Statistical approaches on the analyses and interpretations of a scientometric database

A dissertation
submitted to the Faculty of
the School of Multidisciplinary Sciences
the Department of Statistical Science
The Graduate University for Advanced Studies, SOKENDAI
by

Chang Lin-Hsuan

In partial fulfillment of the requirements
for the Degree of
Doctor of Philosophy

Satoshi Kuriki, Advisor

March 2020

Acknowledgements

I am grateful to all people who have supported me in the process of my Ph.D. course. My supervisor Professor Satoshi Kuriki taught me a lot of things and helped me a lot during my Ph.D course. Also, I would like to express my sincere gratitude to my former supervisor, Professor Junji Nakano. Professor Nakano spent a lot of time for discussing the content of papers with me. I would like to thank Dr. Frederick Kin Hing Phoa of the Institute of Statistical Science, Academia Sinica, who is my former supervisor while I was working in Academia Sinica. Finally, a big thanks to Clarivate Analytics for providing the Web of Science database and the URA team of ISM, especially Dr. Keisuke Honda and Mrs. Hiroka Hamada, for making the database easy to access and maintaining the database. I am also grateful to all other students and professors in ISM.

Abstract

Nowadays, the research funds are very limited, and all research institutes and universities are competing against one another for these resources. Government and private funding agencies, on the other hand, make their decisions on distributing these limited resources mainly based on the applicants' research performances. Therefore, it is very important to have a fair method to judge the research performance, so that the resources are distributed in a reasonable and efficient way. The main theme of this thesis is to study how to evaluate the research performance of an entity in a fair way using a large-scale citation network called the Web of Science. In specific, the phenomenon of cross-disciplinary citation is studied first, then the article network influence (ANI) method is introduced to evaluate the research performance of an entity, including but not limited to research articles, individual researchers, scientific journals, international conferences, universities and research institutes, or even countries. Lastly, the construction of the generative model for a citation network is investigated. Instead of studying the whole Web of Science, we particularly study a subset under the topic of statistics for demonstration purposes.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Nowadays, the research funds are very limited, and all research institutes and universities are competitive against one another for these resources. Government and private funding agencies, on the other hand, make their decisions on distributing these limited resources mainly based on the applicants' research performance. Therefore, it is very important to have a fair method to judge the research performance, so that the resources are distributed in a reasonable and efficient way. As a result, bibliometrics (Pritchard, 1969), which is an analysis of research performance, becomes a hot topic that governments and research institutes advocate and emphasize recently.

Among all bibliometrics, citation plays an important role in evaluating research performance. In the regime of big data, citation information is gathered from large-scale databases and represented in network forms. A citation network analysis is a quantitative method to identify important and impacted literature of a field on the basis of how often a publication is cited in other publications. This analysis has recently become an essential tool to evaluate scientific achievements at different entities, including but not limited to, research articles, individual researchers, scientific journals, international conferences, universities and research institutes, or even countries. Governments and funding agencies make appropriate decisions on the allocation of their resources to these entities according to their research performances. For example, an education department can allocate its educational funds to schools based on quantitative reports of their performances.

In this thesis, we aim at studying the large-scale citation network obtained from one of the biggest scientometric database Web of Science and reveal some interesting phenomena that are seldom to observe quantitatively. In specific, we demonstrate our analytical methods via a subset of the whole Web of Science database under the topic of statistics. This subset represents the citation information of the statistics community over the past 36 years. The rest of this thesis is divided into the following parts. In the rest of Chapter 1, we will introduce some existing methods of network centrality and research metrics. Then we will introduce the Web of Science database

and observe the phenomenon of how other fields use statistics paper as the reference in Chapter 2. Chapter 3 is the main part of this thesis, which introduces a newly developed method called the article network influence (ANI) method for evaluating the research performance. We will further investigate the construction of the generative model for statistics citation network in Chapter 4. The last chapter will consist of discussion and summary of the whole thesis.

1.1 Network Centrality

Evaluating the research performance of the article or identifying the influential article in the citation network is similar to finding the mean value among a group of numbers, and the center of a network, which usually plays an important role in the network, is always of interest from business, government, and academia. Traditionally, the centrality measures provide information on the importance of a target node in a network. There have been many methods in the literature to study the centrality. Traditionally, the center of a network can be detected via three common measures: degree centrality, betweenness centrality, and closeness centrality. There are some notations that need to be introduced first. Given a network $G = (V, E)$ with n nodes and m edges, where $V = \{v_1, v_2, ..., v_n\}$ represents all nodes in the network G and E is a set of undirected edges in G.

1.1.1 Degree Centrality

The most common method for defining the centrality of the network is the degree centrality. In a network, a degree is the number of connections a node has. It is intuitive to understand that a node with the highest degree is the center of the network.

$$C_D(v_t) = degree(v_t), \tag{1.1}$$

where $degree(v_t)$ is the number of degree that node v_t has.

1.1.2 Closeness Centrality

Bavelas (1950) developed the closeness centrality to calculate the mean distance of each node to all others. Below is the equation for calculating the closeness centrality.

$$C_C(v_t) = \frac{1}{\sum_j d(v_t, v_j)}, \tag{1.2}$$

where $d(v_i, v_j)$ is the distance from node v_i to other nodes.

1.1.3 Betweenness Centrality

The idea of Betweenness is to see how the information flow in the network (Freeman, 1977), which is different from the degree centrality. This method is to identify the node that can connect different groups.

$$C_B(v_t) = \frac{g_{v_t}(v_i, v_j)}{g(v_i, v_j)}, \tag{1.3}$$

where $g(v_i, v_j)$ is the number of paths between v_i and v_j. $g_{v_t}(v_i, v_j)$ is the number of paths between v_i and v_j that passed v_t. However, it takes a lot of computation time to obtain this quantity. (Newman, 2005) modified this method and developed a betweenness centrality with a random walk.

1.1.4 Eigenvector Centrality

Eigenvector centrality is a measure of the influence of a node in a network. For each node, a relative score is assigned based on an idea that higher contribution is rewarded for nodes connecting to higher-scoring nodes. Thus, a node with high eigenvector centrality implies that it is connected to many high-score (influential) nodes in a network. Mathematically, let A be the adjacency matrix of a network, i.e. $a_{v,t} = 1$ if node v is linked to another node t, and 0 otherwise. Then the eigenvector centrality x can be defined as the solution of the following eigenvector equation:

$$Ax = \lambda x. \tag{1.4}$$

1.1.5 Pagerank Centrality

Page et al. (1998) developed another version of centrality that shares a similar concept with the eigenvector centrality. If both nodes have the same number of neighbors, the node which has important neighbors will have more value than the other node. Pagerank can be calculated from the following equation,

$$C_p(v_t) = c \sum_{u \in M(v_t)} \frac{C_p(v_i)}{degree(v_i)} + \frac{1-c}{n}, \tag{1.5}$$

where c is a damping factor, $M(v_t)$ is v_t's neighbor nodes, $C_p(v_i)$ is the pagerank of node v_i, $degree(v_i)$ is the number of out-degree from v_i, and n is the number of nodes. The pagerank score of neighbors will be used to calculate the pagerank score of the target node.

1.1.6 Focus Centrality

Most of the centrality methods do not have a procedure of statistical test for verification. Wang and Phoa (2016) developed a new centrality method that provides a statistical testing procedure. The definition of the focus centrality is

$$C_F(v_t) = \sum_{d=1}^{m} g(d)(o(d) - e(d)), \tag{1.6}$$

where $o(d)$ and $e(d)$ are the observed numbers of nodes and expected numbers of nodes with distance d from node v_t, and m is the maximum distance of interest. $g(d)$ is the weighting function which is decay with the distance increase. The main concept of this method is that the center node should have a higher connection probability than the rest of other nodes. The hypothesis is $H_0 : p_c = p_0$ and $H_a : p_c > 0$, where p_c is the connection probability of the center node, and p_0 is the connection probability of the other nodes.

Assume the network follows the assumption of a Poisson random graph, $e(d)$ can be derived from this assumption (Bogu and Pastor-Satorras (2003), Blondel et al. (2008), and Fronczak et al. (2004)). To statistically verify the network centrality, another likelihood ratio test is proposed. Following the derivation in Wang and Phoa (2016), they considered the distance d between v_t and all other nodes follows a multinomial distribution of size 1, with citation probabilities $p_0(1), ..., p_0(d_M)$, where $\sum_{d=1}^{d_M} p_0(d) = 1$. Based on the properties of the multinomial distribution, the variance-covariance matrix can be obtained. Then the asymptotic variance is

$$Var(e_t(d) - o_t(d)) = g(d)^T \sum g(d), \tag{1.7}$$

and the test statistics under the null hypothesis $H_0 : e_t(d) - o_t(d) = 0$ is

$$T_d = \frac{e_t(d) - o_t(d)}{\sqrt{Var(e_t(d) - o_t(d))}}. \tag{1.8}$$

A failure of accepting the null hypothesis means that there is a significant difference at distance d between the number of nodes from v_t and the average number of nodes from any individual nodes in the network.

1.2 Research Metrics

In the citation network, when a researcher published a scientific article, others read it and followed his/her step. When these followers completed their own research that

was influenced or inspired by the original article, they cited it in the reference when they wrote their scientific articles. This led to a citation, which is officially defined as an abbreviated alphanumeric expression embedded in the body of an intellectual work that denotes an entry in the bibliographic reference section of the work for the purpose of acknowledging the relevance of the work of others to the topic of discussion at the spot where the citation appears. After years of progress, a citation network has resulted. We introduce some metric for the importance of an article in this section and a new influence-based metric for the importance of an article in the later chapter, and the importance is in terms of the influence of an article towards all other articles in its associated field or the whole citation network.

1.2.1 Citation count

This method is the most simple method to investigate the performance of the article/ journal. The number of citations a node received is simply represented how important the article or journal is. The idea of this method is similar to degree centrality that only the direct connections are being considered. However, this method lacks time-balance. Newman (2008) modified the citation count method and developed a rescaled-citation count method.

1.2.2 Impact factor

Impact factor is the most famous method to evaluate the research performance of journals. It is very easy to understand the idea and do the calculation. Impact factor only uses the data in 2 years, and it can be calculated from below,

$$IF_y = \frac{Citations_{y-1} + Citations_{y-2}}{Publications_{y-1} + Publications_{y-2}}. \tag{1.9}$$

The impact factor of this year is based on the citation number and article number of the past 2 years. Although it is the most common method in the research metrics, it seems like this method is less fair to some certain fields. For example, medical and biology fields tend to receive many citations while the papers come out. Whereas, mathematics papers or any theoretical papers need more time to receive citations.

1.2.3 Pagerank

Pagerank can be applied to the citation network and be used to find the influential paper. However, the method of Pagerank and most of other methods lack time-balance. For example, the older paper will receive more citations than the recent paper, which will cause the older paper to have a higher impact score than the recent

paper. Mariani et al. (2016) modified the traditional Pagerank method and developed the Rescaled Pagerank:

$$R_i(p) = \frac{p_i - \mu_i(p)}{\sigma_i(p)}, \tag{1.10}$$

where $\mu_i(p)$ and $\sigma_i(p)$ is the mean and standard deviation of the Pagerank in a certain time frame. The timeframe they used is not in days or years but in the number of published papers. They use this method to identify the milestone papers and can be able to find out the recent milestone papers.

1.2.4 Field-weighted citation impact(FWCI)

Field-Weighted Citation Impact (FWCI) is a new measure introduced by Elsevier (2018) to evaluate the entity's impacts in Scopus, the database of Elsevier since 1996. It indicates how the number of citations received by an entity's publications compared to the average number of citations received by all other similar publications. Mathematically, the FWCI of an entity is defined as:

$$FWCI = \frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{e_i},$$

where N is the number of publications by an entity, c_i is the citations received by publication i, and e_i is the expected number of citations received by all similar publications in the publication year plus the following three years. When a similar publication is allocated to more than one discipline, the harmonic mean is used to calculate e_i.

FWCI can be viewed as a simple modification from the impact factor. It differs in two places: (1) the inclusion of the number of citations changes from two backward years backward to three forward years; (2) instead of the number of publications as the denominator, FWCI considers the average number of citations by similar entities. Thus, two differences between ANI and impact factor also inherit in the difference to FWCI, and they are indirect citations and subsets in terms of years.

Chapter 2

Citations of Academic Articles and Statistics Articles in Fields of Sciences

Statistics, a useful tool which using data to do modeling, analysis, interpretation, and inference, can be applied in different areas. For example, engineering field needs the skill of experimental design, reliability, and quality control, like the knowledge of six-sigma is essential for the quality check of factory production. In the biology or medical area, it is important to know whether the new drug is affected or not. The clinical trial or survival analysis will be used to analyze the effect of the drug. In the marketing area, A/B test is conducted a lot to check if the UI/UX can attract more people to go to the website or if the promotion is affected or not. Time series is also a useful method in the financial area or to check the purchase trend of e-commerce. It would be interesting to know how other fields use statistics as the reference. In this chapter, we analyze the citations within each academic field and citations from other fields toward statistics articles. We use the database of academic articles "Web of Science" to obtain the information we need.

2.1 Web of Science and the academic field

Web of Science (WoS), owned by Clarivate Analytics, is one of the most complete bibliographic databases in the world. It contains relevant attributes of published scientific articles, such as the journals where the articles published, their publication years, their authors, their reference lists, and many others. The Institute of Statistical Mathematics has a contract with Clarivate Analytics, so we are able to use the data from WoS. The data we are able to access is from 1981 to 2016. Technically, the WoS database we can access is stored in neo4j graph database. Details about neo4j can be

Table 2.1: Number of journals assigned with different number of subjects

# of Assigned subjects	1	2	3	4	5
# of Journals	6792	6794	3062	1474	600
# of Articles	16270044	16674268	6618622	3818521	1540125
# of Assigned subjects	6	7	8	9	10
# of Journals	188	55	23	7	4
# of Articles	492942	183419	48700	25634	38414

referred to Baton and Bruggen (2017). We use the Cypher query language to extract the required data from the database.

In the WoS database, each journal is assigned with one or multiple different subject areas, such as statistics, social science, or chemistry. However, there are 140 journals which do not have any subjects assigned with (such as International Review of Connective Tissue Research). After removing all the duplicated subjects (such as Legal Medicine and Medicine, Legal) and duplicated journals (2D Materials and 2D MATERIALS), there are 266 subjects, 19138 journals, and 45769924 articles left in the data.

The number of subjects a journal assigned with is shown in Table 2.1. We are interested in the subject area of "Statistics & Probability". There is no journal assigned with only statistics but also with other subjects. There are totally 159 journals assigned to statistics subject. Table 2.2 shows the article number and the journal number with different numbers of subjects. 74 journals are assigned with 2 subjects, Statistics & Probability and Mathematics. However, there are 220 journals which only assigned with Mathematics.

The articles published in the journal which assigned with statistics are considered as the statistics article. All the other articles in other subjects will follow this rule. Since it is almost impossible to find out which single subject the article belongs to. Articles will be treated as the articles in different subjects based on which subject the journal belongs to. If a journal is assigned with 5 different subjects, the articles in this journal will be treated as 5 papers in 5 different subjects when doing the calculation.

Note that the network database we used is from 1981 to 2016, so any articles which published before 1981 is not in the database.

Table 2.2: Number of statistics related journal

Number of Assigned subjects	2	3	4	5	6	7	8	9
Number of Journals	74	20	17	14	13	1	3	2
Number of Articles	104229	14890	26234	17769	20204	1158	13817	4997

2.2 Characteristics of citation situation in each academic field

In this section, we will analyze the citation situation of each field. First, let us see the average cited rate of each field.

$$\text{Average cited rate} = \frac{\text{Number of citations in the subject}}{\text{Total number of articles in the subject}}. \tag{2.1}$$

Figure 2.1 shows the average cited number of each subject. Since it is hard to see the detail of this figure, Figure 2.2 shows the top 20 and bottom 20 subjects of the average cited number. In the Astronomy & Astrophysics field, each article received an average of 29.08 citations. In some other subjects, each article received around 13.16 citations. The average citations an article received within Statistics & Probability is 5.42, and it seems like a little less than other subjects. It might be because of the lack of data. Since the data we used is from 1981 to 2016, and there is no any data before 1980. Even if the articles published before 1980 are used as references, those citations will not be counted in the calculation. Those subjects which reach to the top in these metrics might be considered to have a tendency of using newly published articles as references than the older published articles. Most of the subjects in the bottom 20 are literature related subjects, and the average cited number is below 1. It can be considered as either they tend to use the old articles as reference or they tend to not to use many references.

