarity there will be (Qian, Sural, Gu & Pramanik, 2004). According to the calculation method and measurement characteristics of Euclidean distance and cosine similarity, they are applicable to various data analysis models: Euclidean distance can reflect the absolute difference of individual numerical features, so more is used in the numerical value of the dimension and analysis of differences, such as using user behavior indicators to analyze the similarity or difference of user value; while cosine similarity is more to distinguish the difference from the

direction, but not sensitive to the absolute value (A. Huang, 2008).

To estimate how similar between two services, we extract key information about API description, which is related with word stemming (Jivani et al., 2011), we calculate the cosine similarity with the use of term frequency–inverse document frequency (TF-IDF) algorithms (Ramos et al., 2003).

Word stemming is the process of removing the prefixes and suffixes to get the word roots by segmenting the API service descriptions into terms. For example, 'connects', 'connected' and 'connection' should be all stemmed and regarded as 'connect', which refers to any different form of a word sharing the same frequency in the calculation. Moreover, in order to get the result with higher accuracy, prepositions with little meaning should be eliminated, such as 'a', 'in' and 'the'.

As for TF-IDF, it is extensively applied for the weighting technique for information mining and information retrieval. TF-IDF is a statistical approach applied for the evaluation of the significance of a word for the file set or at least one file of the corpus (Aizawa, 2003). The significance of a world has proportional increase with the number of times it shows up in the file. However, it also reduces inversely with frequency it shows up in corpus (Vijayarani, Ilamathi & Nithya, 2015). There are a variety of forms of TF-IDF weighting being used by search engines as a rating or measure about the level of correlation between user queries and files.

$$tfidf_{i,j} = tf_{i,j} \times idf_{i,j} \tag{3.4}$$

In a specific document, frequency means to appear times of the word in the document. Such frequency refers to the normalization of the term count to avoid it from the bias regarding long files, which has the same amount of words in a long file than a short file (Turney & Pantel, 2010), regardless of if the word is significant or not. For a word ti in

a particular file, its importance can be expressed as:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3.5}$$

where $n_{i,j}$ is occurrence times of the word in a document while the denominator refers to the sum of all the words' occurrences.

Inverse document frequency serves as the measure of the universal importance of a word. The IDF of a specific word can be realized by classifying the total volume of the files through number of files having the word. The obtained quotient is listed in below:

$$idf_i = \log \frac{\mid D \mid}{\mid j : t_i \in d_j \mid} \tag{3.6}$$

$\mid D \mid$ is the amount of files in corpus and denominator is the amount of files that target words appears.

The result of cosine similarity calculation with TF-IDF is a n×n matrix, n is number of APIs and each cell is the similarity of two services.

3.6.3 Matrix Normalization

For the multi-index evaluation mechanism, because of various natures of every assessment index, it usually has various dimensions and orders of magnitude. When there are great differences of the levels of the indicators and the analysis is conducted directly with the values of the original indicator, the function of the higher-value indicators of the comprehensive analysis will be emphasized. The impact of lower-level indexes will be weakened relatively. Thus, to guarantee the reliability of outcomes, the original indicator statistics need the process of standardization.

Before implementation of the data analysis, the data should be normalized and the data needs to be standardized for the analysis of data. There are two aspects of

data normalization process: dimensionless processing and homogenization processing (Al Shalabi, Shaaban & Kasasbeh, 2006). The trending process and the statistics mainly target on the data with different natures. The comprehensive outcomes of various forces cannot be reflected by the direct combination of various nature indicators. It is of great importance to first take into account of the changes of the nature of inverse index data to ensure all the indicators can be applied the same. The total is added to obtain correct outcomes. Data dimensionless processing mainly deals with the comparability of the statistics. After the standardization processing above, the original statistics can be converted to the dimensionless index assessment value, which means every index value has the same quantity level. The comprehensive assessment analysis can be implemented. The benefit of normalization is to enhance the accuracy, which has high efficiency in terms of some distance calculation algorithms.

The main process for API similarity result include:

1. Convert all symmetric normalized Laplacian values to positive values.
2. To ignore self similarity, set diagonal to 0.
3. Scale similarity values between (0, 1).
4. Convert to distance matrix , that is make the large number small similarity.

3.6.4 Isometric Feature Mapping Dimensionality Reduction

In this thesis, our method builds relationships between nodes based on two dimensions of popularity and similarity in a low dimensional space, which optimizes these two dimensions in hyperbolic spaces. Nevertheless, the result of the normalized matrix is a high dimensional similarity dataset. A high-dimension matrix is difficult to be visualized in the low space and the plots are much less intuitive. The reason why we mapping nodes to the simplest manifold is to get similar data points closer together.

Moreover, it is easier to measure latent distance between each data points to estimate how similar a point is with respect to others.

Manifold learning serves as a classical method for non-linear dimensionality reduction. The concept of manifold learning is that the statistics for observation is actually obtained from the low-dimensional manifold to a high-dimensional space. Because of the constrains of the internal features of the statistics, some high-dimensional statistics will lead to dimensional redundancy and actually only require a lower dimension to be represented specifically (McCallum, Nigam & Ungar, 2000).Imaging that the statistics is a low-dimensional manifold uniformly sampled in a high-dimensional Euclidean space, it is used for the recovery of the low-dimensional manifold structures of the high-dimensional sampled data, which implies to find low-dimensional manifolds in high-dimensional space, and corresponding embedded mapping for dimensional reduction or data visualization (Saul & Roweis, 2003).

Isometric feature mapping (ISOMAP) algorithm is applied for the reduction of the dimension of similarity matrix (Tenenbaum, De Silva & Langford, 2000).ISOMAP is used as one of the earliest methods for manifold learning and is applied in a variety of real applications. Isomap is developed on the basis of Multidimensional scaling (MDS) (Borg & Groenen, 2003)and constrains the key geometry of nonlinear statistics, ie the geodesic distance between any pair of points. The core algorithm is in line with MDS. The computation of the distance matrix of the original space leads to the differences. There are many statistics having non-linear structures and cannot fit to Principal Component Analysis (PCA) (Jolliffe, 2011) and MDS algorithms since only the geodesic distance reflects the true low-dimensional geometry of the manifold. In a nonlinear data structure, the geodesic distance between the two data points on the manifold could be far while the Euclidean distance in the high-dimensional space is close.

The Isomap algorithms which we used includes three main steps:

1. Construct neighborhood graph: Determine which points on the manifold M are adjacent, the distance between the two points (i, j) is represented by $D_x(i, j)$; i, j belong to the space X; the distance of $D_x(i, j)$ is defined as the distance of Euclidean (Roweis & Saul, 2000). The adjacency relationship can be set to a fixed radius e or K nearest neighbor. These neighborhood relations are represented as a weighted graph G over the data points, with edges of weight $D_X(i, j)$ between neighboring points.

2. Compute shortest path: Estimate the geodesic distance $D_M(i, j)$ on manifold M by calculating the shortest path $D_G(i, j)$ between two points on graph G.

3. Construct d-dimensional embedding: A classical MDS is used to construct an intrinsic embedded manifold geometry that retains the most complete d-dimensional Euclidean space Y.

3.6.5 Popularity-Similarity Network Construction

With the completion of above tasks, now we construct popularity-similarity network. We input N: Number of Nodes, gamma: Popularity fading controlling parameter and m: Number of links and network model is constructed with following procedures.

1. Determine the value of beta, a parameter controlling popularity fading

2. Initialize network

3. Import Raw sim.

4. Estimate Hyperbolic distance between new node and all existing node

5. Connect to the m-hyperbolically closest nodes.

An overview of Web-API network constructed by PS model is shown with popular nodes labeled [Figure 3.6].

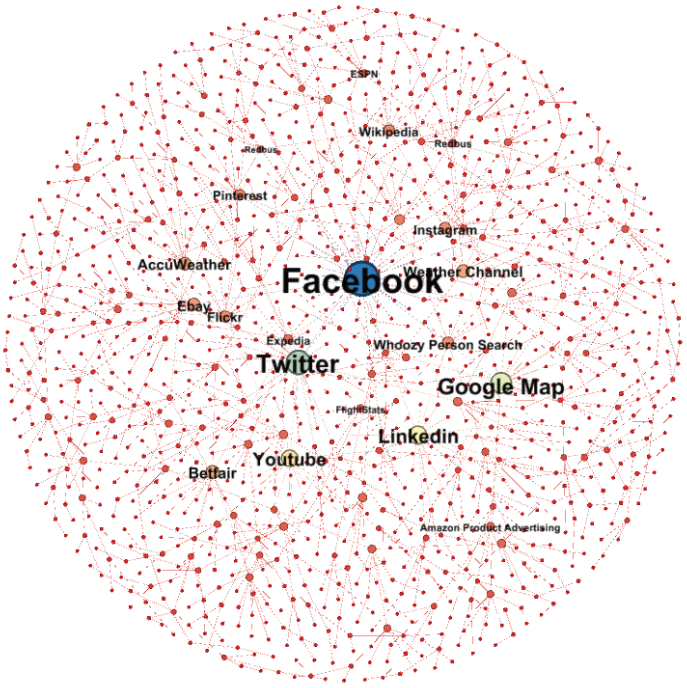


Figure 3.6: Overview of the Web-API network constructed by PS model

3.7 Data Fitting & Parameter Estimation

We already know that random variables, if they obey the power-law distribution, appear to be a straight line in the double logarithmic coordinates. However, the tail of the line fluctuates due to statistical errors. One way to solve the tail volatility is to remove the tail data and only fit the first half. However, doing so is a way to get rid of the important information contained in it. Thus the usual method is to calculate the cumulative distribution. The use of a cumulative distribution avoids the problem of determining the packing width and results in better use of all data without losing any information.

The cumulative distribution function (CDF) is the integral of the probability density function and can fully describe the probability distribution of a real random variable X. For continuous random variable:

$$F_x(X) = Pr(X \leq x) \tag{3.7}$$

while for discrete random variables, the CDF is a piecewise function.

There are many ways to fit the data and estimate the parameters, including calculating the cumulative distribution, Loglog, Log-binning, OLS, MLE, etc. The current view is that it is correct to consider the cumulative distribution function under double logarithmic coordinates. The exponent of the power law distribution is estimated by the maximum likelihood method. Finally, the KS test is performed. The basic steps are as follows:

1. Construct the complementary cumulative distribution function (CCDF), which is the cumulative distribution in statistical physics, the complement distribution of the cumulative distribution in mathematics.