The next metric shows the proportion of the unique number of articles on the subject being cited.

$$\text{Cited article rate} = \frac{\text{Unique number of articles in the subject being cited}}{\text{Total number of articles in the subject}}. \tag{2.2}$$

Figure 2.3 shows the cited article rate in each subject, and Figure 2.4 shows only

top 20 and bottom 20. These metrics indicate the portion of the article which is not independent in the network. At least half of the subjects have more than half of their article being cited at least once. Astronomy & Astrophysics also reach the top 1 in this metric that around 84 % of the article in this field received at least one citation. Unsurprisingly, literature related subjects are listed in the bottom 20. It seems like there are many independent articles exist in the literature related subjects.

The rank is a little different between the top 20 of Figures 2.2 and 2.4. In Figure 2.2, Neurosciences, Neurosciences & Neurology, Biochemistry & Molecular Biology, Virology, Geochemistry & Geophysics, Management, and Psychology reached the top 20 list, but they did not enter the top 20 in Figure 2.4. It indicates that even though there are many citations occurred within their own subject, there are quite an amount of independent articles which do not receive any existing citations. On the other hand, Chemistry, Inorganic & Nuclear, Parasitology, Polymer Science, Chemistry, Analytical, Fisheries, Oceanography and Materials Science, Biomaterials reached the top 20 in Figure 2.4 but did not enter the top 20 in Figure 2.2. It shows that there is not that many independent articles exist compare to their average cited rate.

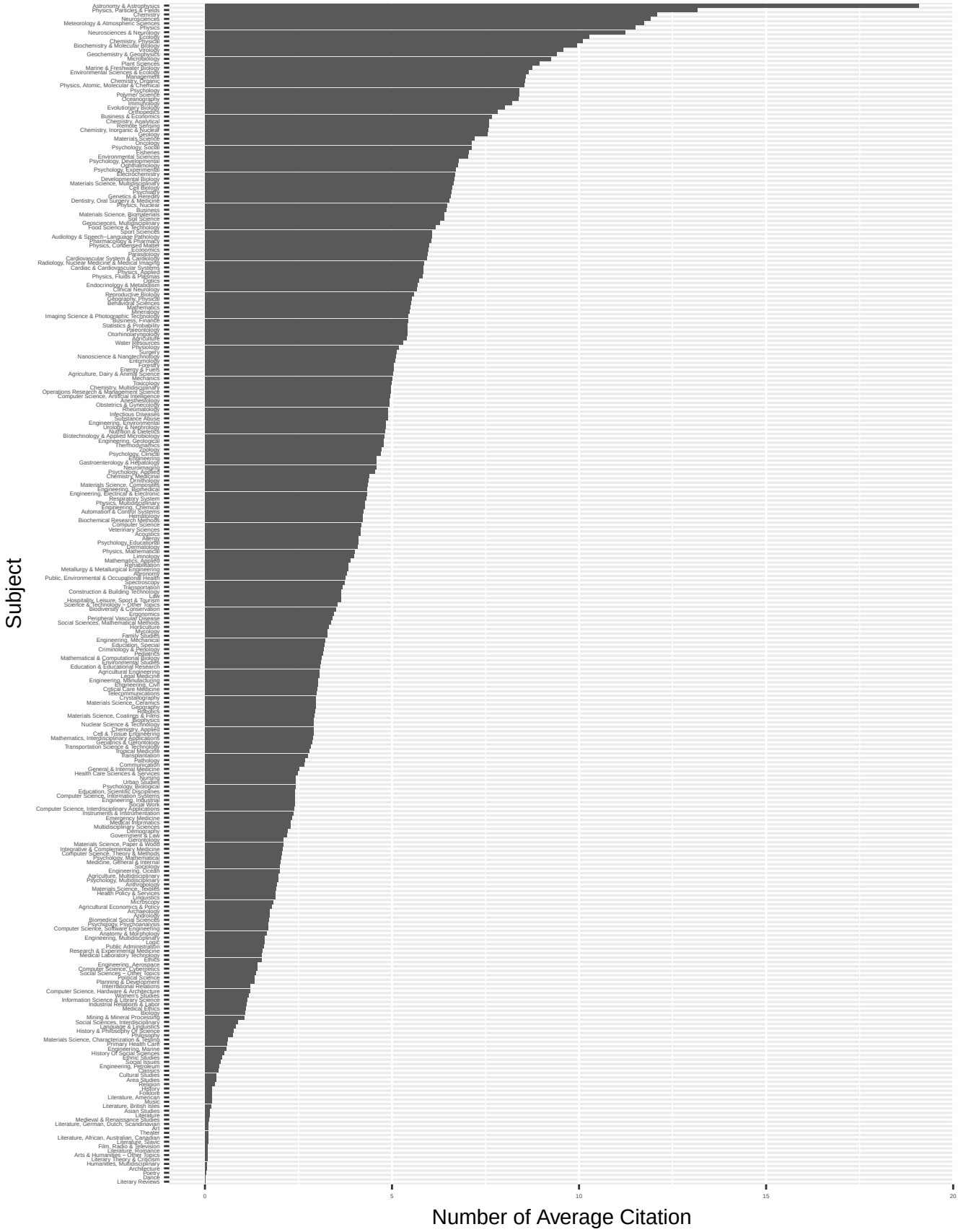

Figure 2.1: Average cited number of each subject

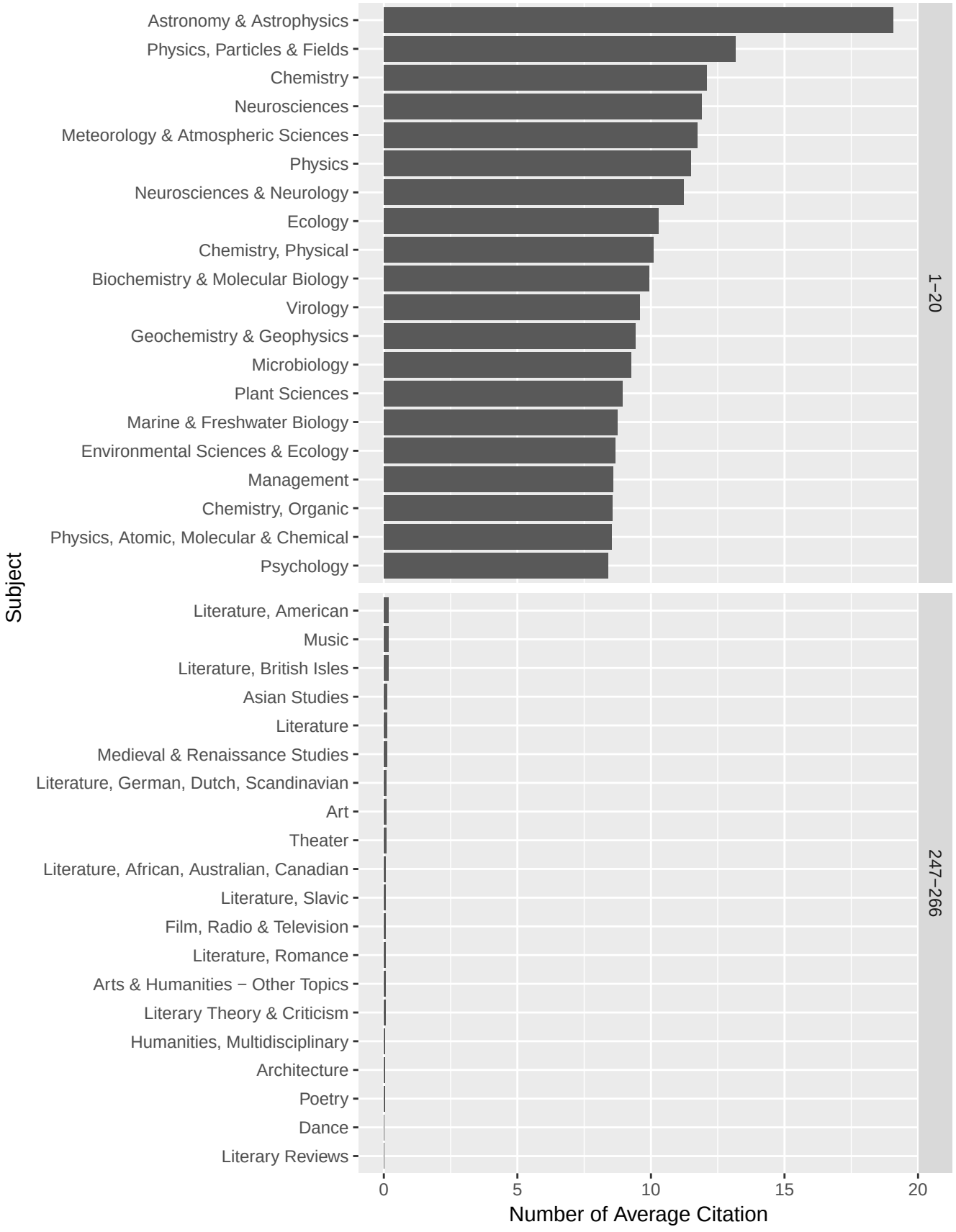

Figure 2.2: Top 20 and bottom 20 average cited number of each subject

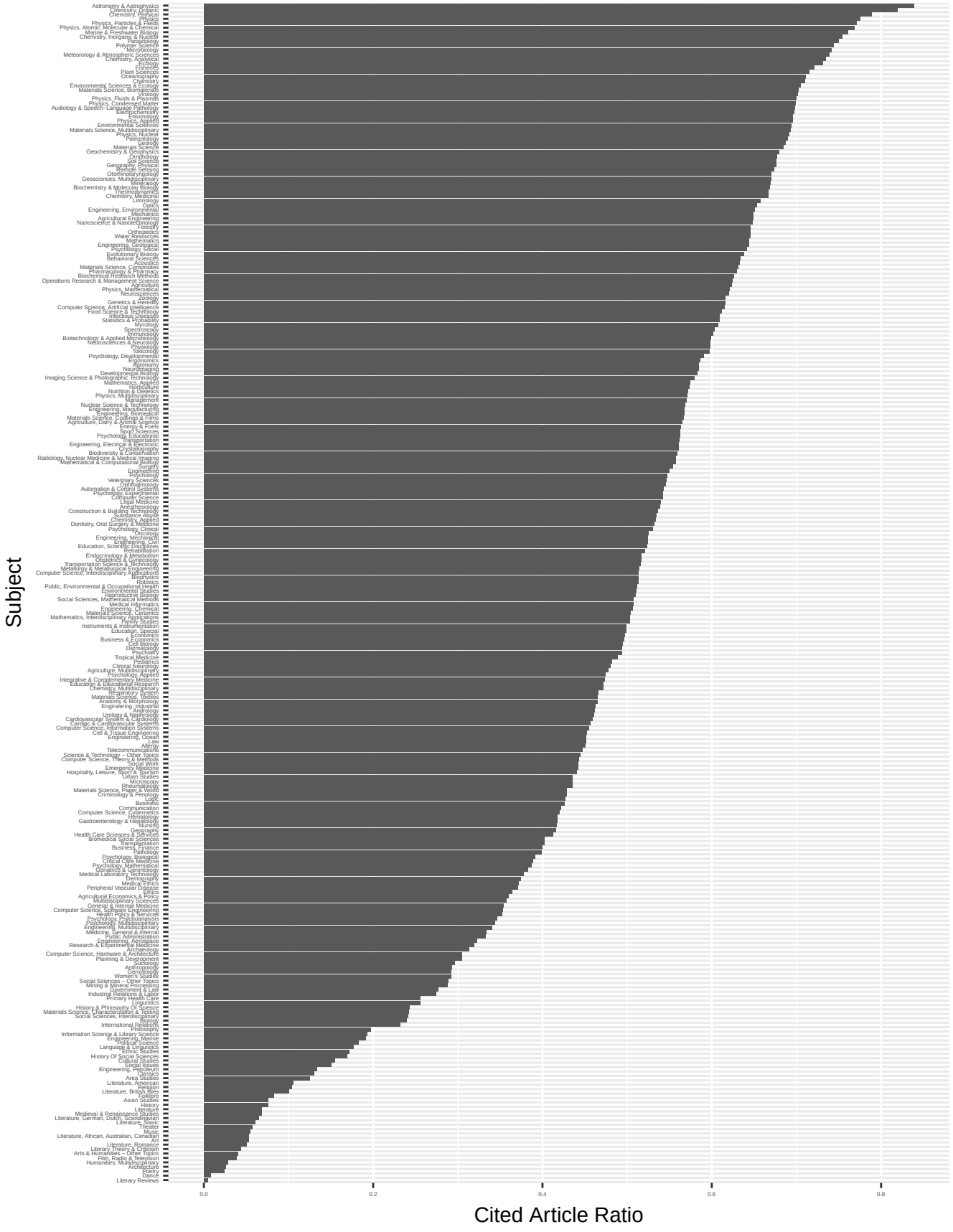

Figure 2.3: Cited article rate in each subject

Figure 2.4: Top 20 and bottom 20 of Cited article rate in each subject

2.3 The phenomenon of statistics article usage in in different academic field

Now we want to know how other fields use statistics articles as the reference. First, we investigate the number of statistics paper one can use as the reference in each subject.

$$\text{Average statistics cited rate} = \frac{\text{Number of citations from each subject to Statistics}}{\text{Total number of articles in the subject}}. \tag{2.3}$$

This metrics shows the statistics usage of different subjects.

Figure 2.5 shows the number of average statistics citations. Figure 2.6 shows the top 20 and bottom 20. Social Sciences, Mathematical Methods is the subject which using statistics article as the reference the most. On average, a Social Sciences, Mathematical Methods article cites more than 2 statistics articles. Among the top 20 subjects, there are 6 subjects related to mathematics. The bottom 20 are all literature related subjects. In addition, the hottest topic of computer science, artificial intelligence reaches the 9th place in this metric, and it is also well known for utilizing statistics technique in the computation.

However, this metric is effected by the average citation in each subject. If articles in one subject tend to use fewer references, then they will also use fewer statistics articles as the reference. The citation occurred within each subject will be used for standardization.

$$\begin{aligned}&\text{Standardized average statistics cited rate}\\ =&\frac{\text{Number of citations from subject to Statistics}}{\text{Number of citations within the subject}}.\end{aligned} \tag{2.4}$$

Figure 2.7 shows the standardized statistics usage of each subject. Figure 2.8 shows the top 20 and bottom 20 of standardized statistic usage. Mathematical & Computational Biology reach the top 1 in this metric, the rate of Mathematical & Computational Biology articles cite themselves and statistics is 1 versus 0.8. Also, Biology, Computer Science, Cybernetics, Research & Experimental Medicine, and Industrial Relations & Labor enter the top 20 in Figure 2.8 but do not enter the top 20 in Figure 2.6. Those 4 subjects can be considered as using quite a lot of statistics articles as reference compares to themselves. Oppositely, Business &Economics, Evolutionary Biology, Automation & Control Systems, and Mathematics, Applied enter the top 20 in Figure 2.6 but did not show up in the top 20 list in 2.8. It indicates that even if they use many statistics articles as reference, but it actually does not have that many when compares to the citations within its own subject.

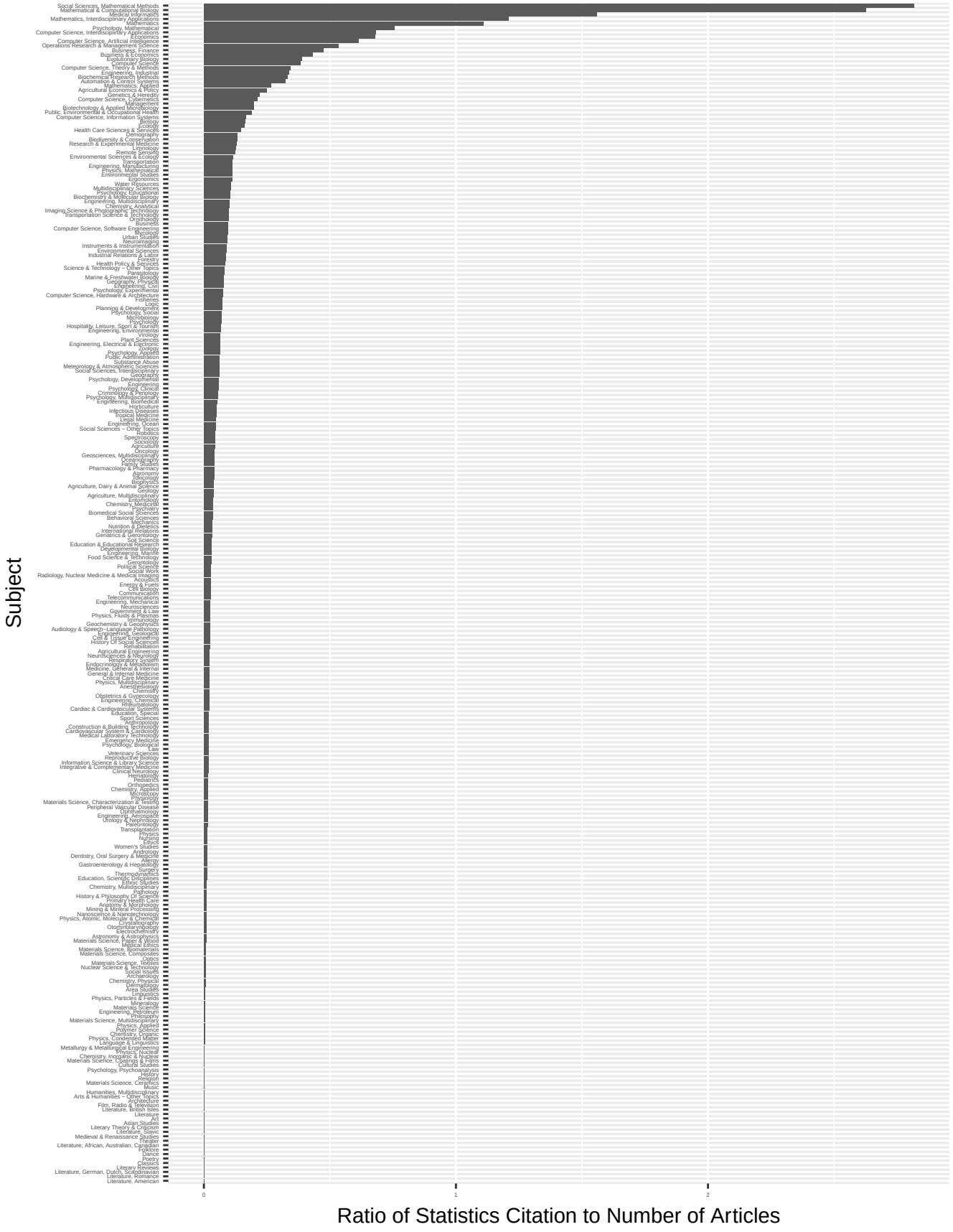

Figure 2.5: Average citations from other subject to statistics

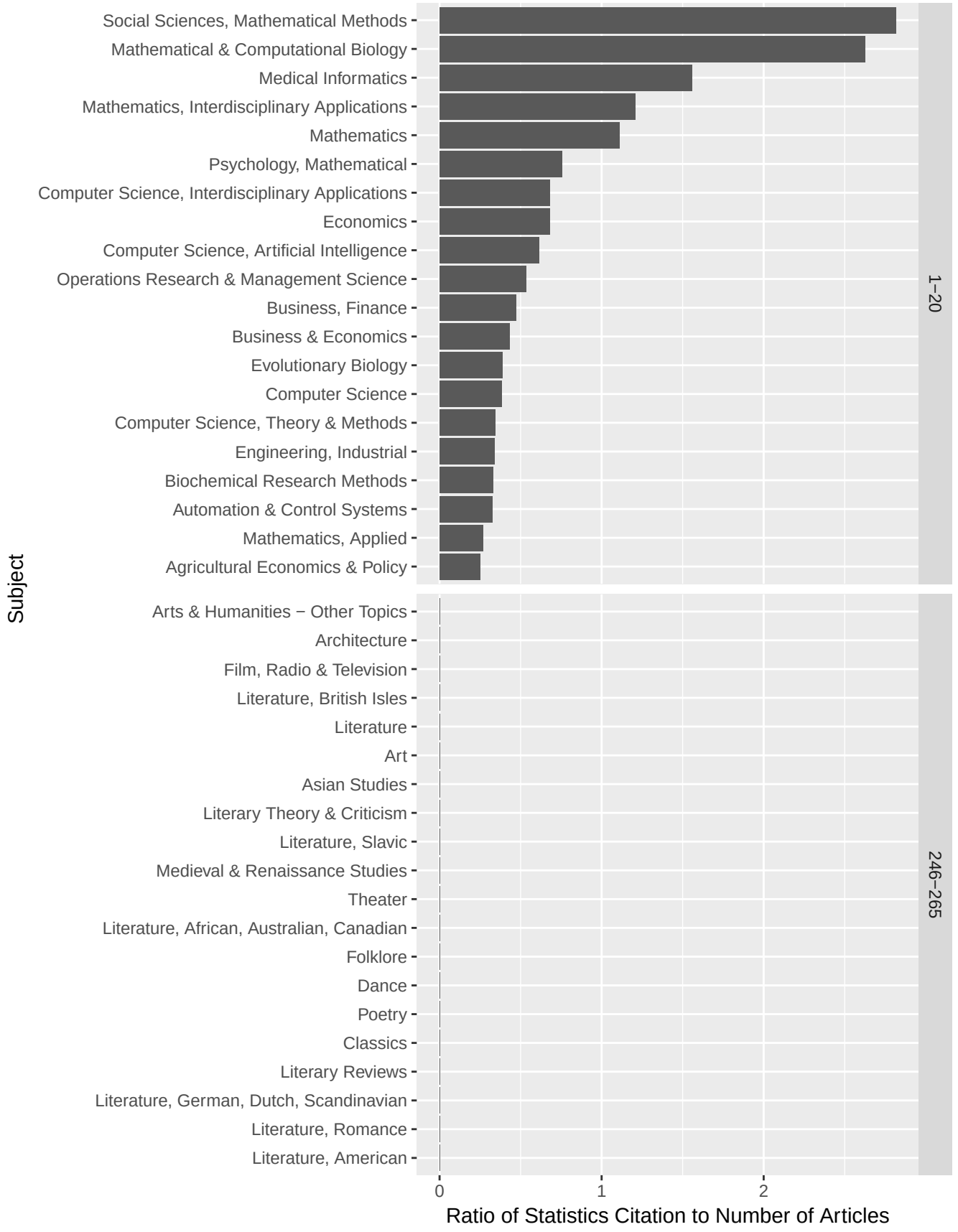

Figure 2.6: Top 20 and bottom 20 of average citations from other subject to statistics

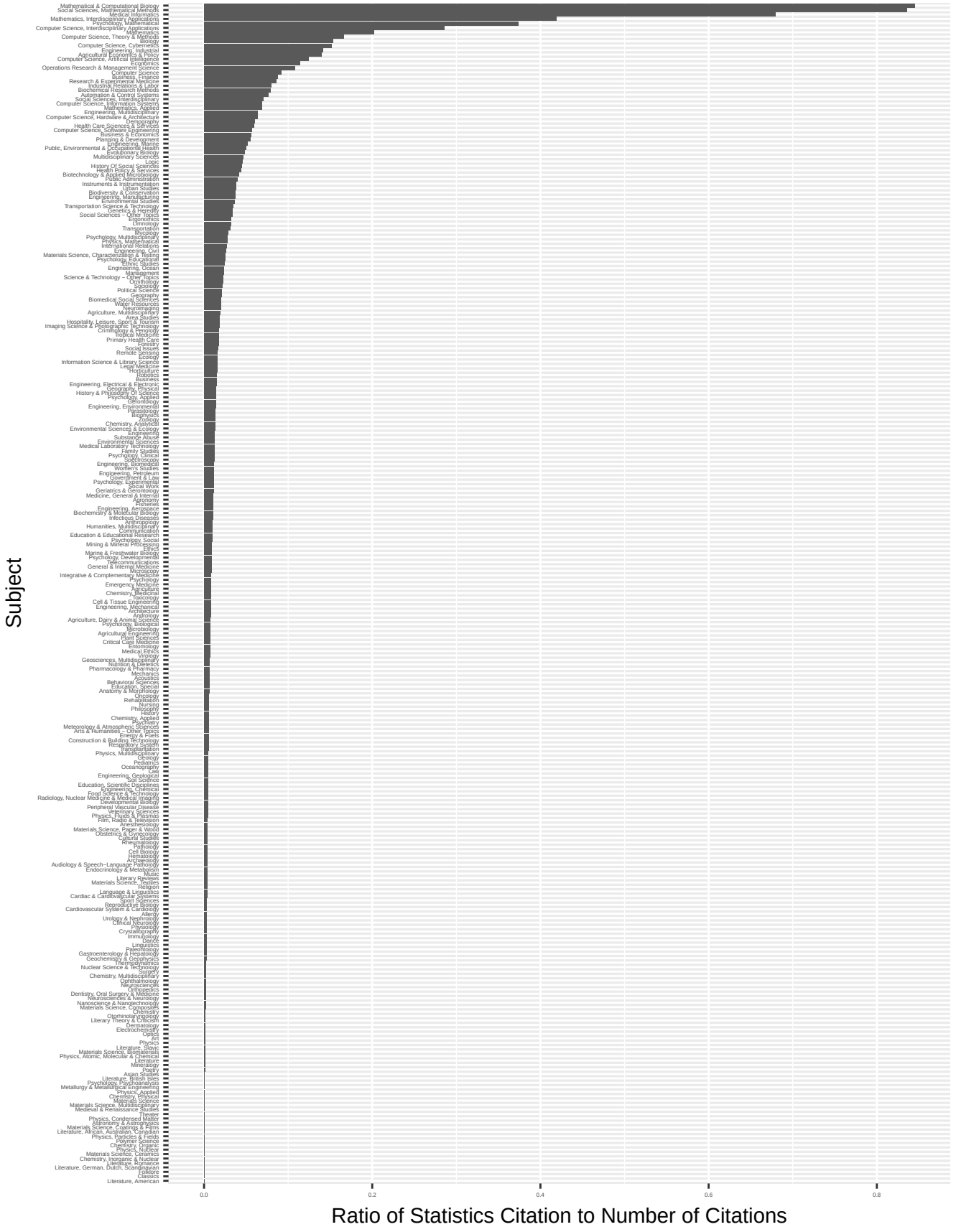

Figure 2.7: Standardized statistics usage

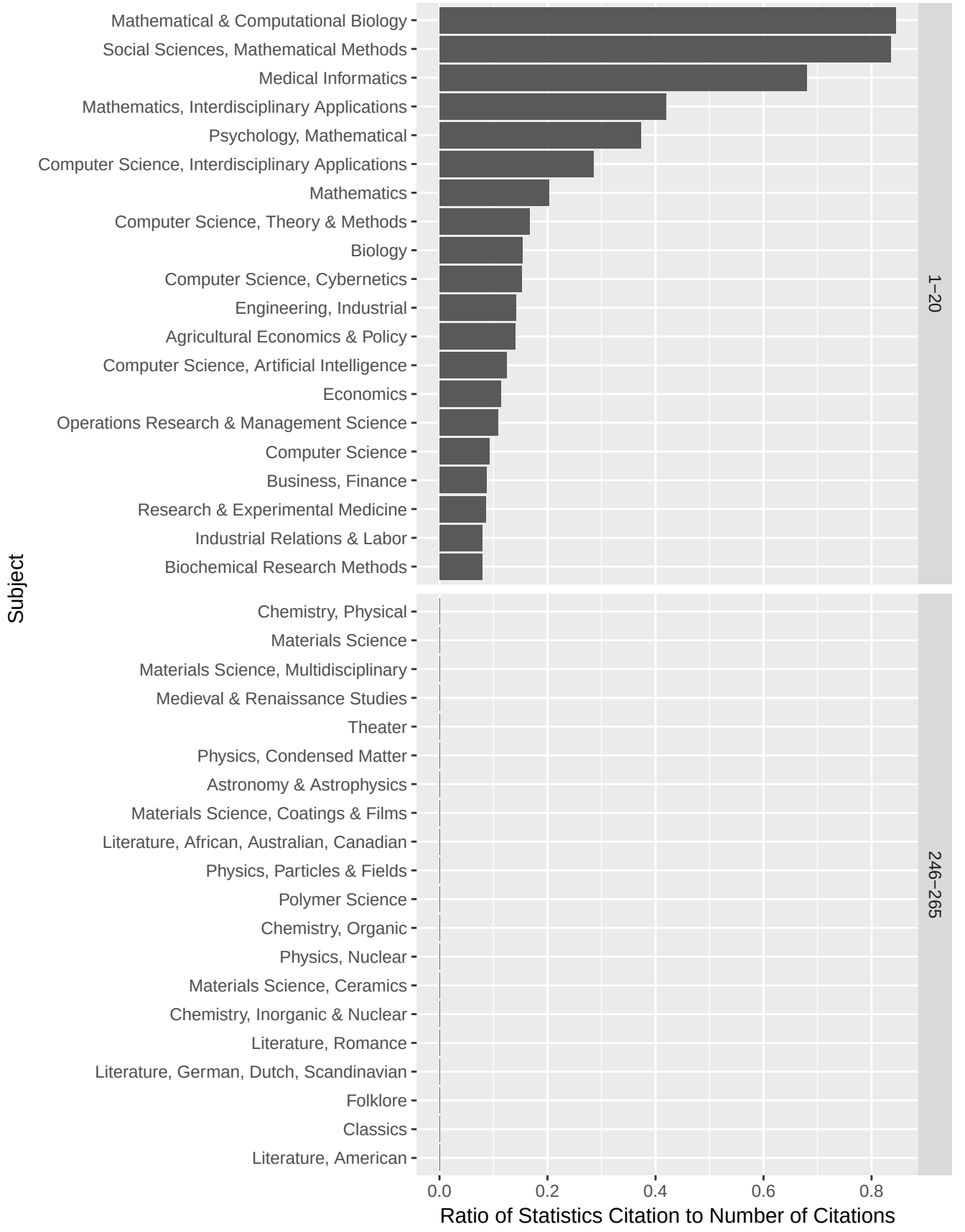

Figure 2.8: Top 20 and bottom 20 of standardized statistics usage

Now, we want to know the number of unique statistics articles received citations.

$$\text{Unique statistics usage rate} = \frac{\text{Number of unique statistics article being cited}}{\text{Total number of articles}}. \tag{2.5}$$

In Figure 2.10, Computer Science, Cybernetics, Computer Science, Information Systems, Ergonomics, and Management enter the top 20 list, but they did get into the top 20 list in Figure 2.6. It indicates that those 4 subjects tend to use different statistics articles as reference, but the citation number is not that many. On the opposite, Economics, Business & Economics, Evolutionary Biology, and Biochemical Research Method enter the top 20 list in Figure 2.6, but did not get into the top 20 list in Figure 2.10. It indicates that the statistics paper those subjects used as the reference tends to concentrate on certain papers.

Next, the unique number of articles on the subject being cited is used to do the standardization.

$$\begin{aligned}&\text{Standardized unique statistics usage rate}\\ =&\frac{\text{Number of statistics paper being cited by papers of a subject}}{\text{Unique number of articles in the subject being cited}}.\end{aligned} \tag{2.6}$$

Figure 2.11 and Figure 2.12 showed the result of metrics. Most of the articles tend to use articles from the same subject as the reference. However, the metric of Social Sciences, Mathematical Methods is larger than 1, it indicates that they cite more statistics papers than from their own subject. Also, there are 6 subjects related to mathematics among the top 20 list. Biology, Demography, and Computer Science, Software Engineering do not rank high in Figure 2.10, but reach the top 20 list in this metric. It indicates that even though they do not use a lot of articles as the reference, they use relatively many statistics article as references. In the last few metrics, Computer Science, Ergonomics, and Management ranked in the top 20 but do not rank high in this metric. It shows that they do not use the varieties of statistics paper as the reference compare to the unique paper they used in their own field. In addition, Automation & Control Systems, Demography, Computer Science, Information Systems, Computer Science, Software Engineering, and Mathematics, Applied do not rank high in Figure 2.8 but reach the top 20 list in this metric. It indicates that there are not many citations toward statistics, but they use the varieties of statistics papers as the reference. On the opposite, Economics, Computer Science, Research & Experiment Medicine, Industrial Relations & Labor, and Biochemical Research Methods ranked high in Figure 2.8 but do not reach the top 20 list in this metric. It indicates that citations are concentrate on certain articles.