2. The original distribution must also be studied since the CCDF is not perfect. It is difficult to judge because of the interval of the power law. Therefore, it is necessary to combine the original distribution to judge whether it is a power law or deviate from the root of the power law.

3. Plotting data in linear, semi-logarithmic, and double logarithmic coordinates to observe if distributions are following power-law [Table 3.4].

Table 3.4: Data Trending in various coordinates

	linear-linear coordinates	**linear-log coordinates**	**log-log coordinates**
power-law	downward convex curve	convex curve	straight line
Gaussian	bell curve	anti-parabola	concave with rapid decay
exponential	convex curve	straight line	concave curve

For example, if a distribution appears as an upward bulge in double logarithmic coordinates, it is not a strict power law distribution. As observed in the linear -logarithmic coordinates, if it is a straight line, it means an exponential distribution. It is noticed that do not just judge in a double logarithmic coordinate only because the distribution of straight lines in double logarithms is not just a power law. There might be other types of distribution.

4. For the power-law distribution, since the power law deviates when the variable tends to 0, it is necessary to determine the minimum value of the variable in accordance with the power law interval. The usual method is either visually judging k_{min} on the graph or drawing a scatter plot of the exponent and k_{min}, both of which are subject to noise and fluctuations. Therefore, the more accurate method is to perform KS estimation.

5. The parameter values are estimated using the maximum likelihood method. The MLE method estimates the parameter values by maximizing the likelihood function of the model, which is more scientific and reasonable.

Chapter 4

Analysis

According to methods we introduced in Chapter 3, this section illustrates the results of experiment.

First of all, Web-API networks are visualized, including an affiliation network, BA model ,and PS model. Then, we test models if they fit in the power-law distribution from the perspective of nodes degree distribution, power-law fitting and estimation of parameter k_{min} and γ. At last, we measure the PA to prove Web-API networks follows the linear-PA.

4.1 Visualization

The networks are visualized using the Force-Atlas 2 layout in gephi, which includes Mashup-API affiliation network [Figure 4.1] [Figure 4.2], and Web-API network constructed by BA model [Figure 4.3] and PS model [Figure 4.4].

Force-Atlas 2 is a force-directed layout typically used in social networking, which is a method producing a fairly beautiful network layout with fully displaying the overall structure of the network and its automorphism characteristics. Therefore, this method occupies a dominant position in the relevant literatures of network node layout technology. Force-directed layout imitates the gravity and repulsion of the physical world and automatically lays out until force balance. Force-Atlas 2 layout makes the graph more compact and readable. It shows that the centralized permission (attraction force) is larger than the hub, which automatically and steadily improves the cohesion of layout. The attraction force between two connected nodes depends on their linear distance, expressed as:

$$F_{\alpha}(n1, n2) = d(n1, n2) \tag{4.1}$$

Moreover, the consideration of nodes degree helps to reduce the visual cluttering since power-law distribution of degrees characterizes real-word data. The concept is to build closer connection between the poorly connected nodes and very connected ones. The solution is to change the repulsive force between the poorly connected ones and the very connected ones. Therefore, they can finally reach to a balance.

$$F_{r}(n1, n2) = k_{r}\frac{(deg(n1) + 1)(deg(n2) + 1)}{d(n1, n2)} \tag{4.2}$$

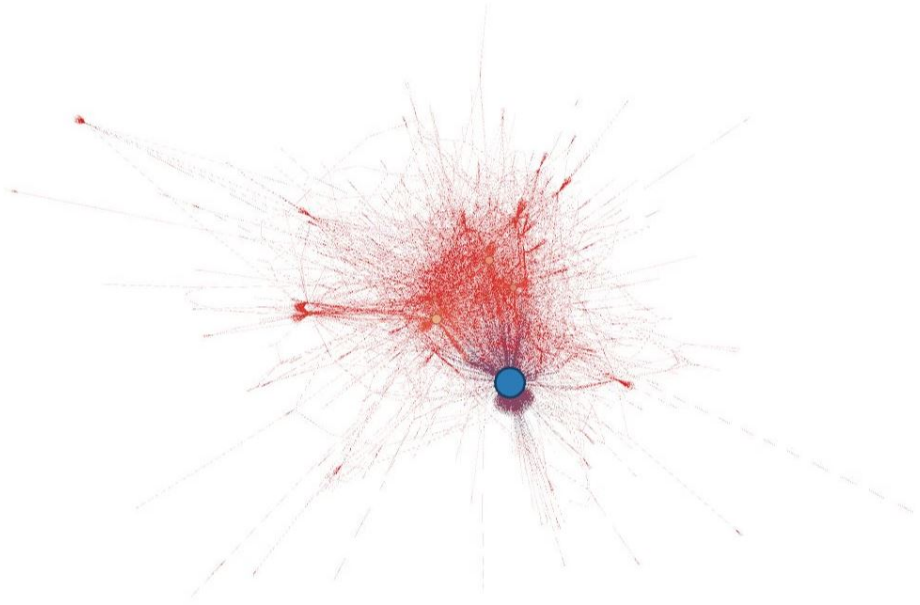


Figure 4.1: Visualization of the Mashup-API Affiliation Network

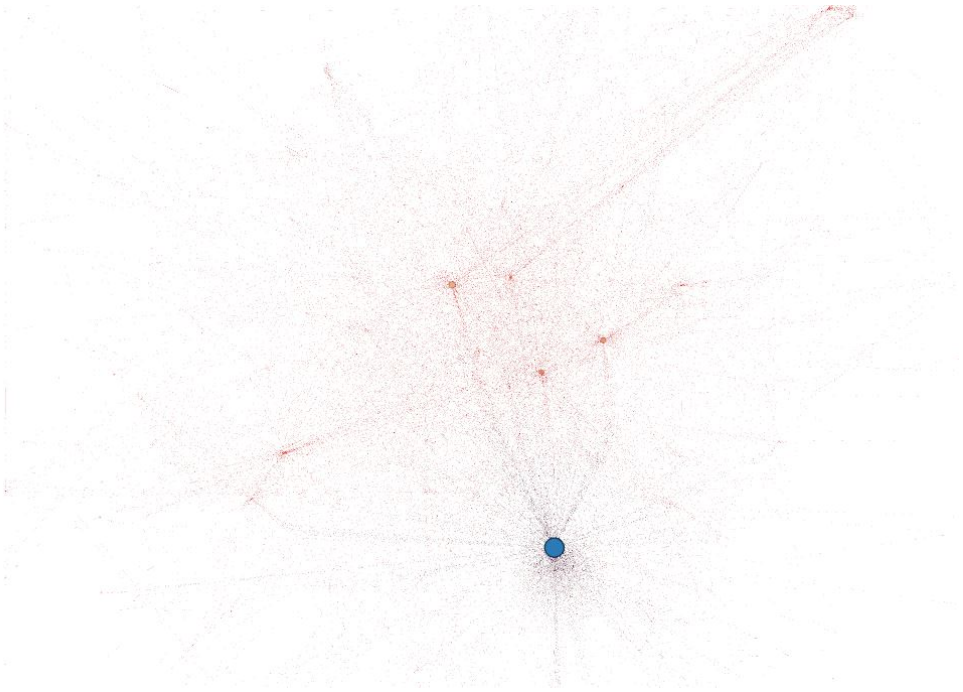


Figure 4.2: Visualization of the Mashup-API Affiliation Network(lin-log mode)

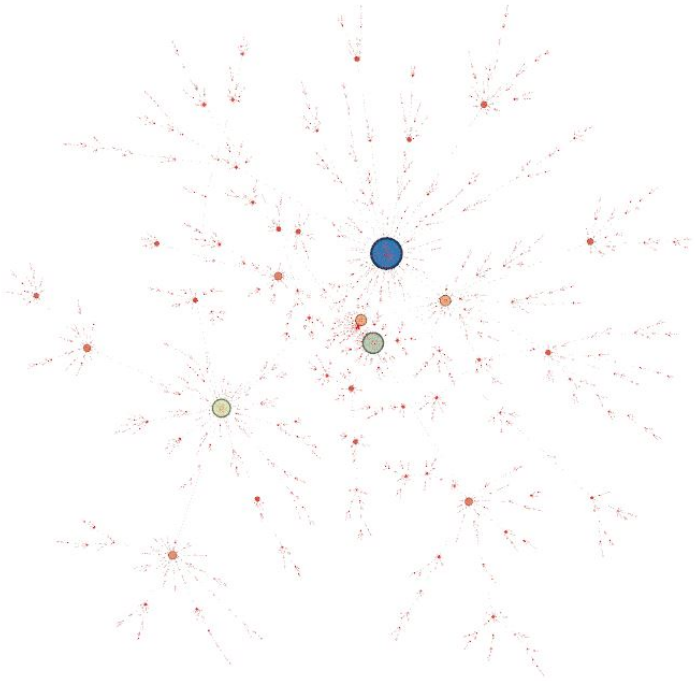


Figure 4.3: Visualization of the Web-API Network (BA model)

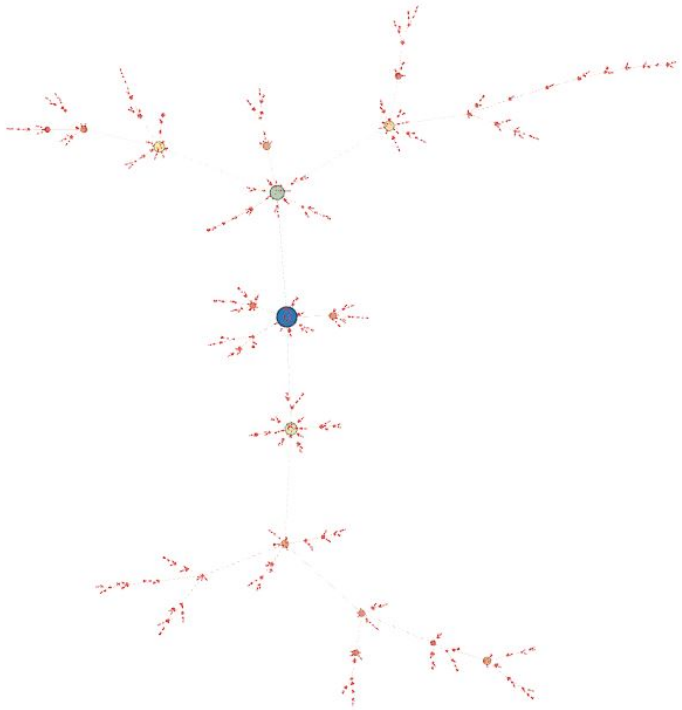


Figure 4.4: Visualization of the Web-API Network (PS model)

4.2 Nodes Degree Distribution

Based on the study of (Barabási et al., 2016) (2016), it proposed the power-law distribution, which include log-log scale with cumulative, log-log scale with log-binning, log-log scale with linear binning and linear scale with linear binning [Figure 4.5]. [Figure 4.6][Figure 4.7] [Figure 4.8] [Figure 4.9] illustrate the trend of the node degree distribution is consistent with the theory.