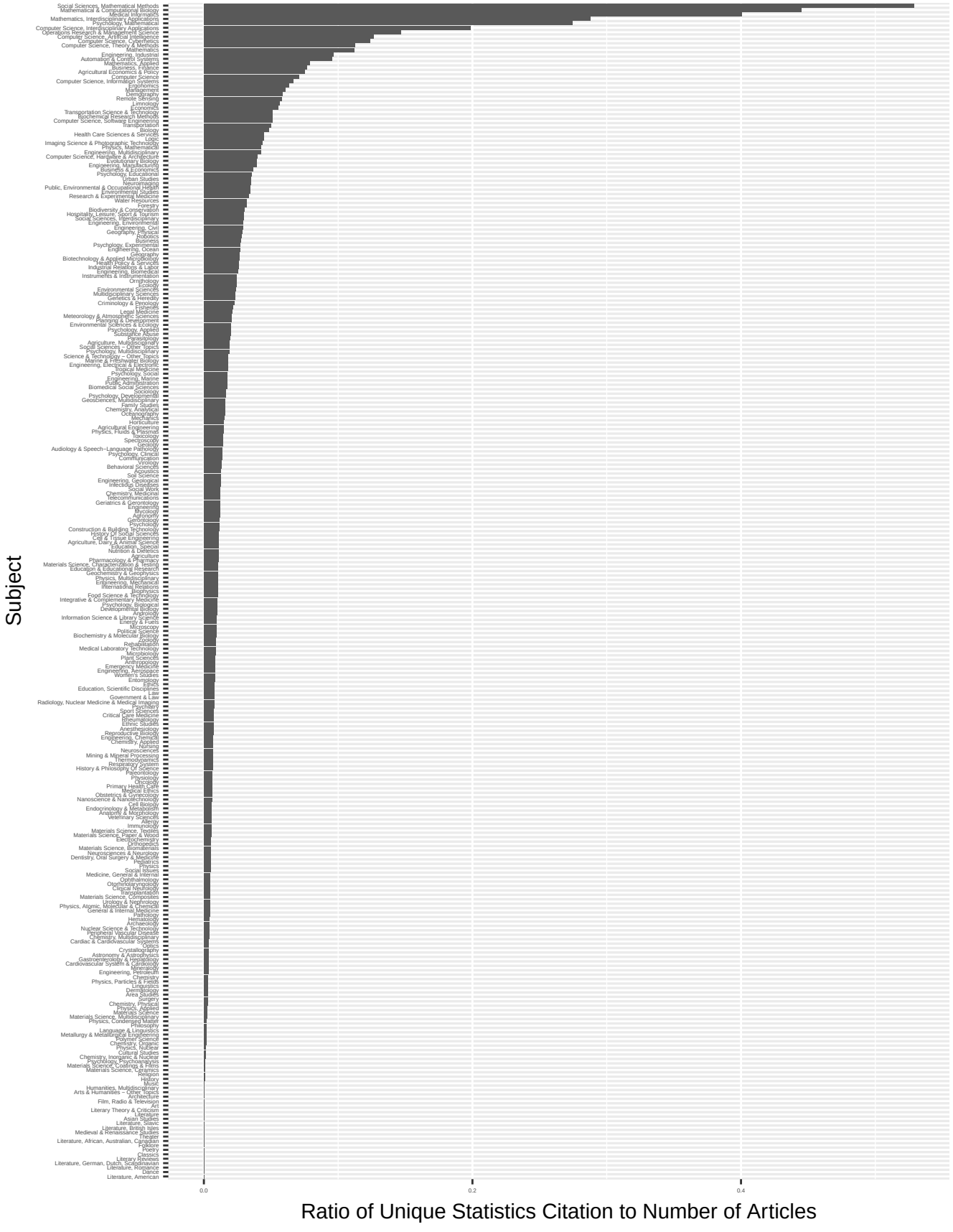

Figure 2.9: Number of unique statistics article being cited

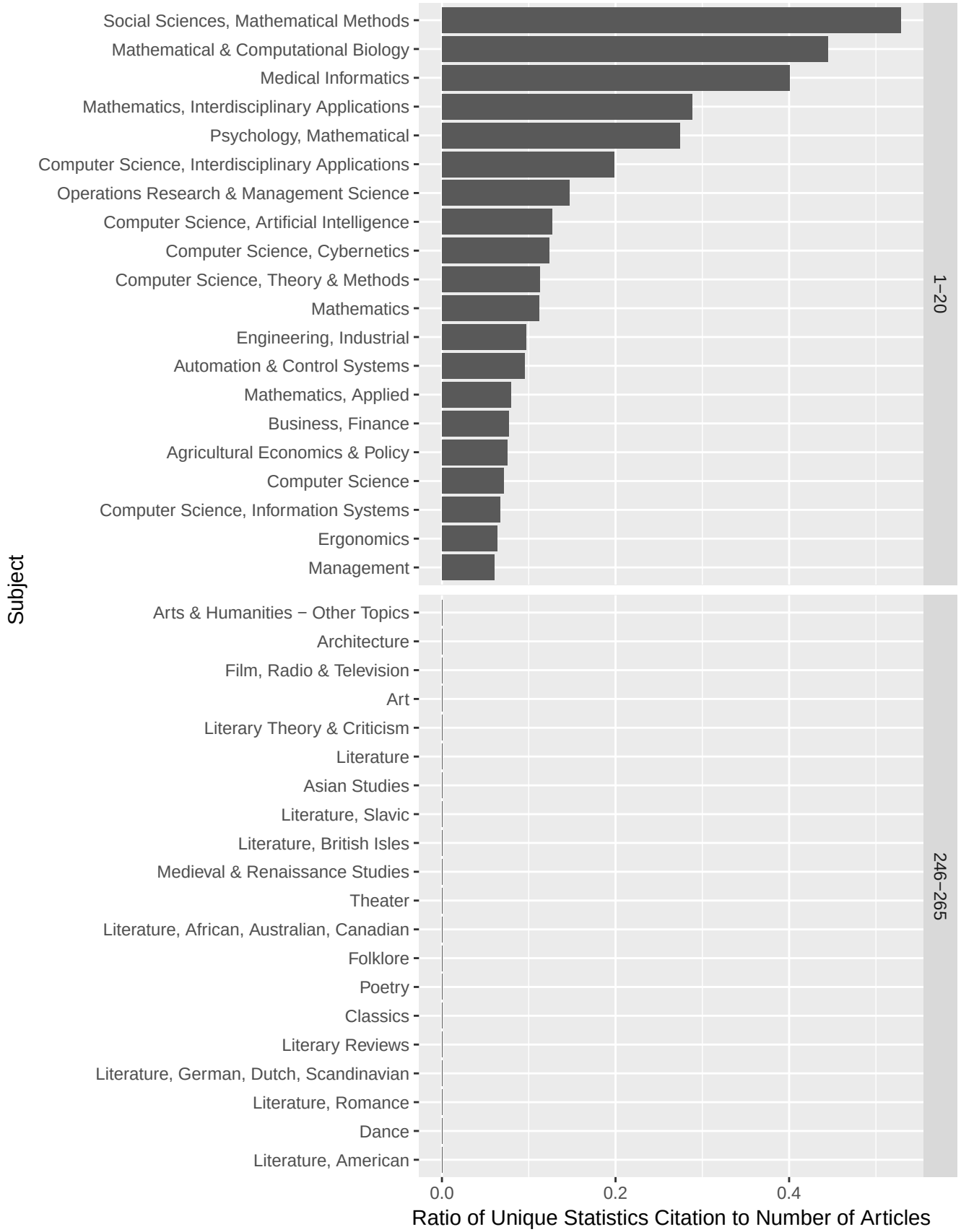

Figure 2.10: Top 20 and bottom 20 of number of unique statistics article being cited

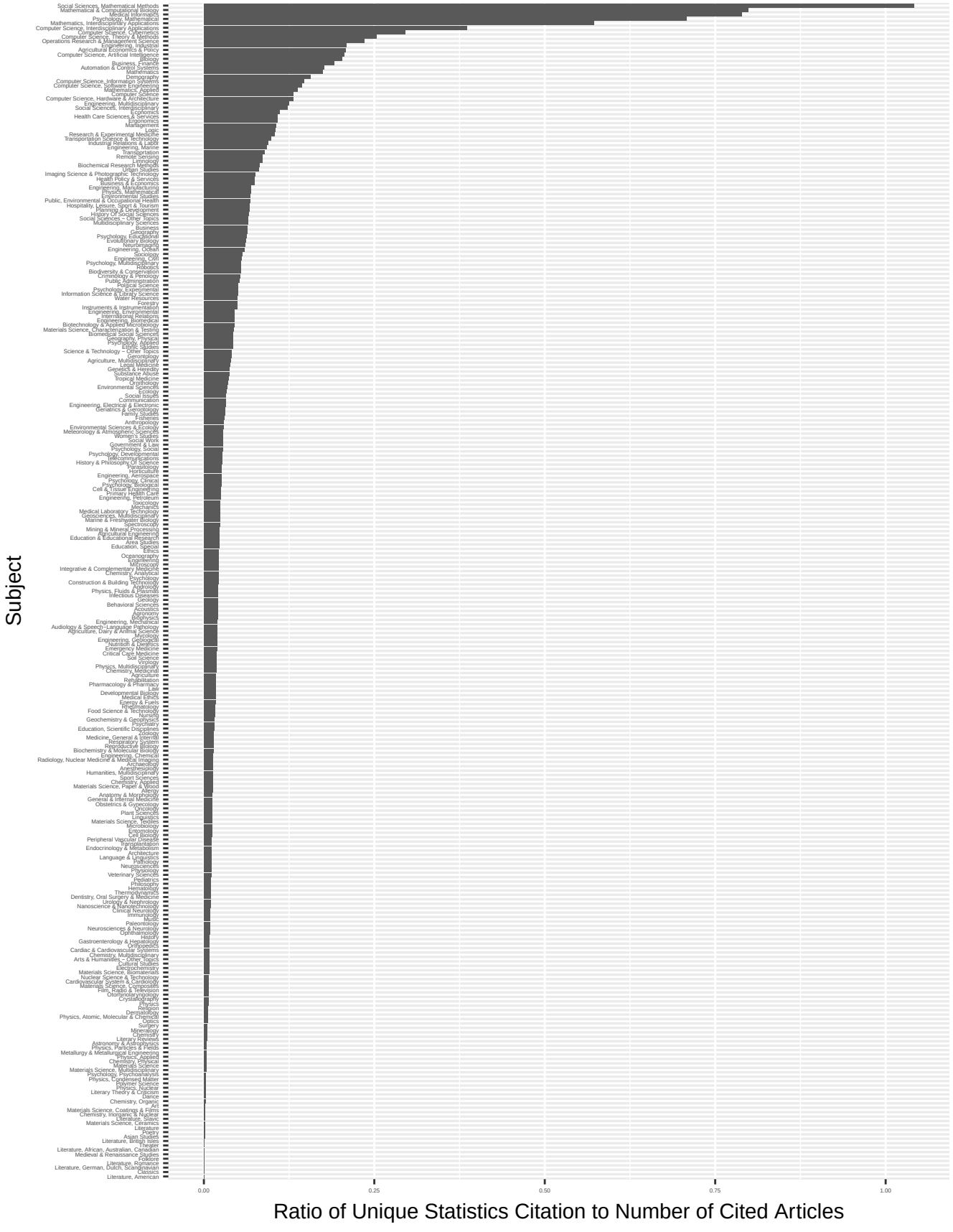

Figure 2.11: Number of statistics paper being cited by papers of a subject

Figure 2.12: Top 20 and bottom 20

2.4 Summary

It is clear that the citation situation is different for each subject. Some subjects tend to use a lot of citations and some are not. For example, around 53% of subjects have more than half of their articles being cited at least once. It indicates that there are not many isolated articles in those subjects. On the contrary, there are many isolated articles, which do not receive any citations, in the subjects related to Literature or Art.

Next, the relationship between statistics and other subjects is investigated. We can find that some subjects have a close relationship with statistics and use statistics articles as the reference a lot. We observe that the articles in Social Science, Mathematical Methods use many statistical articles as reference. In fact, we know that there is no journal which is only assigned to statistics from Table 2.2. That is, journals which are assigned to Statistics & Probability also come with other subjects. We may guess that the subjects which have higher statistics usage have many of their articles overlapped with statistics.

We now know how important statistics is toward other subjects. In the next step, we will focus on the statistics field and find out the influential article within the statistics subject.

Chapter 3

A New Metric for the Analysis of the Scientific Article Citation Network

Many conventional methods listed in Chapter 1 consider only the "direct citation" and ignore the "indirect citation". The formal definitions of the citation directness will be given in the next section, and we use an example to describe the scenario here. Assume we have a small citation network with three articles and two citations, where Article A was cited by Article B and Article B was cited by Article C. If we measure the influence of paper A in this network, conventional methods only take the direct citation into consideration and conclude that Article A has influences only on Article B. However, it is very likely that Article A inspired Article C through Article B. This scenario might happen when one of the above two citations (or both) was interdisciplinary, or Article B was a review-type article. For the past hundred years, scientific advancement was built on the continuous propagation of one's research results to others, and conventional methods are limited to quantify this phenomenon.

We aim at analyzing the citation network from the network perspective. In fact, there is much research investigates the influence from the network perspective, such as Gualdi et al. (2011). Influence is a capacity or power of things to be a compelling force or to produce effects on the actions or opinions of others. In the citation network, when a researcher published a scientific article, others read it and followed his step. When these followers completed their own research that was influenced or inspired by the original article, they cited it in the reference when they wrote their scientific articles. This led to a citation, which is officially defined as an abbreviated alphanumeric expression embedded in the body of intellectual work that denotes an entry in the bibliographic reference section of the work for the purpose of acknowledging the relevance of the work of others to the topic of discussion at the spot where the citation appears. After years of progress, a citation network has resulted. We introduce

a new influence-based metric for the importance of an article, and the importance is in terms of the influence of an article towards all other articles in its associated field or the whole citation network.

3.1 Article Network Influence (ANI)

Let $G = (V, E)$ be a citation network, where $V = \{v_1, \ldots, v_m\}$ is a set of m articles (nodes), and E is a set of citations (directed edges). In specific, we denote $v_t \to v_s$ in G when Paper v_t cites Paper v_s. Here are some special properties of the citation network:

1. All edges in G are directed with only one direction. Since an edge between two nodes in G represents a citation relationship between two articles, it is obvious that the direction of the edge indicates which article cites one another. We do not expect to see a two-way edge in G, because it is unlikely to have a situation of "A cites B while B cites A".

2. There is no self-connection on any nodes in G. The self-connection of a node in G represents the self-citation of an article which is unlawful in nature. We ignore the existence of self-connection in our network and treat them as incorrect entries.

We define a *citation path* between two articles as a finite sequence of citations that connect a set of distinct articles. Let $S \subseteq G$ be an ordered sequence of articles appearing along the path, i.e., the first and last articles in S are the articles of interest, the length of a citation path between two articles, or *pathlength*, in short, is defined as $l_S = |S| - 1$, where $|S|$ is the number of nodes of S. Note that the citation path between two articles is not necessarily unique. Assume that there are k different citation paths between two articles, denoted as $l_{S_1}, \ldots, l_{S_k}$, we define the *influence range* between two articles as $r = \min(l_{S_1}, \ldots, l_{S_k})$.

Let v_t be an article of interest, where $1 \leq t \leq m$. Then the Article Network Influence (ANI) of v_t, denoted as ANI_t, is

$$ANI_t = \sum_{r=1}^{r_M} g(r) o_t(r), \tag{3.1}$$

where r is the influence range from v_t to all other articles in G, r_M is the maximum influence range of interest from v_t, $1 \leq r_M \leq m$, $o_t(r)$ is the observed number of articles with the influence range r from v_t, and $g(r)$ is a weighted function to represent the decay effect of article citations as the influence range increases.

$o_t(r)$ can be easily obtained from the citation database via calculating the number of citations of v_t received from articles with the influence range r. To obtain $g(r)$, we first denote $a_t(r)$ as the number of articles which pathlengths from v_t are exactly r for $t = 1, \ldots, m$. The estimated average number of citations of all m articles within the pathlength r over the whole citation network G is

$$e(r) = \frac{\sum_{t=1}^{m} a_t(r)}{m}. \tag{3.2}$$

Then we define the weight function $g(r)$ as the normalization of a function $g'(r)$ below. If $e(r-1) \leq e(r)$ for $r = 1, ..., r_X$ and $e(r_X) > e(r_X + 1)$, then $g'(r) = e(r)^{-1}$ for $r = 1, ..., r_X$ and $g'(r) = e(r_X)^{-1}$ for $r = r_X + 1, ..., r_M$. Note that it is possible that $r_X = r_M$, i.e., $g(r)$ is a monotonically decreasing function within $\{1, \ldots, r_M\}$.

Generally speaking, the number of citations with the influence range r will increase as r increases. Therefore, the value of a single citation, which is proportional to the reciprocal of the average number of citations becomes less important as the influence range r increases. However, this phenomenon does not always occur because a paper may be cited by minor papers that do not receive any other citations. It leads to a stop in the increase in the average number as the influence range r increases. The boundary indicates that the paper becomes less important at the point r_X, so we fix $g'(r)$ as $g'(r_X)$ when the value starts to increase.

3.2 An Analysis of the Citation Network of Statistics Research Articles

3.2.1 Data Preparation

Data from the WoS database will be used to conduct the analysis. The introduction of the WoS database was explained in the previous chapter. Given a targeted subject, we first extract from the database that all articles are labeled as the targeted subject. The data subset is then downloaded, and the resulting file is in the csv file that is structured as a table of two columns. Each row of the table represents a citation between two articles. In specific, the articles listed in the first and second columns are the articles cited and the articles being cited. The rest of the analyses are conducted using R (R Core Team (2017)), especially the R package "igraph" (Csardi and Nepusz (2006)).

3.2.2 Interpretation and Analysis

We applied our method to the citation network of statistics research articles, which can be obtained via the extraction of data with labeled "Statistics" in the whole WoS database. All the data are first divided into the 11-year time span, which is equivalent to a 10-year period after publication. By dividing the total data into small segments, the historical trend might be able to be observed, and we can also avoid the situation that all early published papers are on the top, and the rest of papers are invisible from the list. Therefore, we have 26 networks from the year 1981-1991 network, the year 1982-1992 network,..., until the year 2006-2016 network. ANI is applied to each network, and the results are listed in the appendix. There are totally 26 tables in the supplementary material, each table lists the top-20 influential articles from each network. Below are some observations from these tables.

First, most of the top-20 articles were published early when compared to the range of years being considered. For example, all 20 articles in Table 3.1 (1981-1991) were published in either 1981, 1982, or 1983. The most extreme of this happened in Table 3.15 (1995-2005), where only one article was published in 1996, and the rest were published in 1995. Similar extreme tables include Tables 3.16 (1996-2006) and 3.17 (1997-2007). This phenomenon is reasonable because the influence of an article, especially in statistics or most mathematical sciences, needs time to accumulate. It is highly unlikely to have an article which influence is high enough to jump into the top 20 list in its first year of publication. From this perspective, we can actually identify some exceptional articles from these tables. For example, "Sampling-based approaches to calculating marginal densities"(Gelfand and Smith (1990)) , published in 1990, was the only article being on the list six times in all possible eleventh years. It entered the top-20 list (ranked the 17th) in its fifth year (Table 3.5), and its rank improved to the 7th in its sixth year (Table 3.6) and became the 1st in the next four years (Tables 3.7-3.10). Several excellent articles entered the list 4-5 times with consecutive first ranks before their eleventh year of publication.

Second, some articles with a smaller number of direct citations are ranked higher than others with a larger number of direct citations. For example, in Table 3.25 (2005-2015), the first-rank article received 552 direct citations from other articles while the second-rank article received 691 direct citations. For most research metrics that consider only direct citations, the latter article should rank on top of the former article. Oppositely, ANI considers indirect citations with various pathlengths, and it reverses the rank between these two articles because of the much larger number of indirect citations of the first-rank article, which can be interpreted as an implicit spread of its idea to a broader group of audiences or authors in the statistics community. In fact, such phenomena appear quite often in these 26 tables.

For a deeper insight, the topics of these top-20 articles listed in these 26 tables can be viewed as the historical development of statistical research in the past 36

years. For example, the first article related to lasso that reached the top-20 list was published in 2006, and it appeared early in Table 3.22 (2002-2012). It was its fourth year after publication, and it indicated the importance and attractiveness of lasso in the statistics community. We found many articles related to lasso appeared in the top-20 lists in the latter years. For example, there were seven lasso-related articles in Table 3.26 (2006-2016). Similar hot and representative topics, including regression, the Bayesian method, and many others, governed the recent developments of statistics.

In opposite, the first article related to the Gibbs sampling that reached the top-20 list appeared in Rable 3.7 (1987-1997). It received numerous attention since then and reached its peak in Tables 3.10 (1990-2000) to 3.12 (1992-2002). We observed a diminishing interest towards this topic thereafter, and this topic disappeared completely from the top-20 list after Table 3.19 (1999-2009). A similar phenomenon also appears in articles related to microarray. The first article appeared in the top-20 list was in Table 3.18 (1998-2008), and it reached its peak in Table 3.20 (2000-2010), then the interest diminishes until its disappearance from the top-20 list in Table 3.25 (2005-2015). It is not trivial to explain in deep why such phenomena of diminishing interests appear, especially in these two topics that might still be popular until now. Some believe that their attractions remain constant while new and hot topics arise and attract more attention, but this guess requires careful evaluations by experts in related fields.

Table 3.1: Top 20 Articles from 1981 to 1991 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Natural exponential-families with quadratic variance functions	1982	41	229	524	466	235	81	192.82
2	Logistic-regression diagnostics	1981	90	310	424	289	108	17	188.80
3	A large sample study of cox regression-model	1981	48	300	383	287	187	67	173.47
4	2 graphical displays for outlying and influential observations in regression	1981	42	259	419	303	119	22	159.33
5	Projection pursuit regression	1981	62	269	305	252	177	47	158.37
6	Efficient bounded-influence regression estimation	1982	63	223	283	263	157	68	147.61
7	Censored-data and the bootstrap	1981	43	138	257	310	296	255	147.37
8	Some asymptotic theory for the bootstrap	1981	116	281	196	86	43	32	145.24
9	Parametric empirical bayes inference - theory and applications	1983	81	192	253	252	110	62	140.29
10	Empirical choice of histograms and kernel density estimators	1982	62	229	306	214	101	31	139.77
11	Survival times - aspects of partial likelihood	1981	22	261	314	251	163	41	138.90
12	Asymptotic theory of non-linear least-squares estimation	1981	32	137	477	305	105	39	136.96
13	Monte-carlo study of 3 data-based nonparametric probability density estimators	1981	28	160	352	306	186	97	135.86
14	On the asymptotic accuracy of efrons bootstrap	1981	103	258	182	117	49	25	135.78
15	Odds ratio estimators when the data are sparse	1981	52	147	279	309	214	75	135.48
16	Nonparametric maximum-likelihood estimation by the method of sieves	1982	31	144	340	369	181	52	133.98
17	An optimal selection of regression variables	1981	40	257	291	188	89	24	130.61
18	Cox regression-model for counting-processes - a large sample study	1982	111	213	191	124	40	12	129.84
19	Data-based optimal smoothing of orthogonal series density estimates	1981	21	119	323	334	203	107	125.33
20	On a formula for the distribution of the maximum-likelihood estimator	1983	53	209	273	213	84	14	125.07

Table 3.2: Top 20 Articles from 1982 to 1992 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Natural exponential-families with quadratic variance functions	1982	47	288	723	700	348	154	268.27
2	Efficient bounded-influence regression estimation	1982	72	279	393	413	267	120	202.82
3	Parametric empirical Bayes inference - theory and applications	1983	93	253	380	391	215	107	197.63
4	Empirical choice of histograms and kernel density estimators	1982	72	288	444	331	170	62	191.05
5	Nonparametric maximum-likelihood estimation by the method of sieves	1982	33	178	447	538	331	101	186.62
6	Cox regression-model for counting-processes - a large sample study	1982	132	289	290	200	98	32	178.44
7	On a formula for the distribution of the maximum-likelihood estimator	1983	60	259	385	349	194	50	175.46
8	Quasi-likelihood functions	1983	55	406	330	139	42	14	163.28
9	Regression diagnostics, transformations and constructed variables	1982	52	218	428	354	101	11	158.56
10	Minimax aspects of bounded-influence regression	1983	28	168	360	403	245	116	154.36
11	Smoothing splines - regression, derivatives and deconvolution	1983	42	291	332	145	105	127	149.10
12	Optimal global rates of convergence for nonparametric regression	1982	79	239	275	147	144	51	144.44
13	Cross-validation in density-estimation	1982	19	147	422	372	185	74	142.70
14	Robust regression using repeated medians	1982	19	95	191	404	415	275	141.91
15	A leisurely look at the bootstrap, the jackknife, and cross-validation	1983	51	140	234	244	293	207	140.47
16	Gini contributions to the theory of inference	1982	3	111	258	379	385	211	136.66
17	Distributions of maximal invariants using quotient measures	1982	11	97	278	388	349	192	135.92
18	A robust comparison of biological shapes	1982	4	91	242	424	423	172	134.93
19	Nearest neighbor (NN) analysis of field experiments	1983	47	200	315	271	125	24	134.84
20	Information and asymptotic efficiency in parametric nonparametric models	1983	77	206	271	183	89	20	132.54

Table 3.3: Top 20 Articles from 1983 to 1993 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Parametric empirical Bayes inference - theory and applications	1983	103	343	528	560	336	165	264.06
2	On a formula for the distribution of the maximum-likelihood estimator	1983	70	317	506	529	353	101	235.42
3	Quasi-likelihood functions	1983	59	520	522	264	94	32	228.41
4	Minimax aspects of bounded-influence regression	1983	30	196	474	586	386	204	203.62
5	Smoothing splines - regression, derivatives and deconvolution	1983	45	347	468	252	162	198	196.03
6	A leisurely look at the bootstrap, the jackknife, and cross-validation	1983	55	160	329	347	426	316	181.16
7	Information and asymptotic efficiency in parametric nonparametric models	1983	86	260	394	302	167	54	180.79
8	Nearest neighbor (NN) analysis of field experiments	1983	49	234	447	410	216	64	178.84
9	On the convergence properties of the EM algorithm	1983	52	220	416	414	196	87	174.10
10	Consistent cross-validated density-estimation	1983	31	244	474	419	159	70	173.20
11	Local sufficiency	1984	39	344	393	265	112	46	169.88
12	Large sample optimality of least-squares cross-validation in density-estimation	1983	48	247	438	343	142	57	168.41
13	An alternative method of cross-validation for the smoothing of density estimates	1984	62	261	447	230	86	38	163.49
14	Projection pursuit	1985	76	240	353	271	121	30	159.52
15	Estimating the error rate of a prediction rule - improvement on cross-validation	1983	52	153	245	406	336	189	157.15
16	Some aspects of the spline smoothing approach to non-parametric regression curve fitting	1985	88	280	237	233	124	38	157.06
17	Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum-likelihood estimator	1984	59	257	394	289	62	12	156.26
18	Regression, prediction and shrinkage	1983	47	144	333	391	247	54	145.53
19	Bandwidth choice for nonparametric regression	1984	89	270	280	117	42	15	142.95
20	A note on the modified likelihood for density-estimation	1983	4	93	372	496	339	120	141.19

Table 3.4: Top 20 Articles from 1984 to 1994 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Local sufficiency	1984	42	410	570	411	231	100	231.14
2	Some aspects of the spline smoothing approach to non-parametric regression curve fitting	1985	99	384	372	383	236	87	223.92
3	An alternative method of cross-validation for the smoothing of density estimates	1984	66	308	610	394	191	81	218.91
4	Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum-likelihood estimator	1984	62	327	563	496	148	32	216.71
5	Projection pursuit	1985	84	310	495	422	204	65	215.69
6	Accurate approximations for posterior moments and marginal densities	1986	104	427	444	196	30	1	202.90
7	Bootstrap confidence-intervals for a class of parametric problems	1985	37	349	455	348	175	54	189.21
8	Bandwidth choice for nonparametric regression	1984	99	335	366	211	84	52	182.88
9	Least median of squares regression	1984	105	175	309	396	238	75	176.77
10	Cross-validation in nonparametric-estimation of probabilities and probability densities	1984	18	184	560	444	229	84	174.99
11	An asymptotically optimal window selection rule for kernel density estimates	1984	69	268	446	238	124	45	170.84
12	A comparison of kriging with nonparametric regression methods	1985	14	153	420	369	368	233	168.03
13	Asymptotics of graphical projection pursuit	1984	19	121	399	556	312	160	166.57
14	Almost complete convergence properties of the kernel predictor	1984	21	140	418	363	359	227	166.37
15	The hat matrix for smoothing splines	1984	12	149	391	415	366	192	163.35
16	Probability-inequalities for empirical processes and a law of the iterated logarithm	1984	43	150	393	421	208	166	162.26
17	Projection pursuit density-estimation	1984	15	133	340	486	387	166	161.10
18	Time-series analysis	1984	7	124	406	376	372	229	158.09
19	The analysis of transformed data	1984	30	240	455	297	124	47	154.79
20	Approximate regression-models and splines	1984	8	102	379	372	383	236	152.02

Table 3.5: Top 20 Articles from 1985 to 1995 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Some aspects of the spline smoothing approach to non-parametric regression curve fitting	1985	112	502	549	556	365	153	307.04
2	Projection pursuit	1985	98	402	673	625	343	122	297.15
3	Accurate approximations for posterior moments and marginal densities	1986	115	562	614	322	84	8	272.07
4	Bootstrap confidence-intervals for a class of parametric problems	1985	40	403	590	537	304	100	250.16
5	A comparison of kriging with nonparametric regression methods	1985	14	172	549	553	538	357	229.71
6	Transformations in regression - a robust analysis	1985	17	136	497	669	358	150	195.30
7	Analysis of field experiments by least-squares smoothing	1985	47	297	509	383	205	39	194.97
8	Jackknife, bootstrap and other resampling methods in regression-analysis - discussion	1986	103	346	343	230	126	51	189.79
9	Assessment of local influence	1986	64	291	505	325	124	45	189.07
10	Edgeworth corrected pivotal statistics and the bootstrap	1985	50	296	451	363	188	67	188.29
11	Longitudinal data-analysis using generalized linear-models	1986	231	268	197	92	16	4	185.31
12	Double exponential-families and their use in generalized linear-regression	1986	40	280	528	383	135	36	184.70
13	Parameter orthogonality and approximate conditional inference	1987	154	394	263	66	13	0	183.45
14	Linear-models for the analysis of longitudinal-studies	1985	63	432	349	184	87	17	180.98
15	Estimating optimal transformations for multiple-regression and correlation	1985	93	347	305	182	107	50	175.36
16	Common structure of smoothing techniques in statistics	1985	21	229	496	426	165	41	169.19
17	Sampling-based approaches to calculating marginal densities	1990	217	283	92	15	3	0	161.77
18	Maximum-likelihood estimation in a class of nonregular cases	1985	32	272	450	287	92	39	158.63
19	Testing for independence in a 2-way table - new interpretations of the chi-square statistic	1985	17	104	385	512	352	119	158.56
20	A diagnostic for cox regression and general conditional likelihoods	1985	17	121	338	549	333	120	158.18

Table 3.6: Top 20 Articles from 1986 to 1996 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Accurate approximations for posterior moments and marginal densities	1986	131	759	856	500	184	33	376.98
2	Longitudinal data-analysis using generalized linear-models	1986	296	394	335	177	52	7	270.28
3	Jackknife, bootstrap and other resampling methods in regression-analysis - discussion	1986	115	445	533	377	207	81	260.79
4	Double exponential-families and their use in generalized linear-regression	1986	44	355	755	578	280	76	260.54
5	Assessment of local influence	1986	76	358	704	511	231	87	260.25
6	Parameter orthogonality and approximate conditional inference	1987	182	527	440	152	36	6	257.60
7	Sampling-based approaches to calculating marginal densities	1990	287	457	203	47	13	0	248.84
8	The calculation of posterior distributions by data augmentation	1987	124	463	435	162	47	4	220.10
9	Automatic smoothing of regression-functions in generalized linear-models	1986	51	305	509	386	227	84	203.61
10	Inference on full or partial parameters based on the standardized signed log likelihood ratio	1986	52	272	558	429	140	37	196.62
11	An extended quasi-likelihood function	1987	60	330	468	333	164	32	194.45
12	Better bootstrap confidence-intervals	1987	130	311	311	210	118	74	191.07
13	Predictive likelihood inference with applications	1986	14	238	588	446	151	34	178.77
14	Model robust confidence-intervals using maximum-likelihood estimators	1986	53	328	372	318	160	42	178.59
15	Robust empirical Bayes analyses of event rates	1987	22	384	530	224	50	12	177.07
16	Semiparametric estimates of the relation between weather and electricity sales	1986	48	282	454	315	135	52	175.37
17	Longitudinal data-analysis for discrete and continuous outcomes	1986	106	275	290	236	167	37	173.81
18	Spline smoothing in a partly linear-model	1986	37	211	441	400	206	84	169.45
19	Approximate predictive likelihood	1986	14	208	545	439	151	36	166.91
20	Likelihood and observed geometries	1986	17	215	532	420	151	31	166.08