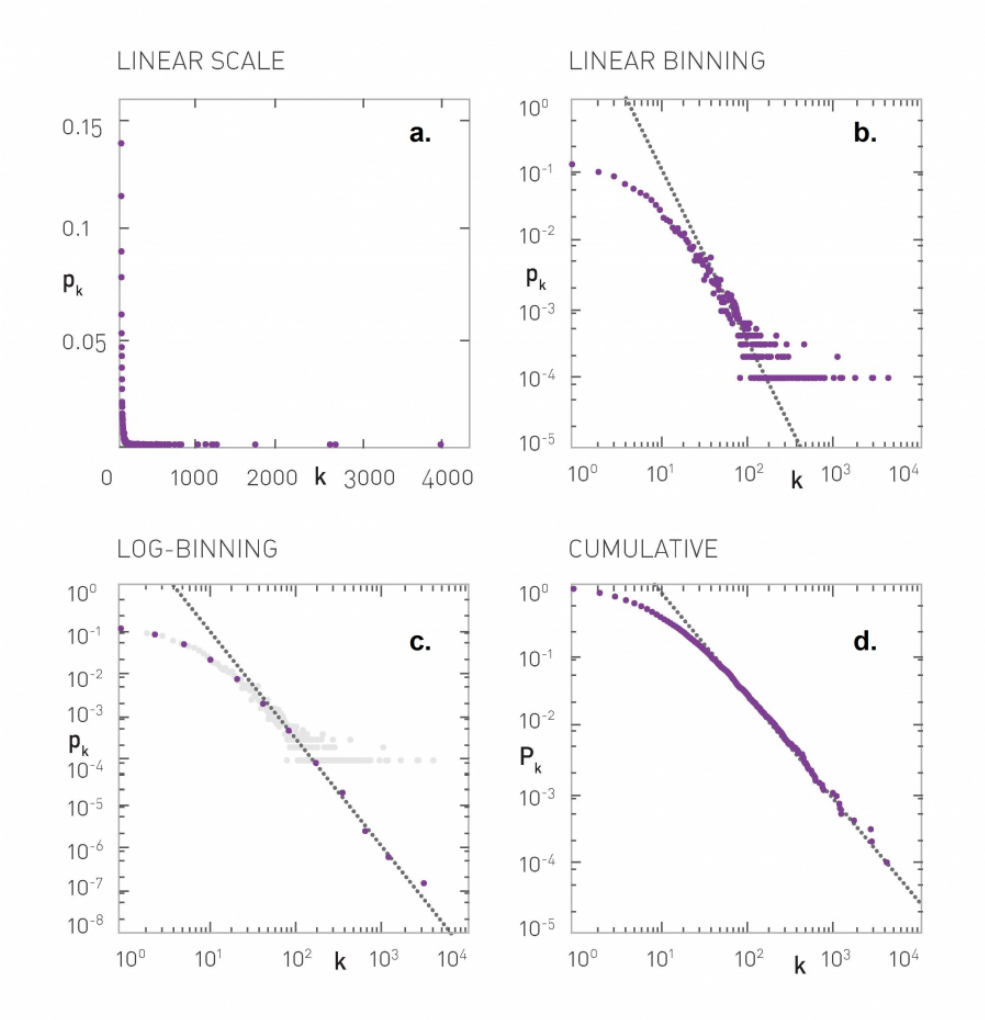


Figure 4.5: Power Law Plotting
Source, Barabási et al., 2016

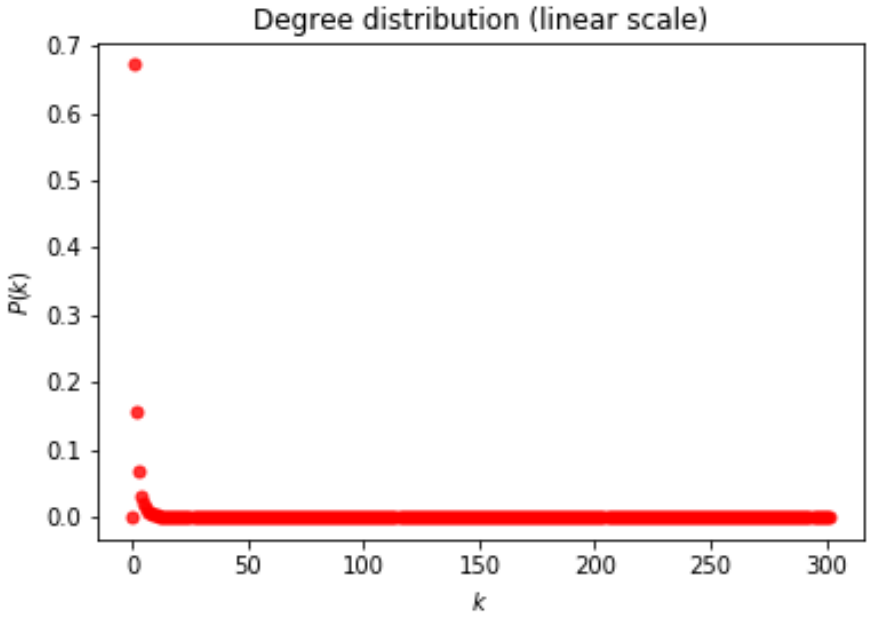


Figure 4.6: Popularity-Based Network Model Node Degree Distribution (linear scale)

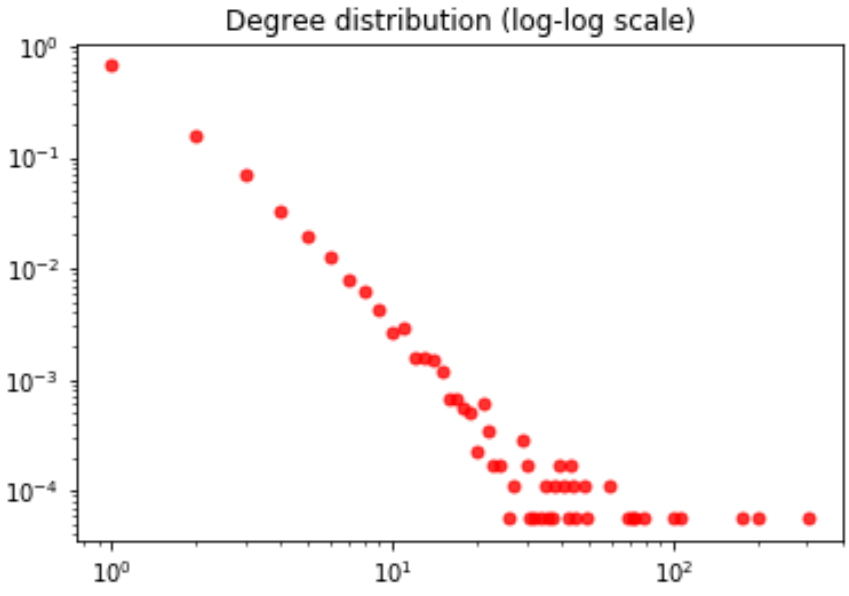


Figure 4.7: Popularity-Based Network Model Node Degree Distribution (log-log scale)

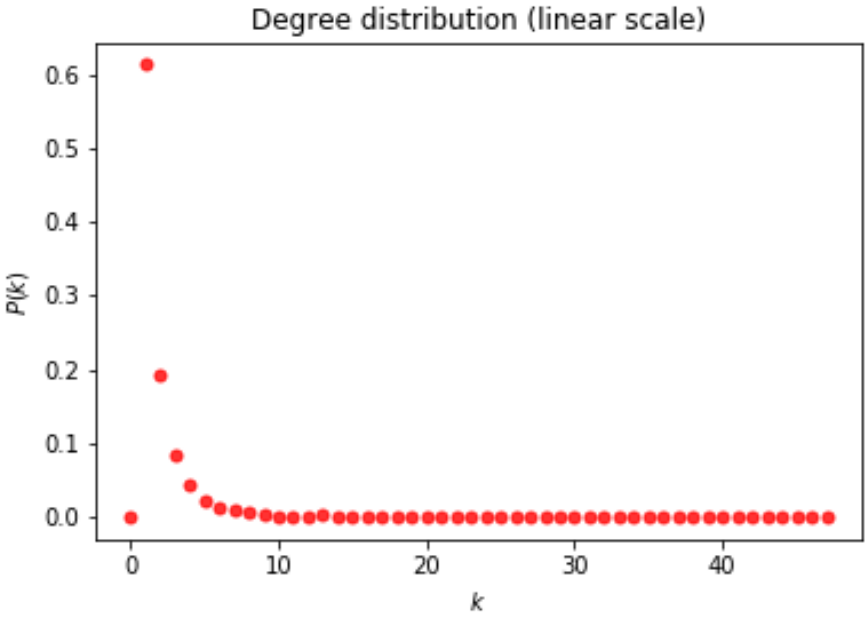


Figure 4.8: Similarity-Based Network Model Node Degree Distribution (linear scale)