Table 3.7: Top 20 Articles from 1987 to 1997 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Sampling-based approaches to calculating marginal densities	1990	356	685	379	117	22	6	363.50
2	Parameter orthogonality and approximate conditional inference	1987	207	662	707	291	77	18	347.52
3	The calculation of posterior distributions by data augmentation	1987	141	662	656	302	93	9	312.67
4	An extended quasi-likelihood function	1987	66	409	635	559	271	77	259.42
5	Better bootstrap confidence-intervals	1987	144	369	463	381	230	131	252.99
6	Robust empirical Bayes analyses of event rates	1987	24	494	776	390	119	22	247.42
7	Models for longitudinal data - a generalized estimating equation approach	1988	138	422	434	251	104	28	232.35
8	Progress with numerical and graphical methods for practical Bayesian statistics	1987	28	439	677	341	109	18	221.11
9	The calculation of posterior distributions by data augmentation - comment	1987	20	438	729	294	79	18	217.59
10	Illustration of Bayesian-inference in normal data models using Gibbs sampling	1990	99	477	420	188	52	14	215.02
11	Correlated binary regression with covariates specific to each binary observation	1988	121	323	423	270	110	24	204.01
12	A quasirandom approach to integration in Bayesian statistics	1988	8	381	678	364	108	21	201.80
13	Generalized linear-models with random effects - a Gibbs sampling approach	1991	106	429	344	137	35	5	193.31
14	On the amount of noise inherent in bandwidth selection for a kernel density estimator	1987	35	246	445	425	325	165	187.13
15	Biased and unbiased cross-validation in density-estimation	1987	54	271	391	339	257	149	182.25
16	Local likelihood estimation	1987	36	201	402	469	413	132	180.22
17	How far are automatically chosen regression smoothing parameters from their optimum	1988	74	301	327	273	178	139	178.98
18	Prepivoting to reduce level error of confidence sets	1987	77	282	316	294	186	102	174.80
19	Maximum-likelihood computations with repeated measures - application of the em-algorithm	1987	63	209	330	399	306	119	172.17
20	Time-series regression with a unit-root	1987	154	309	219	74	47	17	171.74

Table 3.8: Top 20 Articles from 1988 to 1998 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Sampling-based approaches to calculating marginal densities	1990	453	952	610	203	41	14	509.61
2	Models for longitudinal data - a generalized estimating equation approach	1988	158	534	664	395	190	53	313.54
3	Illustration of Bayesian-inference in normal data models using Gibbs sampling	1990	121	640	662	325	85	25	305.24
4	A quasirandom approach to integration in Bayesian statistics	1988	8	488	941	596	188	39	280.67
5	Correlated binary regression with covariates specific to each binary observation	1988	140	428	637	403	206	51	280.53
6	Generalized linear-models with random effects - a Gibbs sampling approach	1991	125	607	528	260	66	17	277.45
7	How far are automatically chosen regression smoothing parameters from their optimum	1988	82	369	461	422	273	234	235.57
8	Locally weighted regression - an approach to regression-analysis by local fitting	1988	70	370	568	395	272	119	234.19
9	Newton-raphson and EM algorithms for linear mixed-effects models for repeated-measures data	1988	45	295	703	541	276	64	230.83
10	Bayesian-inference in econometric-models using Monte-Carlo integration	1989	60	393	559	433	211	58	229.18
11	Linear smoothers and additive-models	1989	86	399	509	310	208	97	229.16
12	Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates	1988	41	294	531	612	370	177	226.04
13	Theoretical comparison of bootstrap confidence-intervals	1988	131	297	376	364	237	76	217.43
14	Conditional logistic-regression models for correlated binary data	1988	28	224	611	625	308	81	206.13
15	Markov regression-models for time-series - a quasi-likelihood approach	1988	25	201	545	656	385	184	204.89
16	Fully exponential laplace approximations to expectations and variances of nonpositive functions	1989	42	393	535	344	132	37	204.87
17	Bayesian computation via the Gibbs sampler and related Markov-chain Monte-Carlo methods	1993	180	388	234	45	14	4	200.58
18	Correlated binary regression using a quadratic exponential model	1990	90	280	443	379	180	58	199.07
19	Analyzing repeated measurements with possibly missing observations by modeling marginal distributions	1988	25	166	482	496	531	339	195.05
20	Local computations with probabilities on graphical structures and their application to expert systems	1988	54	190	499	641	246	56	193.81

Table 3.9: Top 20 Articles from 1989 to 1999 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Sampling-based approaches to calculating marginal densities	1990	526	1254	892	355	87	26	675.48
2	Illustration of Bayesian-inference in normal data models using Gibbs sampling	1990	132	816	967	502	147	45	403.05
3	Generalized linear-models with random effects - a Gibbs sampling approach	1991	143	799	771	406	126	37	374.16
4	Bayesian-inference in econometric-models using Monte-Carlo integration	1989	69	509	772	630	324	117	301.59
5	Linear smoothers and additive-models	1989	95	487	734	502	363	174	300.66
6	Bayesian computation via the Gibbs sampler and related Markov-chain Monte-Carlo methods	1993	215	588	372	119	32	11	291.70
7	Fully exponential laplace approximations to expectations and variances of nonpositive functions	1989	45	516	771	512	228	73	276.10
8	Correlated binary regression using a quadratic exponential model	1990	103	376	608	590	286	104	265.05
9	Multivariate adaptive regression splines	1991	119	491	419	320	184	65	251.25
10	Bayesian image-restoration, with 2 applications in spatial statistics	1991	66	460	685	365	104	16	245.50
11	Adaptive rejection sampling for Gibbs sampling	1992	112	405	620	286	84	17	242.41
12	Spatial statistics and Bayesian computation	1993	87	560	448	175	39	18	234.50
13	Bayesian-analysis of constrained parameter and truncated data problems using Gibbs sampling	1992	52	392	657	356	112	20	220.08
14	A Monte-Carlo method for Bayesian-inference in frailty models	1991	41	271	795	542	172	31	220.05
15	The robust inference for the cox proportional hazards model	1989	47	348	716	371	171	41	219.43
16	Bayesian statistics without tears - a sampling resampling perspective	1992	60	399	606	346	108	26	219.13
17	Approximate Bayesian-inference in conditionally independent hierarchical-models (parametric empirical Bayes models)	1989	59	304	566	514	252	83	214.24
18	Flexible parsimonious smoothing and additive modeling	1989	59	350	473	316	309	181	206.69
19	Non-parametric and semi-parametric maximum-likelihood estimators and the von mises method .1.	1989	106	334	378	263	154	53	199.16
20	On substantive research hypotheses, conditional-independence graphs and graphical chain models	1990	36	283	726	383	111	25	197.82

Table 3.10: Top 20 Articles from 1990 to 2000 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Sampling-based approaches to calculating marginal densities	1990	605	1602	1248	569	151	36	860.20
2	Illustration of Bayesian-inference in normal data models using Gibbs sampling	1990	150	1007	1339	748	259	75	510.80
3	Generalized linear-models with random effects - a Gibbs sampling approach	1991	164	1026	1086	599	235	74	484.70
4	Bayesian computation via the Gibbs sampler and related Markov-chain Monte-Carlo methods	1993	245	833	604	239	57	17	400.81
5	Correlated binary regression using a quadratic exponential model	1990	112	480	849	837	456	176	335.67
6	Bayesian image-restoration, with 2 applications in spatial statistics	1991	97	581	960	584	216	39	332.75
7	Multivariate adaptive regression splines	1991	141	635	598	485	326	140	330.30
8	Adaptive rejection sampling for Gibbs sampling	1992	138	517	873	496	161	33	322.92
9	Spatial statistics and Bayesian computation	1993	100	736	714	319	87	24	320.13
10	Bayesian statistics without tears - a sampling resampling perspective	1992	69	496	867	569	205	47	287.50
11	Bayesian-analysis of constrained parameter and truncated data problems using Gibbs sampling	1992	60	478	933	593	204	44	287.21
12	A Monte-Carlo method for Bayesian-inference in frailty models	1991	47	331	1083	831	302	59	282.89
13	Constrained Monte-Carlo maximum-likelihood for dependent data	1992	72	483	751	501	208	62	269.91
14	Multivariate regression-analyses for categorical-data	1992	150	545	381	258	99	25	264.43
15	Markov-chains for exploring posterior distributions	1994	245	464	190	62	14	0	259.76
16	On substantive research hypotheses, conditional-independence graphs and graphical chain models	1990	37	362	1007	612	207	55	258.67
17	Gibbs sampling for marginal posterior expectations	1991	17	325	961	625	287	83	239.62
18	On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions	1991	14	412	863	490	175	38	232.52
19	Structural image-restoration through deformable templates	1991	14	381	886	499	166	37	228.24
20	Explaining the Gibbs sampler	1992	87	448	518	241	88	24	222.51

Table 3.11: Top 20 Articles from 1991 to 2001 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Generalized linear-models with random effects - a Gibbs sampling approach	1991	180	1279	1452	865	345	121	607.13
2	Bayesian computation via the Gibbs sampler and related Markov-chain Monte-Carlo methods	1993	261	1098	893	381	109	29	509.13
3	Bayesian image-restoration, with 2 applications in spatial statistics	1991	119	716	1275	842	354	78	427.52
4	Multivariate adaptive regression splines	1991	157	766	833	689	492	235	412.93
5	Adaptive rejection sampling for Gibbs sampling	1992	158	640	1186	738	261	63	411.63
6	Spatial statistics and Bayesian computation	1993	112	909	1014	530	138	46	410.36
7	Bayesian statistics without tears - a sampling resampling perspective	1992	73	589	1174	829	336	89	362.46
8	A Monte-Carlo method for Bayesian-inference in frailty models	1991	52	384	1427	1163	496	100	359.44
9	Bayesian-analysis of constrained parameter and truncated data problems using Gibbs sampling	1992	62	539	1274	877	334	77	358.55
10	Markov-chains for exploring posterior distributions	1994	299	653	357	115	32	4	353.17
11	Constrained Monte-Carlo maximum-likelihood for dependent data	1992	84	618	1002	709	345	94	349.70
12	Multivariate regression-analyses for categorical-data	1992	170	700	594	388	180	53	343.40
13	On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions	1991	16	485	1165	730	288	76	299.32
14	Gibbs sampling for marginal posterior expectations	1991	17	351	1238	944	449	141	299.17
15	Explaining the Gibbs sampler	1992	96	578	755	404	146	58	293.35
16	Structural image-restoration through deformable templates	1991	16	437	1189	759	277	73	292.42
17	Approximate inference in generalized linear mixed models	1993	278	396	327	158	45	13	284.78
18	Comparing sweep strategies for stochastic relaxation	1991	24	376	1081	836	357	108	280.49
19	Modeling complexity - applications of Gibbs sampling in medicine	1993	69	505	873	480	144	42	279.19
20	Hierarchical Bayesian-analysis of change-point problems	1992	42	300	814	1064	674	247	275.36

Table 3.12: Top 20 Articles from 1992 to 2002 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Bayesian computation via the Gibbs sampler and related Markov-chain Monte-Carlo methods	1993	280	1342	1273	603	189	48	625.77
2	Adaptive rejection sampling for Gibbs sampling	1992	179	763	1484	1098	422	111	516.36
3	Spatial statistics and Bayesian computation	1993	126	1068	1352	849	233	77	513.90
4	Bayesian statistics without tears - a sampling resampling perspective	1992	79	695	1475	1182	533	151	460.79
5	Bayesian-analysis of constrained parameter and truncated data problems using Gibbs sampling	1992	66	615	1597	1248	531	138	453.59
6	Constrained Monte-Carlo maximum-likelihood for dependent data	1992	91	748	1332	945	548	162	448.61
7	Markov-chains for exploring posterior distributions	1994	338	884	584	209	53	11	448.22
8	Multivariate regression-analyses for categorical-data	1992	183	827	813	576	284	93	418.07
9	Explaining the Gibbs sampler	1992	103	708	1069	605	241	100	378.35
10	Hierarchical Bayesian-analysis of change-point problems	1992	45	377	1040	1342	957	371	366.23
11	Modeling complexity - applications of Gibbs sampling in medicine	1993	72	601	1184	726	249	69	358.52
12	Approximate inference in generalized linear mixed models	1993	317	533	487	274	86	24	356.94
13	Design-adaptive nonparametric regression	1992	145	585	560	309	144	33	291.74
14	Bayesian computation and stochastic-systems	1995	168	572	439	221	78	10	278.00
15	Approximating point process likelihoods with glim	1992	16	140	784	1322	947	544	276.98
16	Assessment and propagation of model uncertainty	1995	85	493	660	419	135	24	257.75
17	Bayes factors	1995	218	434	361	116	43	4	256.90
18	Asymptotic-behavior of the Gibbs sampler	1993	19	450	923	565	211	53	256.79
19	Local linear-regression smoothers and their minimax efficiencies	1993	134	460	469	293	121	42	247.23
20	Facilitating the Gibbs sampler - the Gibbs stopper and the griddy-Gibbs sampler	1992	53	330	717	631	381	171	243.25

Table 3.13: Top 20 Articles from 1993 to 2003 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Bayesian computation via the Gibbs sampler and related Markov-chain Monte-Carlo methods	1993	298	1626	1693	891	296	80	763.40
2	Spatial statistics and Bayesian computation	1993	130	1239	1749	1215	391	118	630.68
3	Markov-chains for exploring posterior distributions	1994	375	1123	847	389	102	24	555.49
4	Modeling complexity - applications of Gibbs sampling in medicine	1993	75	701	1521	1023	418	128	451.71
5	Approximate inference in generalized linear mixed models	1993	356	717	683	408	164	52	447.16
6	Bayesian computation and stochastic-systems	1995	185	745	642	350	161	27	356.69
7	Assessment and propagation of model uncertainty	1995	102	609	931	615	239	56	340.64
8	Bayes factors	1995	260	611	541	211	78	13	340.44
9	Asymptotic-behavior of the Gibbs sampler	1993	20	517	1166	828	383	101	329.38
10	Local linear-regression smoothers and their minimax efficiencies	1993	153	568	609	439	217	91	311.24
11	Bayesian-inference for generalized linear and proportional hazards models via Gibbs sampling	1993	73	345	696	934	733	342	306.08
12	Variable selection via Gibbs sampling	1993	93	531	778	570	251	57	300.38
13	Bayesian-analysis of binary and polychotomous response data	1993	132	397	737	600	359	108	297.31
14	Ideal spatial adaptation by wavelet shrinkage	1994	183	505	590	345	115	40	296.54
15	A language and program for complex Bayesian modeling	1994	50	378	911	868	432	178	296.28
16	Approximate Bayesian-inference with the weighted likelihood bootstrap	1994	57	476	809	688	385	166	294.30
17	On the irreducibility of a Markov-chain defined on a space of genotype configurations by a sampling scheme	1993	16	254	975	1094	618	241	287.92
18	Representations of knowledge in complex-systems	1994	52	476	880	568	257	68	279.64
19	Model selection and accounting for model uncertainty in graphical models using occams window	1994	62	555	761	510	171	35	278.53
20	Fractional Bayes factors for model comparison	1995	90	330	688	786	462	157	277.00

Table 3.14: Top 20 Articles from 1994 to 2004 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Markov-chains for exploring posterior distributions	1994	417	1367	1215	632	183	47	679.44
2	Bayesian computation and stochastic-systems	1995	192	917	926	552	271	73	447.89
3	Assessment and propagation of model uncertainty	1995	110	748	1222	931	415	113	446.53
4	Bayes factors	1995	304	780	786	364	138	50	433.44
5	A language and program for complex Bayesian modeling	1994	55	431	1160	1176	693	305	396.22
6	Approximate Bayesian-inference with the weighted likelihood bootstrap	1994	65	560	1054	972	604	297	394.55
7	Ideal spatial adaptation by wavelet shrinkage	1994	225	643	819	556	202	68	390.52
8	Fractional Bayes factors for model comparison	1995	100	389	926	1060	702	290	374.11
9	Model selection and accounting for model uncertainty in graphical models using occams window	1994	74	667	1019	782	316	79	370.35
10	Representations of knowledge in complex-systems	1994	56	562	1122	866	419	136	367.93
11	Bayesian model choice - asymptotics and exact calculations	1994	95	660	962	622	267	96	358.61
12	Bayesian-analysis of linear and nonlinear population-models by using the Gibbs sampler	1994	59	292	781	1113	963	546	355.65
13	Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes	1994	83	546	1024	697	270	92	341.61
14	Reversible jump Markov chain Monte Carlo computation and Bayesian model determination	1995	244	599	590	343	121	23	341.12
15	Regeneration in Markov-chain samplers	1995	31	426	1081	891	471	196	332.32
16	Wavelet shrinkage - asymptopia	1995	179	565	665	420	228	43	325.21
17	Generalized linear-models with unknown link functions	1994	20	192	875	1184	853	376	313.89
18	On the convergence of Monte-Carlo maximum-likelihood calculations	1994	28	287	1028	977	548	236	310.16
19	Estimation of finite mixture distributions through Bayesian sampling	1994	84	422	800	746	414	156	308.69
20	Robust priors for smoothing and image-restoration	1994	8	218	951	1073	630	300	294.36