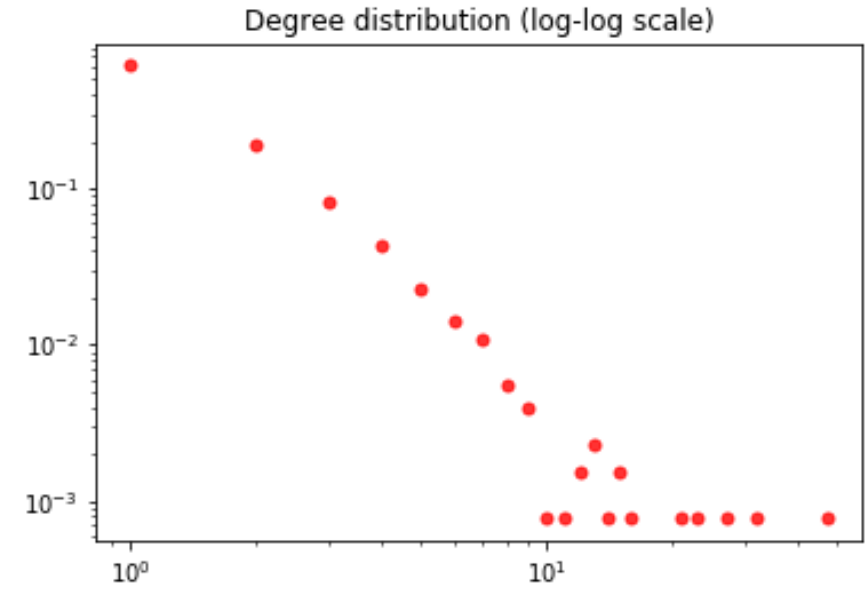


Figure 4.9: Similarity-Based Network Model Node Degree Distribution (log-log scale)

4.3 Power Law

The existence of power law distribution is very extensive and is of great significance for many scientific research problems. However, since the long tail of the power law

distribution has large fluctuations, the range of the determination of the long tail is particularly complex. The least squares method estimates the power law distribution with a large error. Even if there is no error, it is not possible to determine whether this distribution is a power law distribution because it is not compared with other distribution forms, such as exponential, log-normal and truncated power law. The comparison is whether or not a data conforms to one of the multiple distributions, usually using the KS test method.

Clauset, Shalizi and Newman (2009) proposed the idea of fitting power-law distribution data. This idea fits by using the maximum likelihood method and the results of the fit are evaluated using the KS test statistic and likelihood ratio. The research of power law fitting in this paper uses poweRlaw package to compare the curve of varies distribution forms to fit heavy-tailed distributions for both discrete and continuous power law distributions.

Continuous Power Law

Assuming $\alpha > 1$, the density function of the continuous power law is as follows:

$$p(k) = \frac{\alpha - 1}{k_{min}}(\frac{k}{k_{min}})^{-\alpha} \tag{4.3}$$

where k_{min} gives the lower bound of the k range, because $k^{-\alpha}$ tends to infinity when k approaches 0, which is not appear in reality.

Discrete Power Law

Similar to the continuous power law, the expression of the discrete power law of is:

$$p(k) = \frac{k^{-\alpha}}{\zeta(\alpha, k_{min})} \tag{4.4}$$

We fit power law distribution and observe if it is suitable for power law. As what is demonstrated in [Figure 4.10] [Figure 4.11], four classical models was chosen to fit the data. Exponential and the Poisson fit poorly to the network, while power-law and log-normal is good.

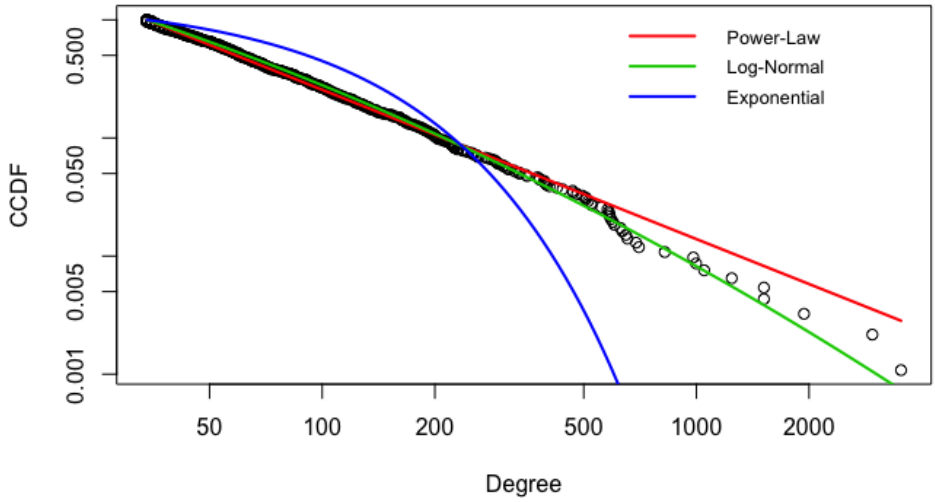


Figure 4.10: Continuous Distributions of Web-API Network

4.4 Estimation of Parameter k_{min} & γ

4.4.1 Exponent γ

In this paper, the maximum likelihood estimate (MLE) method is used to estimate the degree exponent γ. The MLE method is more scientific and reasonable by maximizing the likelihood function of the model to estimate the value of parameter. For continuous power-law distributions, the scaling parameter γ and its corresponding standard deviation are estimated according to the following formulas:

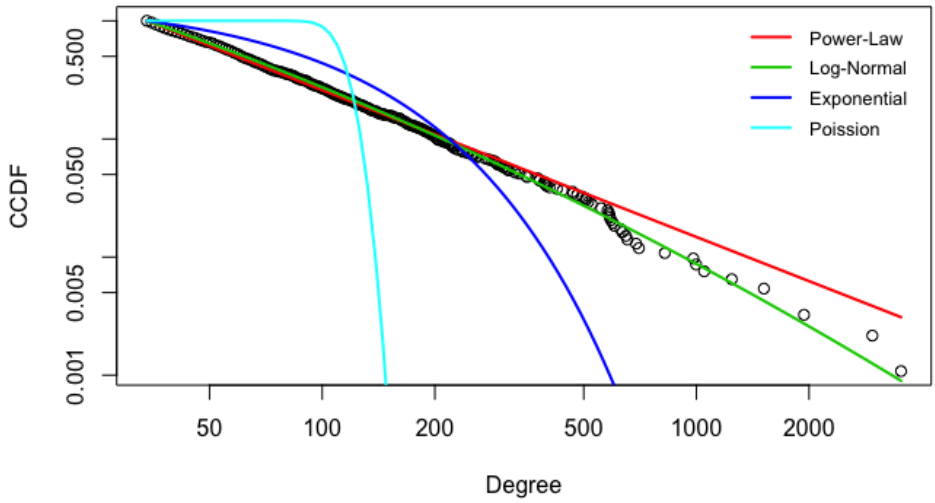


Figure 4.11: Discrete Distributions of Web-API Network

$$\hat{\gamma} = 1 + n\left[\sum_{i=1}^{n} \ln \frac{k_i}{k_{min}}\right]^{-1} \tag{4.5}$$

$$\sigma = \frac{\hat{\gamma} - 1}{\sqrt{n}} + O\left(\frac{1}{n}\right) \tag{4.6}$$

where k_i are the observed values of k that $k_i \geq k_{min}$, and k_{min} is the minimum value that corresponds to the power-law. The estimation value of parameter is that $\gamma > 1$ by default since $\gamma \leqslant 1$ does not exist in real world.

As for the discrete case, the estimate value of γ:

$$\hat{\gamma} \simeq 1 + n\left[\sum_{i=1}^{n} In \frac{k_i}{k_{\min} - \frac{1}{2}}\right]^{-1} \tag{4.7}$$

When $k_{min} = 1$ the proper estimation of γ is (Goldstein, Morris & Yen, 2004):

$$\frac{\zeta'(\hat{\gamma})}{\zeta(\hat{\gamma})} = -\frac{1}{n}\sum_{i=1}^{n} \ln k_i \tag{4.8}$$

where $\zeta(\hat{\gamma})$ is the Riemann Zeta function.

When $k_{min} > 1$, the appropriate estimator for γ is (Clauset, Young & Gleditsch, 2007):

$$\frac{\zeta'(\hat{\gamma}, k_{min})}{\zeta(\hat{\gamma}, k_{min})} = -\frac{1}{n}\sum_{i=1}^{n} \ln k_i$$

$$\sigma = \frac{\hat{\gamma} - 1}{\sqrt{n\left[\frac{\varsigma''(\hat{\gamma}, k_{\min})}{\varsigma(\hat{\gamma}, k_{\min})} - \left(\frac{\varsigma'(\hat{\gamma}, k_{\min})}{\varsigma(\hat{\gamma}, k_{\min})}\right)^2\right]}} \tag{4.9}$$

The estimation is a gradual normal. When the sample is $n \to \infty$, the variance tends to 0. In fact, due to the limited sample size (especially for the data at the tail), the distribution function of CCDF is more robust than the density function of PDF. Although it is generally believed that the best methods are CCDF and MLE, these two methods are not perfect. An important defect of the cumulative distribution is that the tail tends to deviate from the power law, so the commonly used method is to cut off the data from the tail before closing. Similarly, the head data is often removed, because the power-law distribution in reality is difficult to obey the power law well throughout the interval.

4.4.2 Lower Bound k_{min}

We use the KS statistic to estimate k_{min}. We choose $\hat{k}_{min}$ so that the distribution function we fit will be closest to the distribution function of the data. The KS statistic is defined as follows:

$$D = \max_{k \geq k_{min}} |S(k) - P(k)| \tag{4.10}$$

where $S(x)$ is the cumulative distribution function of the observed data (only data of $k > k_{min}$ is needed), and $P(k)$ is the cumulative distribution function of the power law distribution fitted. In order to quantitatively estimate the uncertainty of k_{min}, this paper uses the non-parametric Bootstrap method for estimation. The non-parametric Bootstrap

method was firstly proposed byEfron and Tibshirani (1994) and is an important practical method for data processing in modern statistics. It can use the Bootstrap sample to make statistical inferences on the overall N without any assumption about the distribution type [Table 4.1].

Table 4.1: An algorithm for uncertainty in k_{min}
Source, clauset, 2007

1: Set N equal to the number of values in the original data set.
2: **for** i in 1:B:
3: Sample N values (with replacement) from the original data set.
4: Estimate k_{min} and γ using the KS statistic.
5: **end for**

4.4.3 Experiment Result

[Figure 4.12][Figure 4.13] show the parameter estimation result of Web-API network constructed by BA model while [Figure 4.14][Figure 4.15] show the PS model for continuous power-law and discrete power-law. The top row shows the mean estimate of parameters k_{min}, γ and $ntail$. The bottom row shows the estimate of standard deviation for each parameter.