Table 3.15: Top 20 Articles from 1995 to 2005 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Assessment and propagation of model uncertainty	1995	123	902	1628	1317	623	216	602.09
2	Bayesian computation and stochastic-systems	1995	203	1120	1286	806	429	153	577.45
3	Bayes factors	1995	368	1035	1079	595	244	89	571.11
4	Fractional Bayes factors for model comparison	1995	112	485	1211	1399	1043	438	523.68
5	Regeneration in Markov-chain samplers	1995	34	473	1357	1247	712	331	455.93
6	Reversible jump Markov chain Monte Carlo computation and Bayesian model determination	1995	297	781	864	545	212	47	453.98
7	Wavelet shrinkage - asymptopia	1995	192	683	917	683	386	101	424.28
8	Bayesian model choice via Markov-chain Monte-Carlo methods	1995	90	534	1034	777	407	100	372.67
9	Data-driven bandwidth selection in local polynomial fitting - variable bandwidth and spatial adaptation	1995	125	551	856	667	442	172	370.08
10	Inference from a deterministic population-dynamics model for bowhead whales	1995	20	432	1024	1019	582	244	366.41
11	Understanding the Metropolis-Hastings algorithm	1995	158	569	839	524	238	88	351.05
12	A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion	1995	76	518	835	740	334	97	331.17
13	Covariance structure and convergence rate of the Gibbs sampler with various scans	1995	33	333	861	939	634	237	330.22
14	A Bayesian method for combining results from several binomial experiments	1995	14	325	814	852	527	207	296.50
15	Markov chain Monte Carlo convergence diagnostics: a comparative review	1996	152	511	570	429	161	48	288.33
16	Efficient parametrizations for normal linear mixed models	1995	56	448	743	568	284	112	280.24
17	Minorization conditions and convergence-rates for Markov-chain Monte-Carlo	1995	67	330	566	590	503	265	272.53
18	Annealing Markov-chain Monte-Carlo with applications to ancestral inference	1995	60	390	626	549	355	150	263.94
19	Bayesian density-estimation and inference using mixtures	1995	107	348	587	488	284	102	256.95
20	Gibbs sampler convergence criteria	1995	12	346	749	674	351	111	255.45

Table 3.16: Top 20 Articles from 1996 to 2006 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Markov chain Monte Carlo convergence diagnostics: A comparative review	1996	163	672	806	667	303	92	393.50
2	The intrinsic Bayes factor for model selection and prediction	1996	131	432	868	763	333	86	355.08
3	On Bayesian analysis of mixtures with an unknown number of components	1997	191	623	582	285	129	25	305.94
4	Hierarchical generalized linear models	1996	148	531	573	417	149	42	287.62
5	Approximate Bayes factors and accounting for model uncertainty in generalised linear models	1996	61	323	700	667	303	55	271.31
6	The EM algorithm - An old folk-song sung to a fast new tune	1997	83	402	559	544	215	47	256.02
7	The effect of improper priors on Gibbs sampling in hierarchical linear mixed models	1996	77	366	623	442	206	51	243.80
8	Asymptotic equivalence of nonparametric regression and white noise	1996	87	337	499	467	273	86	240.14
9	Posterior predictive assessment of model fitness via realized discrepancies	1996	133	353	470	429	164	51	239.42
10	Using bootstrap likelihood ratios in finite mixture models	1996	33	285	673	558	264	87	238.33
11	Maximum likelihood algorithms for generalized linear mixed models	1997	139	413	503	277	69	14	228.57
12	Stochastic versions of the EM algorithm: An experimental study in the mixture case	1996	12	250	632	558	270	115	222.02
13	The selection of prior distributions by formal rules	1996	96	337	603	343	107	22	220.59
14	Bayesian curve fitting using multivariate normal mixtures	1996	31	214	422	482	372	223	210.97
15	Flexible smoothing with B-splines and penalties	1996	142	257	474	292	143	57	210.03
16	Bayesian model averaging for linear regression models	1997	78	342	575	304	92	18	205.34
17	Asymptotic equivalence of density estimation and Gaussian white noise	1996	63	244	429	386	264	98	197.00
18	Rates of convergence of the Hastings and Metropolis algorithms	1996	60	216	414	418	268	102	192.75
19	Simulating ratios of normalizing constants via a simple identity: A theoretical exploration	1996	84	261	420	373	191	37	192.60
20	Computing Bayes factors by combining simulation and asymptotic approximations	1997	68	300	393	392	161	60	191.70

Table 3.17: Top 20 Articles from 1997 to 2007 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	On Bayesian analysis of mixtures with an unknown number of components	1997	230	862	883	521	239	72	441.66
2	The EM algorithm - An old folk-song sung to a fast new tune	1997	90	523	842	838	396	144	364.24
3	Maximum likelihood algorithms for generalized linear mixed models	1997	162	529	735	509	192	40	323.55
4	Bayesian model averaging for linear regression models	1997	85	423	879	580	255	50	301.51
5	Generalized partially linear single-index models	1997	90	343	524	671	420	177	282.08
6	Computing Bayes factors by combining simulation and asymptotic approximations	1997	72	380	608	675	324	129	278.67
7	1994 Wald memorial lectures - Polynomial splines and their tensor products in extended linear modeling	1997	78	456	799	402	176	73	273.98
8	Hierarchical spatio-temporal mapping of disease rates	1997	98	379	599	465	235	83	256.18
9	Practical Bayesian density estimation using mixtures of normals	1997	75	400	601	435	286	110	255.45
10	Likelihood analysis of non-Gaussian measurement time series	1997	77	357	594	558	263	70	252.70
11	Automatic Bayesian curve fitting	1998	79	339	563	512	268	114	246.69
12	Polychotomous regression	1997	31	179	555	820	401	157	239.89
13	Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler	1997	56	308	593	525	267	88	233.44
14	Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors	1997	35	298	516	536	265	125	218.11
15	Monte Carlo maximum likelihood estimation for non-Gaussian state space models	1997	46	172	436	525	471	234	217.10
16	Wavelet threshold estimators for data with correlated noise	1997	63	279	485	463	248	59	207.90
17	Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models	1997	43	350	536	329	179	57	200.36
18	Disease mapping with errors in covariates	1997	29	186	336	551	429	232	199.44
19	Minimax estimation via wavelet shrinkage	1998	87	304	442	337	161	49	196.94
20	Local polynomial variance-function estimation	1997	47	234	482	394	276	125	194.58

Table 3.18: Top 20 Articles from 1998 to 2008 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Automatic Bayesian curve fitting	1998	86	460	829	838	501	243	360.20
2	Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data	1998	114	327	724	621	306	111	287.02
3	On measuring and correcting the effects of data mining and model selection	1998	54	646	692	426	159	26	283.27
4	Minimax estimation via wavelet shrinkage	1998	90	376	657	611	338	135	282.48
5	Model choice: A minimum posterior predictive loss approach	1998	91	548	560	501	238	58	280.75
6	Analysis of multivariate probit models	1998	92	557	576	398	198	65	272.90
7	Smoothing spline models for the analysis of nested and crossed samples of curves	1998	105	477	615	487	192	45	272.90
8	On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data	2001	133	555	607	246	101	18	268.05
9	General methods for monitoring convergence of iterative simulations	1998	118	288	427	504	417	228	255.90
10	Bayesian measures of model complexity and fit	2002	330	322	224	106	39	15	244.85
11	Empirical Bayes analysis of a microarray experiment	2001	220	444	323	159	39	3	236.63
12	Analysis of variance for gene expression microarray data	2000	128	493	473	218	83	17	234.17
13	Wavelet thresholding via a Bayesian approach	1998	71	222	510	582	425	151	233.31
14	Arcing classifiers	1998	54	381	594	458	210	46	228.77
15	Optimal scaling of discrete approximations to Langevin diffusions	1998	26	218	574	558	375	302	227.06
16	Multiple shrinkage and subset selection in wavelets	1998	54	230	474	589	414	149	222.94
17	Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm	1999	112	356	510	306	145	23	218.20
18	Direct generalized additive modeling with penalized likelihood	1998	47	226	377	615	423	227	217.65
19	Risk bounds for model selection via penalization	1999	96	299	517	430	180	34	215.77
20	Poisson/gamma random field models for spatial statistics	1998	50	254	496	482	367	137	214.02

Table 3.19: Top 20 Articles from 1999 to 2009 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Bayesian measures of model complexity and fit	2002	440	490	434	232	97	30	372.73
2	On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data	2001	145	717	887	469	193	74	367.75
3	Empirical Bayes analysis of a microarray experiment	2001	261	616	515	321	125	13	335.01
4	Risk bounds for model selection via penalization	1999	114	421	785	754	401	99	325.86
5	Analysis of variance for gene expression microarray data	2000	136	608	736	405	190	55	317.93
6	Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm	1999	120	460	750	545	300	67	304.35
7	The control of the false discovery rate in multiple testing under dependency	2001	203	528	535	313	98	14	291.16
8	Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments	2002	142	621	526	345	115	28	286.35
9	Adjusting for nonignorable drop-out using semiparametric nonresponse models	1999	145	478	563	457	229	66	284.88
10	Reference Bayesian methods for generalized linear mixed models	2000	47	628	615	488	233	83	280.83
11	Clustering gene expression patterns	1999	51	543	743	470	245	92	280.44
12	Bayesian analysis of mixture models with an unknown number of components - An alternative to reversible jump methods	2000	71	275	699	681	492	182	275.46
13	Bayesian detection of clusters and discontinuities in disease maps	2000	49	273	805	650	472	193	273.92
14	A direct approach to false discovery rates	2002	228	539	372	183	48	4	270.46
15	Comparison of discrimination methods for the classification of tumors using gene expression data	2002	202	501	492	241	53	5	270.22
16	Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables	1999	59	325	634	643	402	183	262.45
17	Replicated microarray data	2002	96	422	659	368	215	80	255.27
18	Inference in generalized additive mixed models by using smoothing splines	1999	86	384	598	537	243	103	254.52
19	Bayesian model averaging: A tutorial	1999	137	377	621	413	118	34	252.97
20	When log-normal and gamma models give different results: A case study	1999	10	212	739	862	448	189	252.47

Table 3.20: Top 20 Articles from 2000 to 2010 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Bayesian measures of model complexity and fit	2002	535	691	690	428	214	74	517.48
2	On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data	2001	154	856	1205	761	356	157	472.55
3	Empirical Bayes analysis of a microarray experiment	2001	292	768	742	531	266	68	436.97
4	Analysis of variance for gene expression microarray data	2000	143	714	991	700	346	112	406.76
5	The control of the false discovery rate in multiple testing under dependency	2001	241	664	731	543	225	55	388.33
6	Reference Bayesian methods for generalized linear mixed models	2000	57	772	875	790	425	198	384.70
7	Bayesian analysis of mixture models with an unknown number of components - An alternative to reversible jump methods	2000	76	351	956	968	825	376	375.03
8	Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments	2002	148	756	752	553	251	72	369.60
9	Bayesian detection of clusters and discontinuities in disease maps	2000	52	333	1044	978	774	363	366.08
10	Comparison of discrimination methods for the classification of tumors using gene expression data	2002	228	649	776	431	120	15	364.77
11	A direct approach to false discovery rates	2002	266	691	573	354	136	18	363.74
12	Markov chain Monte Carlo methods for computing Bayes factors: A comparative review	2001	48	656	821	725	385	162	340.77
13	Additive logistic regression: A statistical view of boosting	2000	103	572	816	599	321	123	331.58
14	Replicated microarray data	2002	104	515	862	588	370	165	331.54
15	Support vector machine classification and validation of cancer tissue samples using microarray expression data	2000	100	449	947	715	309	83	325.53
16	Missing value estimation methods for DNA microarrays	2001	106	478	750	694	384	105	317.83
17	Marginal likelihood from the Metropolis-Hastings output	2001	88	249	828	819	618	340	316.65
18	Counting degrees of freedom in hierarchical and other richly-parameterised models	2001	29	613	696	648	395	211	307.81
19	Computational and inferential difficulties with mixture posterior distributions.	2000	91	521	704	593	327	132	304.13
20	Variable selection via nonconcave penalized likelihood and its oracle properties	2001	266	532	422	214	70	28	300.28

Table 3.21: Top 20 Articles from 2001 to 2011 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Bayesian measures of model complexity and fit	2002	665	1004	1060	741	424	169	737.94
2	On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data	2001	160	984	1581	1204	640	308	594.76
3	Empirical Bayes analysis of a microarray experiment	2001	321	941	1027	829	494	180	562.47
4	The control of the false discovery rate in multiple testing under dependency	2001	274	842	1012	849	411	141	511.40
5	Comparison of discrimination methods for the classification of tumors using gene expression data	2002	251	843	1134	727	271	48	490.50
6	A direct approach to false discovery rates	2002	307	859	839	600	267	73	481.87
7	Markov chain Monte Carlo methods for computing Bayes factors: A comparative review	2001	58	826	1192	1126	712	341	470.78
8	Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments	2002	153	888	1037	869	466	166	469.90
9	Variable selection via nonconcave penalized likelihood and its oracle properties	2001	361	775	687	450	184	60	458.44
10	Counting degrees of freedom in hierarchical and other richly-parameterised models	2001	33	773	1036	994	693	408	424.15
11	Replicated microarray data	2002	111	604	1128	897	609	303	419.65
12	Marginal likelihood from the Metropolis-Hastings output	2001	100	329	1086	1242	965	622	418.95
13	Missing value estimation methods for DNA microarrays	2001	120	572	1027	1025	636	256	414.70
14	Least angle regression	2004	369	740	513	228	74	17	411.28
15	Regularization of wavelet approximations	2001	90	768	1020	814	436	144	405.21
16	Bayesian varying-coefficient models using adaptive regression splines	2001	5	676	1012	1055	734	417	395.26
17	Bayesian analyses of longitudinal binary data using Markov regression models of unknown order	2001	6	672	1002	1060	734	424	394.64
18	On the relationship between Markov chain Monte Carlo methods for model uncertainty	2001	25	210	1076	1384	1310	683	391.12
19	Nonparametric mixed effects models for unequally sampled noisy curves	2001	103	519	947	1007	665	258	387.40
20	Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics	2001	177	491	801	748	366	176	360.93

Table 3.22: Top 20 Articles from 2002 to 2012 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Bayesian measures of model complexity and fit	2002	767	1376	1526	1131	719	325	968.19
2	Comparison of discrimination methods for the classification of tumors using gene expression data	2002	275	1016	1605	1101	497	141	629.57
3	A direct approach to false discovery rates	2002	340	1054	1223	974	509	161	622.75
4	Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments	2002	161	994	1394	1339	777	329	589.58
5	Least angle regression	2004	463	1081	800	431	166	42	580.48
6	Replicated microarray data	2002	119	689	1449	1312	988	522	533.73
7	Adaptive model selection	2002	47	779	1081	874	534	257	404.31
8	Multiple hypothesis testing in microarray experiments	2003	164	650	867	733	422	229	396.17
9	Operating characteristics and extensions of the false discovery rate procedure	2002	128	584	888	972	592	204	392.00
10	Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach	2004	169	490	762	742	457	196	355.66
11	Model-based clustering, discriminant analysis, and density estimation	2002	226	562	601	523	287	91	350.63
12	Regularization and variable selection via the elastic net	2005	306	613	435	270	82	12	349.44
13	Exploration, normalization, and summaries of high density oligonucleotide array probe level data	2003	188	536	784	561	232	137	347.62
14	Comparison of methods for image analysis on cDNA microarray data	2002	31	299	827	1183	969	527	334.34
15	The positive false discovery rate: A Bayesian interpretation and the q-value	2003	145	498	656	693	460	179	329.80
16	Selecting the number of knots for penalized splines	2002	142	517	820	542	258	89	321.78
17	A comparison of normalization methods for high density oligonucleotide array data based on variance and bias	2003	157	461	681	620	350	193	320.18
18	Varying-coefficient models and basis function approximations for the analysis of repeated measurements	2002	91	472	716	693	463	178	303.77
19	The adaptive lasso and its oracle properties	2006	359	465	213	53	9	2	302.41
20	Tumor classification by partial least squares using microarray gene expression data	2002	81	312	837	814	519	230	293.59