4.5 Preferential Attachment Measurement

Many real world systems are advantageously described as networks that are getting complex over time. Priority dependence and node fitness are two ubiquitous growth mechanisms that can not only explain some structural characteristics commonly observed in real-world systems, but also are associated with many applications in modeling

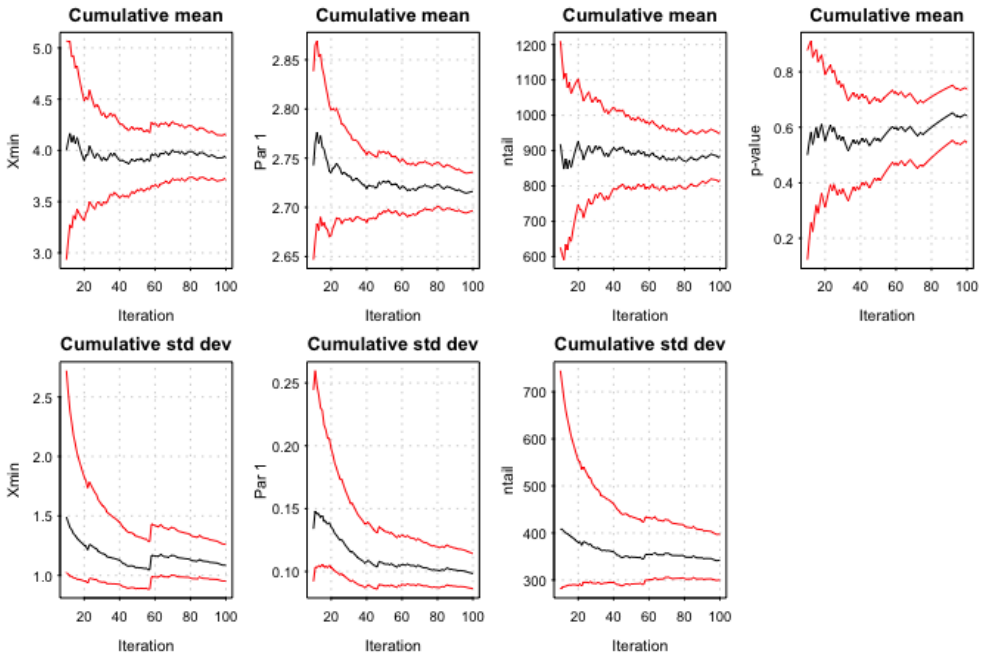


Figure 4.12: Simulation in Continuous Power-Law Pattern (BA model)

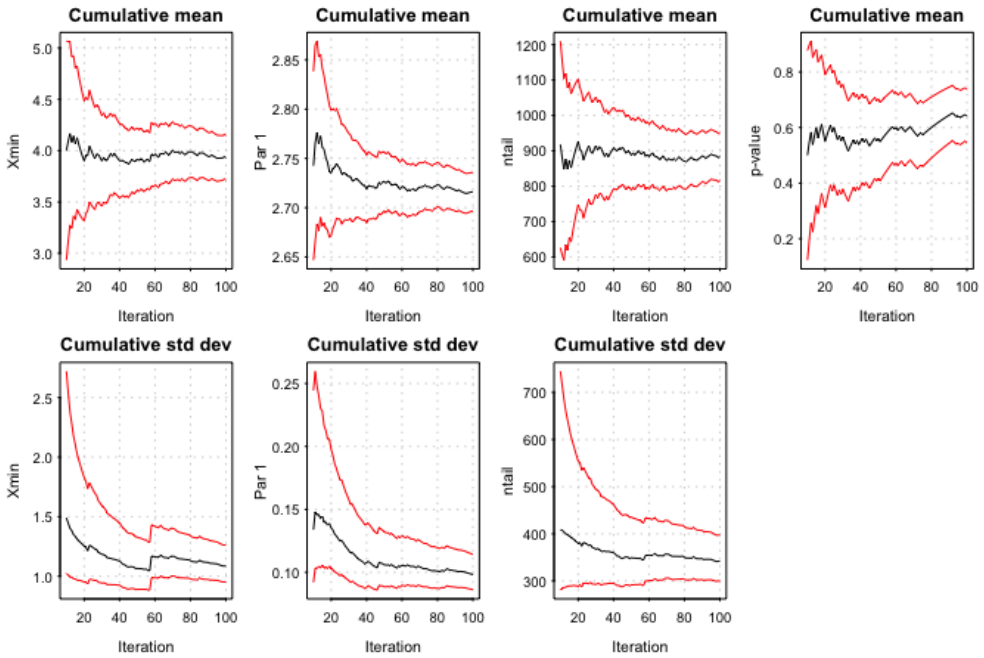


Figure 4.13: Simulation in Discrete Power-Law Pattern (BA model)

and reasoning. This paper uses the R-package PAFit (Pham, Sheridan & Shimodaira, 2015), which implements a well-established statistical method for the estimation of the fitness of preferred attachments and nodes, as well as some functions for generating

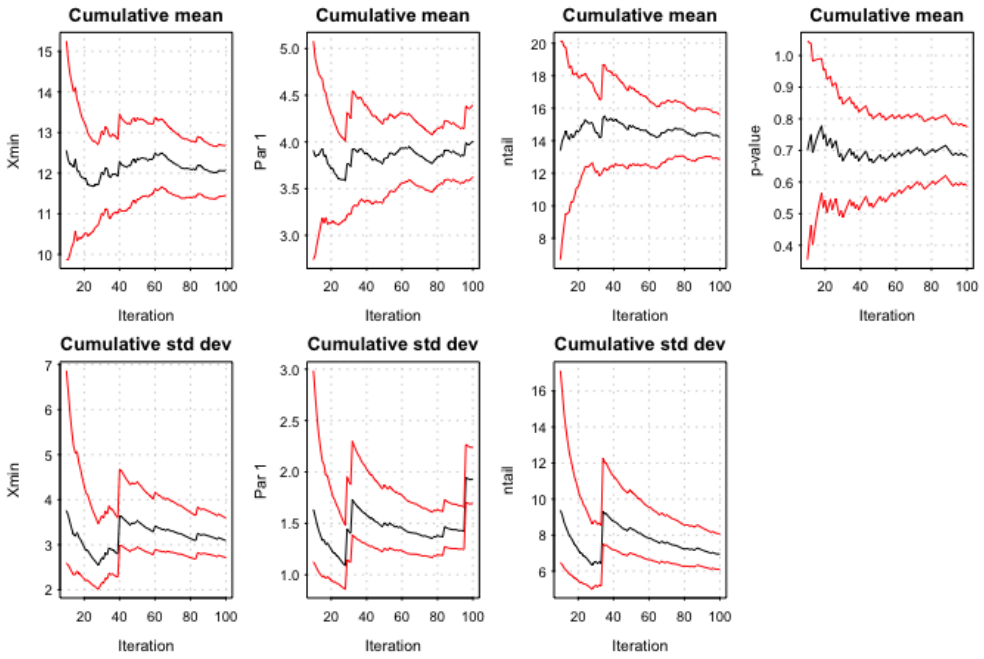


Figure 4.14: Simulation in Continuous Power-Law Pattern (PS model)

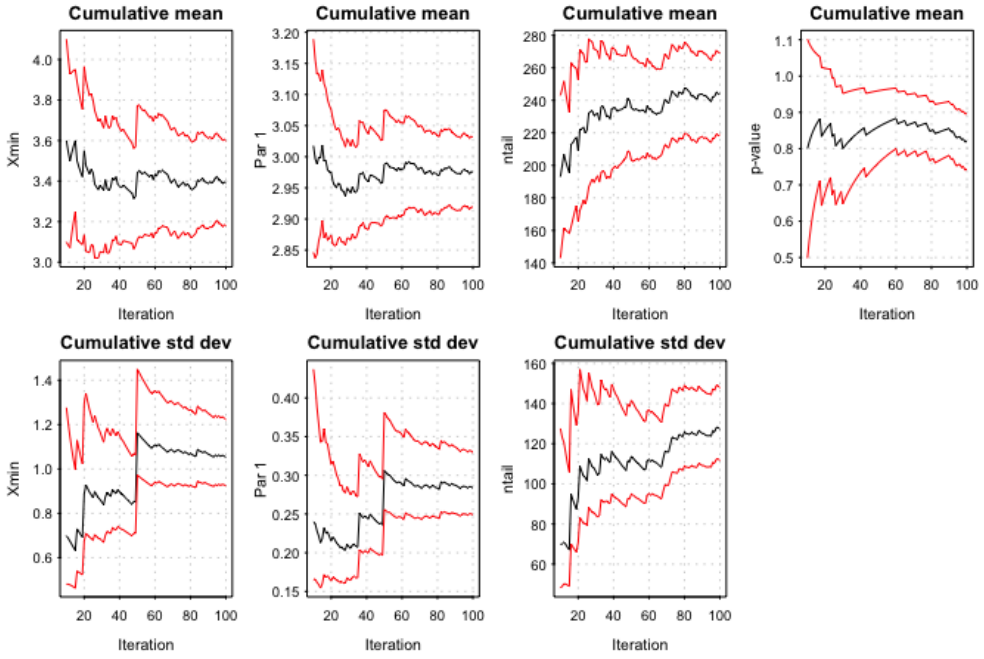


Figure 4.15: Simulation in Continuous Power-Law Pattern (PS model)

complex networks from these two mechanisms. This package ensures good performance for the calculation of large networks, which consist of many approaches to measure preferential attachment [Table. 4.2].

[Table. 4.3] shows the value of estimated attachment exponent results with various methods, including Jeong's method, Newman's method, and the PAFit method. These results were performed using various timestamps and the PA value calculated by PAFit method is approximated to 1, which demonstrated the existence and proved the Web API network follows the linear-PA.

Table 4.2: Existing attachment kernel estimation methods

Method	**Estimation Method**
Newman (2001)	Weighted sum of multiple histograms
Jeong et al. (2003)	Histogram
Massen et al. (2007)	Iterative fixed-point algorithm
Sheridan et al. (2012)	Markov chain Monte Carlo

Table 4.3: Preferential Attachment Measurement

Steps for Timestamp	**Jeong**	**Newman**	**PAFit**
10	0.66±0.21	0.97± 0.05	1.09± 0.06
20	0.66± 0.21	0.96± 0.05	1.08± 0.06
50	0.64± 0.20	0.95± 0.07	1.06± 0.07
100	0.65± 0.20	0.94± 0.06	1.05± 0.08
$Monthly$	0.67± 0.15	0.96± 0.09	1.03± 0.09

Chapter 5

Discussion

In this section, we firstly assess the results of empirical data. According to various parameters to further discuss about the Web-API network model fitting, we mainly discuss the significance of related work from the view of network science and web service discovery. According to our research, we evaluate how these works are essential in the real world and solve issues.

5.1 Assessment for Web-API Network Models

According to the results from above , we discuss if BA and PS network model can meet the characteristics of scale-free networks. [Table 5.1] [Table 5.2] list the network properties of BA network model and PS network model.

First of all, we perform a goodness-of-fit test for probability analysis (p-value). Fitting is used for the analysis of the distribution of existing observed variables and check that their distribution is in good agreement with a desired distribution (or standard distribution). Mathematically, the process of fitting is to find a process that can moderate the mathematical model of the current data sequence. In order to evaluate the degree of fitting, a mechanism for determining the validity of the fit is proposed, which is the goodness of fit. The goodness of fit can also assess the quality of the data fit based on the notion of test probability. The goodness-of-fit test is viewed as one of the key components of the statistical significance test via the chi-square statistic. It computes every category's expected frequency in the categorical variable based on the general distribution. It is compared with the observed frequency of distribution and evaluates if there is a dramatic difference between the observed frequency and the expected one and hence to realize the aim of analyzing the categorical variables. The agreement is tested via the statistical hypothesis test between the theoretical number of calculated and the number of observations based on the hypothesis or distribution model to decide if the model or the hypothesis is in line with the reality of observations.

For some distributions, there is a corresponding $p - value$ closed form expression, and thus an accurate $p - value$ can be obtained. However, for some other distributions that do not have a corresponding closed form expression, there is a p value range table. Normally, $0 < p < 1$ in the scale-free network. The larger the p is, the better the model fit the data. The smaller the p, the more significantly different the probability predicted from the model and observed in the empirical data. In the results of Web-API

network model fitting, it illustrates that both power-law and log-normal distribution can fit well. However, we found that the $p - value$ of continuous power-law distribution and discrete power-law distribution for BA model is 0.6685 and 0.7385, which are larger than $p - value$ of the log-normal distribution (0.6285). Similarly, the $p - value$ of continuous power-law distribution and discrete power-law distribution for PS model is 0.6835 and 0.8257, which are larger than $p - value$ of the log-normal distribution (0.3854). Obliviously, discrete power-law is the most fitting distribution among all.

Moreover, it has been found that $2 < \gamma < 3$ in a scale-free network. The value of γ is 2.7452 for discrete power-law of BA model while the value of γ is 2.7523. Both γ value of BA and PS model is in theoretical range.

With the statistical proof of $p-value$ and exponent γ, it verifies that the BA Web-API network and PS Web-API are scale-free networks.

Table 5.1: Web-API Network Distribution (BA model)

Parameter	**Continuous power-law**	**Discrete power-law**	**Log-normal**
k_{min}	20	5	33
γ	2.6921	2.7452	3.6921
$n - tail$	22	492	8
$p - value$	0.6685	0.7385	0.6285

Table 5.2: Web-API Network Distribution (PS model)

Parameter	**Continuous power-law**	**Discrete power-law**	**Log-normal**
k_{min}	15	4	11
γ	3.0852	2.7523	2.0982
$n - tail$	15	144	15
$p - value$	0.6835	0.8257	0.3854

5.2 Web Service Social Networks

The scale-free network means that the real network is constantly expanding and growing, such as the birth of new web pages in the Internet, the joining of new friends in the human network, the publication of new papers, the construction of new airports in aviation networks, and so on. New nodes will tend to be connected with more nodes when they join the network. For example, new web pages will generally have connections to well-known web sites. Meanwhile, newcomers will want to meet with celebrities in the community. The new paper tends to cite well-known literatures that have been widely quoted. The new airport will give priority to establish routes with the big airports. Complex networks have long been a specialty of network science researchers. They have penetrated into the work of statistical physics, social sciences, economic finance, and computing biologists. Practitioners in various industries use the network to abstract and model the interrelationships among their entities. Data generated from these disciplines encourages machine learning to develop predictive

models for graphs, such as GCN convolutional neural networks.

As the boundary of the information island system in the Internet gradually develops toward cross-integration, it will inevitably present multi-form infrastructure features in the future, enabling effective integration, association and interaction of multiple service interfaces. Therefore, Web services will become the major content of Internet services afterwards, which mainly refers to the integration of various resources such as data, functions and knowledge on the WWW. Therefore, the relationship between Web services and the WWW is very close. As the scale of the WWW gradually expands, the complexity will continue to increase. Meanwhile, as the current network technology research work, technicians will conduct a comprehensive analysis of their structure and dynamic characteristics, and focus on their major research work with typical network characteristics. Researchers use a variety of complex network analysis techniques to explore the complexity of the network and WWW. The emergence of a variety of networked software has created conditions and platforms for the study of network models. It is concluded that the current Web service network is showing the characteristics of a small world and is constantly developing towards a scale-free direction. That is to say, it is a bottom-up network model construction, which mainly refers to the relationship between network services based on the mining of WSDL services, and the interoperability characteristics of the network model.

The two Web-API social networks constructed in this paper are consistent with the characteristics of the scale-free network. However, compared to the BA model, the PS model introducing the concept of node similarity is closer to the state of the network service in reality. The in-depth study of the network ecological network and the optimization of the model can pay more attention to the topological structure of the individual interactions in the system and understand the nature and function of the complex system.

With the deepening of research on complex networks, it has been found that in

practical complex networks, there is a common nature: there is a large similarity between the nodes in a group while a small similarity with others in the network. We call this structure a community structure. The structure of the community was first proposed by Girvan and Newman, which was gradually accepted and applied. At present, the research on community structure has lasted for about ten years, and has become an important research hotspot and direction in the field of complex network research. Researchers can gain a deeper understanding of the dynamic behavior of the network through in-depth study of the community structure, which is an important way to understand the structure and function of the entire network.

5.3 Web Service Discovery & Recommendation

With the continuous development of modern computer network technology and service-oriented architecture, the research of Web service discovery and recommendation has also achieved a great breakthrough. At present, the major research still focuses on publishing, registration, organization and management of the Web service, and strives to meet the actual application needs of end users. A large number of Web services have begun to appear on the WWW. Confronting the current situation of information explosion caused by the rapid growth of Web services, service discovery and service recommendation have become a hot research issue in the field of service computing. With the rapid growth of the number of Web services, the service platform represented by the $ProgrammableWeb$ website has gradually become the major intermediary for Web service publishing and discovery. Because of the great amount of Web services on the service platform, users often find it difficult to choose due to the lack of experience or ability. Therefore, recommending appropriate Web services for users has become a top priority. Most current Web service recommendations focus on service function information or non-functional information such as QoS, time, etc., and do not pay enough

attention to the rich edge information of Web services, such as subject information and service composition information. At the same time, current service recommendation algorithms tend to merely focus on the improvement of accuracy while limited attention has been attached to the diversity of recommendation algorithms, which will result in more long-tail services for the service platform.

In the process of development, the web ecosystem is influenced by the Internet and computer models, which gradually changes towards the direction of big data, diversity and complexity, and finally builds a more complex system type based on Web services. However, it is difficult to form mutual perception between the various services involved. It does not have perfect semantic support, which causes the problem of information islands in current Web services to occur frequently, which affects the service level of the WWW. In order to solve this problem effectively, it is necessary for technicians to strengthen the analysis of Web services interconnection and actively carry out research on the Web service complex network from the perspective of community discovery.

On the other hand, social networks are social entity networks connected by relationships. The popularity of Internet has enabled human activities to be recorded, transmitted and stored in an unprecedented breadth and depth. Various forms of Web social networks provide users with an integrated platform for information dissemination and information sharing, personal opinion expression, thought exchange, emotional communication and economic exchanges. For the Web community network, from the release of content to the way how the network users apply the network, the center of the network offsets the majority of users, and the network has truly become a distributed information system with users as the main body. Therefore, in the future development direction, the Web-API social network should be modeled to reflect the nature of the behavior and preferences of individuals and groups of users. The classical social network model analysis method (Carrington, Scott & Wasserman, 2005) summarizes the property metrics of individual users as nodes by modeling social networks into graph

models, such as betweenness, closeness, centrality, degree and prestige. These metrics represent the association of a single node with other nodes in a social network. In addition, user preference modeling primarily reflects the extent to which how individual users or groups are interested in Web services. The keyword-based user preference model is represented by a set of keywords. Each keyword can represent the subject of interest. A set of keywords can be used to express the user interest. Amalthaea (Moukas, 1997) is a typical keyword-based user preference model. Keyword vectors with weights are widely used in Fab (Balabanović & Shoham, 1997) (web recommendation sites), Letizia (Lieberman et al., 1995) (browser helpers), Syskill, and Webert (Pazzani, Muramatsu, Billsus et al., 1996) (recommended systems).

5.4 Contributions

The main contribution of this paper is to conduct research from two perspectives of network science and Web services.

A great number of Web services widely exist on the Internet, which will inevitably form a complex network with complex structure, numerous nodes and mutual influence. At present, the research on the complex network characteristics of Web services is also in the development stage. Currently, the model and algorithm of building complex networks with Web services as nodes have been proposed. The characteristics of the constructed networks are studied.

The proposed method leads the research of Web services from the traditional SOA model to the complex network model. However, the traditionally constructed complex network is widely used the BA model, which only considers the probability and process of interconnection between nodes yet fails to consider the relationship between nodes. This is a certain difference for real Web service networks. In order to better explore the Web service network ecosystem, based on the BA model, the degree of similarity

described by the Web API is quantified according to the random walk and affects the probability of connecting other nodes in the network construction process, which is the PS model. This model is helpful to establish a better understanding and an in-depth study of Web-API network.

Moreover, optimization of Web service discovery and Web services recommendation still remains as a challenge nowadays. How to improve accuracy and relevance is still the main research direction now. In addition to continuous improvement and optimization of its algorithm, it is also helpful to reasonably construct and reflect the Web service network as realistically as possible.

Chapter 6

Conclusions

In this paper, we construct a Web-API evolving complex network based on typical BA model and PS model in the web service ecosystem. For better understanding of the structure of Web-API social network, a API-Mashup affiliation network is also visualized. The major difference between them is that the BA model focuses on the process of network growing, which depends on nodes popularity, while PS model considers similarity in a community aspect. We achieve this by first collecting data from $ProgrammableWeb$ and extracting the APIs' popularity distribution in the ecosystem. Then, for BA model, we calculate the probability of each new node connected to the initial network according to preferential attachment. On the other hand, for PS model, we use random walk algorithm to estimate the API's similarity according to their description. After network model is constructed, we fit empirical data and experiment result into goodness-of-fit test to verify if this network is scale-free or not. According to the result in quantitative research, the p-value and nodes distribution follow the theoretical expect. The research we did helps to understand the structure of Web service social network. Moreover, the PS model is closer to the real network, which can also help to improve the performance of web service discovery and web service recommendation.

6.1 Limitation and Future Work

The data collected in this paper is from $ProgrammableWeb$ which is a large directory for API only. However, other types of web service are not involved, such as SWS.

In recent years, the concept of network and the related studies have become an important means to reveal the structure and function of various complex systems in nature and human society. In the study of complex dynamic scale-free network degree distribution, it is found that the evolution of truest scale-free network degree distribution is not in equilibrium. However, in a non-equilibrium state, the unbalanced statistical mechanics method can be more accurately analyzed. The evolution law of the degree distribution of real networks is calculated. The concept of geometric preferential attachment explains scale-free degree distributions, strong clustering, and community structure at the same time, which provides a perspective to understand the community structure of social networks.

An important reason to study web service complex networks is to make a great contribution to the web service discovery. In this paper, only two models are constructed and analyzed from the perspective of network science. However, there is none significant contrast between them. In the future research, the performance of the web discovery of these two model services can be compared.

References

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45–65.

Albert, R. & Barabási, A.-L. (2000). Topology of evolving networks: local events and universality. *Physical review letters*, *85*(24), 5234.

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*(1), 47.

Albert, R., Jeong, H. & Barabási, A.-L. (1999). Internet: Diameter of the world-wide web. nature.

Albert, R., Jeong, H. & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *nature*, *406*(6794), 378.

Alberti, M., Cattafi, M., Chesani, F., Gavanelli, M., Lamma, E., Mello, P., ... Torroni, P. (2011). A computational logic application framework for service discovery and contracting. *International Journal of Web Services Research (IJWSR)*, *8*(3), 1–25.

Alonso, G., Casati, F., Kuno, H. & Machiraju, V. (2004). *Web services (pp. 123-149).* Springer Berlin Heidelberg.

Alrifai, M., Skoutas, D. & Risse, T. (2010). Selecting skyline services for qos-based web service composition. In *Proceedings of the 19th international conference on world wide web* (pp. 11–20).

Al Shalabi, L., Shaaban, Z. & Kasasbeh, B. (2006). Data mining: A preprocessing engine. *Journal of Computer Science*, *2*(9), 735–739.

Amatriain, X., Jaimes, A., Oliver, N. & Pujol, J. M. (2011). Data mining methods for recommender systems. In *Recommender systems handbook* (pp. 39–71). Springer.

Ascher, D., Dubois, P. F., Hinsen, K., Hugunin, J., Oliphant, T. et al. (2001). *Numerical python.* Citeseer.

Balabanović, M. & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, *40*(3), 66–72.

Barabási, A.-L. (2003). *Linked: The new science of networks.* AAPT.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *science*, *286*(5439), 509–512.

Barabási, A.-L., Albert, R. & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, *272*(1-2), 173–187.

Barabási, A.-L., Albert, R. & Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: statistical mechanics and its applications*, *281*(1-4), 69–77.

Barabási, A.-L. et al. (2016). *Network science*. Cambridge university press.

Barrat, A., Barabasi, A., Caldarelli, G., De los Rios, P., Erzan, A., Kahng, B., ... others (2004). Virtual round table on ten leading questions for network research. *European Physical Journal B*, *38*(ARTICLE), 143–145.

Barrat, A., Barthélemy, M. & Vespignani, A. (2004). Weighted evolving networks: coupling topology and weight dynamics. *Physical review letters*, *92*(22), 228701.

Barros, A. P. & Dumas, M. (2006). The rise of web service ecosystems. *IT professional*, *8*(5), 31–37.

Ben-Naim, E., Frauenfelder, H. & Toroczkai, Z. (2004). *Complex networks* (Vol. 650). Springer Science & Business Media.

Bizer, C., Heath, T. & Berners-Lee, T. (2008). Linked data: Principles and state of the art. In *World wide web conference* (Vol. 1, p. 40).

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, *424*(4-5), 175–308.

Bollobás, B. & Riordan, O. M. (2003). Mathematical results on scale-free random graphs. *Handbook of graphs and networks: from the genome to the internet*, 1–34.

Borg, I. & Groenen, P. (2003). Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement*, *40*(3), 277–280.

Borgatti, S. P. (1995). Centrality and aids. *Connections*, *18*(1), 112–114.

Brown, A. R. (1960). *Method in social anthropology*. Times Of India; Bombay.

Bu, T. & Towsley, D. (2002). On distinguishing between internet power law topology generators. In *Proceedings. twenty-first annual joint conference of the ieee computer and communications societies* (Vol. 2, pp. 638–647).

Burda, Z., Jurkiewicz, J. & Krzywicki, A. (2004). Network transitivity and matrix models. *Physical Review E*, *69*(2), 026106.

Carrington, P. J., Scott, J. & Wasserman, S. (2005). *Models and methods in social network analysis* (Vol. 28). Cambridge university press.

Chau, D. H., Pandit, S., Wang, S. & Faloutsos, C. (2007). Parallel crawling for online social networks. In *Proceedings of the 16th international conference on world wide web* (pp. 1283–1284).

Chen, Q., Chang, H., Govindan, R. & Jamin, S. (2002). The origin of power laws in internet topologies revisited. In *Proceedings. twenty-first annual joint conference of the ieee computer and communications societies* (Vol. 2, pp. 608–617).

Chen, W., Paik, I. & Hung, P. C. (2013). Constructing a global social service network for better quality of web service discovery. *IEEE transactions on services computing*, *8*(2), 284–298.

Cherifi, C. & Santucci, J.-F. (2012). A comparative study of web services composition networks. In *2012 eighth international conference on signal image technology and internet based systems* (pp. 700–706).

Chesbrough, H. & Spohrer, J. (2006). A research manifesto for services science. *Communications of the ACM*, *49*(7), 35–40.

Cho, J. (2001). Crawling the web: discovery and maintenance of large-scale web data. *A Thesis Nov*.

Clauset, A., Shalizi, C. R. & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, *51*(4), 661–703.

Cowles, M. K. & Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, *91*(434), 883–904.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. & Suri, S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 160–168).

Doar, M. B. (1996). A better model for generating test networks. In *Proceedings of globecom'96. 1996 ieee global telecommunications conference* (pp. 86–93).

Dormann, C. F., Gruber, B. & Fründ, J. (2008). Introducing the bipartite package: analysing ecological networks. *interaction*, *1*(0.2413793).

Dorogovtsev, S. N. & Mendes, J. F. (2013). *Evolution of networks: From biological nets to the internet and www*. OUP Oxford.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Elmeleegy, H., Ivan, A., Akkiraju, R. & Goodwin, R. (2008). Mashup advisor: A recommendation tool for mashup development. In *2008 ieee international conference on web services* (pp. 337–344).

Erds, P. & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, *5*, 17–61.

Eubank, S., Guclu, H., Kumar, V. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z. & Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, *429*(6988), 180.

Everett, M. G. & Borgatti, S. P. (1999). The centrality of groups and classes. *The Journal of mathematical sociology*, *23*(3), 181–201.

Fallatah, H., Bentahar, J. & Asl, E. K. (2014). Social network-based framework for web services discovery. In *2014 international conference on future internet of things and cloud* (pp. 159–166).

Faloutsos, M., Faloutsos, P. & Faloutsos, C. (1999). On power-law relationships of the internet topology. In *Acm sigcomm computer communication review* (Vol. 29, pp. 251–262).

Farzi, P., Akbari, R. & Bushehrian, O. (2017). Improving semantic web service discovery method based on qos ontology. In *2017 2nd conference on swarm intelligence and evolutionary computation (csiec)* (pp. 72–76).

Feng, Z., Lan, B., Zhang, Z. & Chen, S. (2015). A study of semantic web services network. *The Computer Journal*, *58*(6), 1293–1305.

Fern, X. Z. & Brodley, C. E. (2004). Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the twenty-first international conference on machine learning* (p. 36).

Fortunato, S., Latora, V. & Marchiori, M. (2004). Method to find community structures based on information centrality. *Physical review E*, *70*(5), 056104.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215–239.

García, J. M., Ruiz, D. & Ruiz-Cortés, A. (2012). Improving semantic web services discovery using sparql-based repository filtering. *Web Semantics: Science, Services and Agents on the World Wide Web*, *17*, 12–24.

Ghoshal, G. (2009). Structural and dynamical properties of complex networks.

Gilbert, A. (1959). The composition of the blood of the shore crab, carcinus moenas pennant, in relation to sex and body size: Ii. blood chloride and sulphate. *Journal of Experimental Biology*, *36*(2), 356–362.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. (1995). *Markov chain monte carlo in practice*. Chapman and Hall/CRC.

Girvan, M. & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821–7826.

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

Gronmo, R., Skogan, D., Solheim, I. & Oldevik, J. (2004). Model-driven web service development. *International Journal of web Services Research (IJWSR)*, *1*(4), 1–13.

Hagberg, A., Swart, P. & S Chult, D. (2008). *Exploring network structure, dynamics, and function using networkx* (Tech. Rep.). Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Hofman, J. M., Sharma, A. & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488.

Holme, P., Kim, B. J., Yoon, C. N. & Han, S. K. (2002). Attack vulnerability of complex networks. *Physical review E*, *65*(5), 056109.

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (nzcsrsc2008), christchurch, new zealand* (Vol. 4, pp. 9–56).

Huang, K., Fan, Y. & Tan, W. (2012). An empirical study of programmable web: A network analysis on a service-mashup system. In *2012 ieee 19th international conference on web services* (pp. 552–559).

Ifrim, G., Bakir, G. & Weikum, G. (2008). *of proceedings: Kdd 2008: proceedings of the 14th acm kdd international conference on knowledge discovery & data mining*. ACM.

Jeong, H., Néda, Z. & Barabási, A.-L. (2003). Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, *61*(4), 567.

Jhingran, A. (2006). Enterprise information mashups: integrating information, simply. In *Proceedings of the 32nd international conference on very large data bases* (pp. 3–4).

Jiang, H., Yang, X., Yin, K., Zhang, S. & Cristoforo, J. A. (2011). Multi-path qos-aware web service composition using variable length chromosome genetic algorithm.

Information Technology Journal, *10*(1), 113–119.

Jivani, A. G. et al. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, *2*(6), 1930–1938.

Jolliffe, I. (2011). *Principal component analysis*. Springer.

Kay, M. (2001). *Xslt: programmer's reference*. Wrox Press Ltd.

Klemm, K. & Eguiluz, V. M. (2002). Growing scale-free networks with small-world behavior. *Physical Review E*, *65*(5), 057102.

Li, X. & Chen, G. (2003). A local-world evolving network model. *Physica A: Statistical Mechanics and its Applications*, *328*(1-2), 274–286.

Lieberman, H. et al. (1995). Letizia: An agent that assists web browsing. *IJCAI (1)*, *1995*, 924–929.

Loper, E. & Bird, S. (2002). Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Lü, L., Jin, C.-H. & Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, *80*(4), 046122.

Lyu, S., Liu, J., Tang, M., Kang, G., Cao, B. & Duan, Y. (2014). Three-level views of the web service network: an empirical study based on programmableweb. In *2014 ieee international congress on big data* (pp. 374–381).

Maamar, Z., Faci, N., Wives, L., Badr, Y., Santos, P. & de Oliveira, J. P. M. (2011). Using social networks for web services discovery. *IEEE internet computing*, *15*(4), 48–54.

Maamar, Z., Faci, N., Wives, L. K., Yahyaoui, H. & Hacid, H. (2011). Towards a method for engineering social web services. In *Working conference on method engineering* (pp. 153–167).

Maamar, Z., Hacid, H. & Huhns, M. N. (2011). Why web services need social networks. *IEEE Internet Computing*, *15*(2), 90–94.

Maamar, Z., Wives, L. K., Badr, Y., Elnaffar, S., Boukadi, K. & Faci, N. (2011). Linkedws: A novel web services discovery model based on the metaphor of "social networks". *Simulation Modelling Practice and Theory*, *19*(1), 121–132.

Maleshkova, M., Pedrinaci, C. & Domingue, J. (2010). Investigating web apis on the world wide web. In *2010 eighth ieee european conference on web services* (pp. 107–114).

McCallum, A., Nigam, K. & Ungar, L. H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth acm sigkdd international conference on knowledge discovery and data mining* (pp. 169–178).

McPherson, M., Smith-Lovin, L. & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, *27*(1), 415–444.

Menczer, F. (2002). Growing and navigating the small world web by local content. *Proceedings of the National Academy of Sciences*, *99*(22), 14014–14019.

Menczer, F. (2004). Evolution of document networks. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5261–5265.

Merton, R. K. (1968). The matthew effect in science: The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63.

Moukas, A. (1997). Amalthaea information discovery and filtering using a multiagent evolving ecosystem. *Applied Artificial Intelligence*, *11*(5), 437–457.

Newman, M. E. (2000). Models of the small world. *Journal of Statistical Physics*, *101*(3-4), 819–841.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical review E*, *64*(2), 025102.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167–256.

Newman, M. E. (2005). A measure of betweenness centrality based on random walks. *Social networks*, *27*(1), 39–54.

Newman, M. E. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, *69*(2), 026113.

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, *9*(3), 10–20.

Paliwal, A. V., Shafiq, B., Vaidya, J., Xiong, H. & Adam, N. (2011). Semantics-based automated service discovery. *IEEE Transactions on Services Computing*, *5*(2), 260–275.

Pan, Y., Li, D.-H., Liu, J.-G. & Liang, J.-Z. (2010). Detecting community structure in complex networks via node similarity. *Physica A: Statistical Mechanics and its Applications*, *389*(14), 2849–2857.

Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguná, M. & Krioukov, D. (2012). Popularity versus similarity in growing networks. *Nature*, *489*(7417), 537.

Park, J. & Newman, M. E. (2004). Statistical mechanics of networks. *Physical Review E*, *70*(6), 066117.

Pazzani, M. J., Muramatsu, J., Billsus, D. et al. (1996). Syskill & webert: Identifying interesting web sites. In *Aaai/iaai, vol. 1* (pp. 54–61).

Pham, T., Sheridan, P. & Shimodaira, H. (2015). Pafit: A statistical method for measuring preferential attachment in temporal complex networks. *PloS one*, *10*(9), e0137796.

Pitaevskii, L., Lifshitz, E. & Sykes, J. (2017). *Course of theoretical physics: Physical kinetics* (Vol. 10). Elsevier.

Pons, P. & Latapy, M. (2005). Computing communities in large networks using random walks. In *International symposium on computer and information sciences* (pp. 284–293).

Powers, D. M. (1998). Applications and explanations of zipf's law. In *Proceedings of the joint conferences on new methods in language processing and computational natural language learning* (pp. 151–160).

Price, D. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, *27*(5), 292–306.

Qian, G., Sural, S., Gu, Y. & Pramanik, S. (2004). Similarity between euclidean and cosine angle distance for nearest neighbor queries. In *Proceedings of the 2004 acm symposium on applied computing* (pp. 1232–1237).

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, *101*(9), 2658–2663.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, pp. 133–142).

Rangarajan, S. & Chandar, R. K. (2017). Qos-based architecture for discovery and selection of suitable web services using non-functional properties. *EAI Endorsed Transactions on Scalable Information Systems*, *4*(12).

Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, *290*(5500), 2323–2326.

Samir, S., Sarhan, A. & Algergawy, A. (2017). Context-based web service discovery framework with qos considerations. In *2017 11th international conference on research challenges in information science (rcis)* (pp. 146–155).

Saul, L. K. & Roweis, S. T. (2003). Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of machine learning research*, *4*(Jun), 119–155.

Scott, J. (1988). Social network analysis. *Sociology*, *22*(1), 109–127.

Şimşek, Ö. & Jensen, D. (2008). Navigating networks by using homophily and degree. *Proceedings of the National Academy of Sciences*, *105*(35), 12758–12762.

Stephenson, K. & Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social networks*, *11*(1), 1–37.

Strogatz, S. H. (2001). Exploring complex networks. *nature*, *410*(6825), 268.

Tan, W., Zhang, J. & Foster, I. (2010). Network analysis of scientific workflows: A gateway to reuse. *Computer*, *43*(9), 54–61.

Tan, W., Zhang, J., Madduri, R., Foster, I., De Roure, D. & Goble, C. (2011). Servicemap: Providing map and gps assistance to service composition in bioinformatics. In *2011 ieee international conference on services computing* (pp. 632–639).

Tata, S. & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. *ACM Sigmod Record*, *36*(2), 7–12.

Tenenbaum, J. B., De Silva, V. & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, *290*(5500), 2319–2323.

Tong, H., Faloutsos, C. & Pan, J.-Y. (2008). Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, *14*(3), 327–346.

Tsalgatidou, A. & Pilioura, T. (2002). An overview of standards and related technology in web services. *Distributed and parallel databases*, *12*(2-3), 135–162.

Turney, P. D. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, *37*, 141–188.

Uzzi, B., Amaral, L. A. & Reed-Tsochas, F. (2007). Small-world networks and management science research: A review. *European Management Review*, *4*(2), 77–91.

Van Der Walt, S., Colbert, S. C. & Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, *13*(2),

22.

Vázquez, A. & Weigt, M. (2003). Computational complexity arising from degree correlations in networks. *Physical Review E*, *67*(2), 027101.

Vijayarani, S., Ilamathi, M. J. & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7–16.

Wang, H., Feng, Z., Chen, S., Xu, J. & Sui, Y. (2010). Constructing service network via classification and annotation. In *2010 fifth ieee international symposium on service oriented system engineering* (pp. 69–73).

Watts, D. J. (2004). The "new" science of networks. *Annu. Rev. Sociol.*, *30*, 243–270.

Waxman, B. M. (1988). Routing of multipoint connections. *IEEE journal on selected areas in communications*, *6*(9), 1617–1622.

Weiss, G. H. & Rubin, R. J. (1982). Random walks: theory and selected applications. *Advances in Chemical Physics*, 363–505.

Weiss, M. & Gangadharan, G. (2010). Modeling the mashup ecosystem: Structure and growth. *R&d Management*, *40*(1), 40–49.

West, D. B. et al. (1996). *Introduction to graph theory* (Vol. 2). Prentice hall Upper Saddle River, NJ.

Wood, R. K. (1993). Deterministic network interdiction. *Mathematical and Computer Modelling*, *17*(2), 1–18.

Xing, W. & Ghorbani, A. (2004). Weighted pagerank algorithm. In *Proceedings. second annual conference on communication networks and services research, 2004.* (pp. 305–314).

Yang, B. & Liu, J. (2008). Discovering global network communities based on local centralities. *ACM Transactions on the Web (TWEB)*, *2*(1), 9.

Yurke, B. & Denker, J. S. (1984). Quantum network theory. *Physical Review A*, *29*(3), 1419.

Zha, H., He, X., Ding, C., Simon, H. & Gu, M. (2001). Bipartite graph partitioning and data clustering. In *Proceedings of the tenth international conference on information and knowledge management* (pp. 25–32).

Zhou, H. (2003). Distance, dissimilarity index, and network community structure. *Physical review e*, *67*(6), 061901.

Zhou, J., Zhang, T., Meng, H., Xiao, L., Chen, G. & Li, D. (2008). Web service discovery based on keyword clustering and ontology. In *2008 ieee international conference on granular computing* (pp. 844–848).

Appendix A

Code Implementation

All codes and output can be found in the following link:

https://drive.google.com/drive/folders/1UMtO_MZy-sQgmcz2A8jU1xCk3GTT6uYy