Table 3.23: Top 20 Articles from 2003 to 2013 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Least angle regression	2004	555	1490	1156	713	302	97	783.29
2	Regularization and variable selection via the elastic net	2005	387	875	707	452	179	50	496.85
3	Multiple hypothesis testing in microarray experiments	2003	170	762	1110	1096	696	394	485.52
4	The adaptive lasso and its oracle properties	2006	467	750	401	124	27	4	450.84
5	Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach	2004	180	573	1017	1095	738	311	440.23
6	Exploration, normalization, and summaries of high density oligonucleotide array probe level data	2003	198	663	1063	858	404	243	437.18
7	The positive false discovery rate: A Bayesian interpretation and the q-value	2003	163	558	889	1018	750	329	411.75
8	Nonconcave penalized likelihood with a diverging number of parameters	2004	157	928	902	503	166	73	410.38
9	A comparison of normalization methods for high density oligonucleotide array data based on variance and bias	2003	165	536	905	924	594	336	396.20
10	High-dimensional graphs and variable selection with the Lasso	2006	271	775	579	239	59	8	380.21
11	Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions	2003	53	315	1101	1634	999	318	377.08
12	Efficient estimation of covariance selection models	2003	49	488	1129	1054	563	247	350.32
13	Persistence in high-dimensional linear predictor selection and the virtue of overparametrization	2004	68	808	927	508	187	50	341.57
14	New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis	2004	114	547	881	774	417	136	340.43
15	Prior distributions for variance parameters in hierarchical models(Comment on an Article by Browne and Draper)	2006	259	614	536	257	97	27	340.35
16	Model selection and estimation in regression with grouped variables	2006	250	638	536	246	62	21	337.77
17	Large-scale simultaneous hypothesis testing: The choice of a null hypothesis	2004	163	596	658	464	299	245	333.15
18	Classification of gene microarrays by penalized logistic regression	2004	36	605	951	781	507	218	327.67
19	Variable selection using MM algorithms	2005	81	815	793	401	142	39	326.50
20	Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences	2004	69	350	908	1043	730	294	320.11

Table 3.24: Top 20 Articles from 2004 to 2014 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Least angle regression	2004	630	1892	1564	1025	488	184	986.66
2	Regularization and variable selection via the elastic net	2005	455	1138	1001	683	332	109	646.99
3	The adaptive lasso and its oracle properties	2006	572	1018	646	250	72	8	613.15
4	Nonconcave penalized likelihood with a diverging number of parameters	2004	189	1191	1234	752	320	138	537.29
5	High-dimensional graphs and variable selection with the Lasso	2006	312	1020	867	427	116	25	502.89
6	Model selection and estimation in regression with grouped variables	2006	308	850	780	393	130	43	455.72
7	Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach	2004	194	597	1068	1257	790	314	453.55
8	Prior distributions for variance parameters in hierarchical models(Comment on an Article by Browne and Draper)	2006	305	790	827	427	174	57	450.28
9	Persistence in high-dimensional linear predictor selection and the virtue of overparametrization	2004	74	1003	1248	802	332	101	437.04
10	Variable selection using MM algorithms	2005	96	1045	1093	645	279	77	431.00
11	New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis	2004	126	706	1161	1092	648	262	430.53
12	Large-scale simultaneous hypothesis testing: The choice of a null hypothesis	2004	172	714	909	698	435	351	405.35
13	Classification of gene microarrays by penalized logistic regression	2004	39	726	1238	1076	768	376	402.10
14	Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations	2004	83	708	1223	866	455	186	389.42
15	Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences	2004	76	410	1147	1452	1071	484	385.22
16	The Dantzig selector: Statistical estimation when p is much larger than n	2007	260	812	514	189	63	18	379.76
17	Detecting differential gene expression with a semiparametric hierarchical mixture method	2004	135	482	860	894	643	341	350.52
18	The golden chain - Discussion	2004	10	553	1289	965	591	295	337.90
19	Bayesian P-splines	2004	130	588	883	628	295	125	334.76
20	Finding predictive gene groups from microarray data	2004	18	511	1207	1043	644	291	331.99

Table 3.25: Top 20 Articles from 2005 to 2015 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	Regularization and variable selection via the elastic net	2005	552	1504	1380	995	554	198	841.82
2	The adaptive lasso and its oracle properties	2006	691	1389	973	480	145	33	797.02
3	High-dimensional graphs and variable selection with the Lasso	2006	361	1352	1255	697	241	61	655.85
4	Model selection and estimation in regression with grouped variables	2006	385	1141	1111	638	248	86	608.87
5	Prior distributions for variance parameters in hierarchical models(Comment on an Article by Browne and Draper)	2006	347	1012	1192	681	334	123	577.48
6	Variable selection using MM algorithms	2005	109	1303	1536	1013	478	158	569.96
7	The Dantzig selector: Statistical estimation when p is much larger than n	2007	294	1144	797	375	132	44	506.28
8	Sparsity and smoothness via the fused lasso	2005	173	708	1360	803	406	126	446.78
9	PATHWISE COORDINATE OPTIMIZATION	2007	182	935	877	510	183	48	420.28
10	L-1-regularization path algorithm for generalized linear models	2007	118	789	1026	701	272	74	386.32
11	Empirical Bayes selection of wavelet thresholds	2005	85	561	1230	1121	549	203	384.41
12	Profile likelihood inferences on semiparametric varying-coefficient partially linear models	2005	183	552	840	710	484	219	370.89
13	The group lasso for logistic regression	2008	118	898	860	434	107	20	366.83
14	The sparsity and bias of the lasso selection in high-dimensional linear regression	2008	153	846	709	416	137	30	361.00
15	Regularized estimation of large covariance matrices	2008	147	665	937	581	186	33	354.74
16	Sure independence screening for ultrahigh dimensional feature space	2008	229	745	596	242	41	9	354.65
17	Adapting to unknown sparsity by controlling the false discovery rate	2006	83	666	1048	690	357	122	349.24
18	One-step sparse estimates in nonconcave penalized likelihood models	2008	223	722	568	235	57	10	344.22
19	Covariance matrix selection and estimation via penalised normal likelihood	2006	87	599	1014	817	409	113	343.70
20	Piecewise linear regularized solution paths	2007	64	716	966	686	329	93	338.42

Table 3.26: Top 20 Articles from 2006 to 2016 network. 1 to 6 indicate the number of citations with distance 1 to 6.

	Article Name	Year	1	2	3	4	5	6	ANI
1	The adaptive lasso and its oracle properties	2006	834	1893	1419	772	266	83	1060.76
2	High-dimensional graphs and variable selection with the Lasso	2006	417	1789	1774	1066	401	129	858.83
3	Model selection and estimation in regression with grouped variables	2006	469	1545	1559	969	395	147	811.74
4	Prior distributions for variance parameters in hierarchical models(Comment on an Article by Browne and Draper)	2006	416	1255	1702	1045	563	223	747.56
5	The Dantzig selector: Statistical estimation when p is much larger than n	2007	343	1595	1230	589	255	100	691.95
6	PATHWISE COORDINATE OPTIMIZATION	2007	221	1267	1295	768	342	101	573.04
7	L-1-regularization path algorithm for generalized linear models	2007	135	1009	1454	1093	478	157	509.89
8	Sure independence screening for ultrahigh dimensional feature space	2008	301	1059	913	451	118	23	509.87
9	The group lasso for logistic regression	2008	142	1193	1300	719	225	58	504.48
10	The sparsity and bias of the lasso selection in high-dimensional linear regression	2008	179	1158	1105	668	284	70	499.51
11	Regularized estimation of large covariance matrices	2008	186	913	1330	889	352	91	488.66
12	One-step sparse estimates in nonconcave penalized likelihood models	2008	259	1053	887	426	132	32	481.82
13	Piecewise linear regularized solution paths	2007	70	935	1386	1065	515	207	452.66
14	Adapting to unknown sparsity by controlling the false discovery rate	2006	87	822	1514	1050	568	226	452.62
15	Covariance matrix selection and estimation via penalised normal likelihood	2006	97	748	1441	1285	647	213	451.07
16	Relaxed lasso	2007	78	924	1420	904	346	80	438.47
17	Sparse inverse covariance estimation with the graphical lasso	2008	265	751	872	549	220	86	432.23
18	On the degrees of freedom of the lasso	2007	158	961	1042	506	211	55	428.11
19	Strictly proper scoring rules, prediction, and estimation	2007	230	715	1042	638	267	54	426.48
20	Tuning parameter selectors for the smoothly clipped absolute deviation method	2007	218	835	920	463	130	39	418.04

3.3 Extension of Article Network Influence and Comparison to Some Existing Measures of Research Metrics

We have demonstrated how ANI can be used to quantitatively measure the influence of a research article in its scientific field. In Chapter 1, we introduced some common measures of research metrics. In this section, we extend the use of network influence to other levels of citations and discuss the differences between ANI and three common measures of research metrics found in the literature.

3.3.1 A Comparison to Impact Factor in Journal-level Citation Network

Garfield (1955) introduced a pioneered concept to quantitatively measure the quality of a scientific journal using a single index about 60 years ago, and it is now well-known to be the Impact Factor among researchers in scientific fields. The formal definition of an impact factor is

$$IF_y = \frac{Citations_{y-1} + Citations_{y-2}}{Publications_{y-1} + Publications_{y-2}},$$

where y is the year of measurement, $Citations_t$ and $Publications_t$ are the numbers of total citations and the number of published papers of a journal respectively for $t = y - 1$ and $y - 2$.

Following the spirit of the impact factor, we extend the use of ANI to the journal-based level via considering the location of ANI of the articles published in the journal, namely Journal Network Influence (JNI). Since the distributions of ANI in every journal are mostly skewed, the location of distribution we consider is the median value instead of the mean value. The following table lists the top ten journals with the largest median network influence, and we compare them with the values of their impact factors. The network influence is calculated during 2006-2016, and the impact factor is taken in the year 2016.

Although the top-10 order follows quite similarly, it is obvious that some ranks are out of the usual expectation. Here are two observations: (1) Journal of Biostatistics ranked the third and had a higher influence than two of the traditional top-4 journals. It represents that 2006-2016 is a period that this subfield develops quickly and expands its influence towards the whole statistics community over the traditional statistics subfield. (2) Except for Journal of Biostatistics, Biometrika ranked back to top-4 statistics journals with JRSSB, Annals of Statistics, and JASA. This prestigious journal has its impact factor dropped to the sixth rank due to the rise of biostatistics

Table 3.27: The Median Network Influence of Top 10 Statistics Journals.

Journal	Influence	Impact Factor
JRSS SERIES B	29.4321	4.610
ANNALS OF STATISTICS	22.7762	3.023
BIOSTATISTICS	17.4442	1.798
JASA	16.0050	2.016
BIOMETRIKA	14.5430	1.448
JCGS	13.9627	1.735
BERNOULLI	12.2067	1.070
ECONOMETRIC REVIEWS	11.6636	1.333
BIOMETRICS	8.7595	1.329
STATISTICA SINICA	8.4241	0.899

and data visualization. However, if we examine its influence over the whole statistics community, it still remains as top 5. Similar situations appear in other two prestigious journals featuring theoretical statistics, Bernoulli and Statistica Sinica.

The main cause of these discrepancies comes from the conceptual difference between the network influence and the impact factor summarized as follows.

First, as mentioned in previous sections, an impact factor considers only an entity's direct citations and ignores its indirect citations. It is equivalent to set $r_M = 1$ or define weighting functions $g(1) = 1$ and $g(r) = 0$ for all $1 < r < r_M$ in the definition of ANI_t. Such setting completely ignores the indirect citations, so the influence of a potentially highly-influential article may be underestimated if one of its followers wrote another excellent article, and it received many citations after it published.

Second, the definition of the impact factor reveals that it considers a subset of the citation database that dates back for only two years. It is a major source of unfair comparisons between articles or journals in different subjects because different subjects may have different citation habitats for their distinct natures. For example, the journals in biological sciences and the computer science enjoy high impact factors as their advancements are straightforward and time applicable, while those in mathematical sciences always have relatively low impact factors as the contents require a deep understanding on the background and adequate time for full digestion. Instead of imposing a year limit, network influence considers the complete citation database, and it is ready to investigate the long-time influence accumulations across years.

3.3.2 A Comparison to PageRank Algorithm in Article-level Citation Network

Page et al. (1998) introduced a state-of-the-art algorithm called Pagerank to the rank website in Google search engine. It provides a measure of the importance of a web page by counting the number and quality of links to that web page under the assumption of preferential attachments. As a simple definition, the pagerank of a node, denoted as $R(v)$, is defined implicitly as the solution of the following equation,

$$R(v) = c \sum_{u \in M(v)} \frac{Ru}{N_u} + \frac{1-c}{n},$$

where c is a damping factor between 0 and 1, $M(v)$ is a set of nodes that link to v, N_u is the out-degree of node u, and n is the number of nodes in the graph.

If we consider a node as an article, it is possible to implement Pagerank to measure the importance of an article. In fact, Eigenfactor is a journal ranking method similar to Pagerank. Since the calculation of pagerank of a node includes the consideration of the quality of its neighboring nodes, it implicitly aggregates information from all nodes in the whole network rather than considering direct citations only. However, there are still several differences between Pagerank and network influence, as stated below.

First, Pagerank is an algorithm, and it requires multi-step propagations towards convergence. The Pagerank values of nodes are easily altered, and the propagation process is required to run again when a new node is inserted to the network. It may take an unexpectedly large amount of computational resources to obtain the converged values for all nodes if the node addition is very frequent, and it is the reality of the citation network in the Web of Science. Therefore, we suggest using the network influence as it can be obtained via a mathematical function that depends only on the number of citations of a target node and its neighboring nodes of interest. When a new node is added, it is not necessary to recalculate the values of all nodes and only a subset of nodes requires recalculations. Besides, it is not necessary to update the weight function whenever a new node is added because this function aims at providing a decreasing weight towards the pathlengths between two nodes. We suggest updating annually whenever an annual report on the network influence of the Web of Science is published.

Second, in the iterative formula of Pagerank, there is a damping factor (c) that is a user-defined parameter. Its complement ($1 - c$) is technically an adjustment or a weight on the probability of average crediting to all nodes in the network. Conventional wisdom suggests $c = 0.85$ as it provides satisfactory results, but this suggestion is arbitrary and without statistical justification. Even worse, a different setting of c

results in different Pagerank. To the extreme, if we set $c = 0$, all nodes will enjoy equivalent importance. It has been questionable on how to adjust this parameter properly. ANI has a user-defined parameter r_M, but it has a clear and interpretable meaning for users. Its weighting function is composed of terms with statistical meaning.

3.3.3 A Comparison to Field-Weighted Citation Impact (FWCI) in Article-level Citation Network

Field-Weighted Citation Impact (FWCI) is a new measure introduced by Elsevier (2018) to evaluate the entity's impacts in Scopus, the database of Elsevier since 1996. It indicates how the number of citations received by an entity's publications compared to the average number of citations received by all other similar publications. Mathematically, the FWCI of an entity is defined as:

$$FWCI = \frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{e_i},$$

where N is the number of publications by an entity, c_i is the citations received by publication i, and e_i is the expected number of citations received by all similar publications in the publication year plus the following three years. When a similar publication is allocated to more than one discipline, the harmonic mean is used to calculate e_i.

FWCI can be viewed as a simple modification from the impact factor. It differs in two places: (1) the inclusion of the number of citations changes from two backward years backward to three forward years; (2) instead of the number of publications as the denominator, FWCI considers the average number of citations by similar entities. Thus, two differences between ANI and impact factor also inherit in the difference to FWCI, and they are indirect citations and subsets in terms of years.

Moreover, when one considers an aggregated ANI from several articles, FWCI takes the arithmetic means on the ratio of actual to the expected citations of each publication while JNI defined in section 3.3.1 takes the median values of ANI of each publication. Notice that it is highly unlikely for both the ratio and the influence of an article in a citation network to be normally distributed, and as we observe in most cases, they are highly right-skewed. Therefore, FWCI tends to provide an inflated value on the average impact of an entity, and it is highly sensitive to the outlying values. In contrast, we propose to consider the median value of the ANIs of several articles instead. The median value is well-known to be resistant to these outliers, and it provides an informative average value towards a non-normal set of values.

Chapter 4

Statistical Model of citation network in Statistics & Probability

In this chapter, we analyze the structure of the article citation network of a particular subject and propose a generative model on how this citation network is evolved. In the WoS database, each article is assigned to a subject attribute that describes the field of the article. It is natural that different subjects have different characteristics in their citation network. For some theoretical-based subjects, it takes a certain amount of time for readers to digest the theorems in the articles, so their resulting citations will take some years to appear. On the other hand, for some application-based subjects, the results from their articles should be implemented in the shortest time in order to have the best advances, so these articles will receive immediate references after they are published. In order to understand the characteristics of citation networks of different subjects, it is essential to understand their structures and their underlying generative models. Then we might be able to find out the difference between different subjects.

4.1 Network Structure: Model and Characteristics

From the previous chapter, we know that citation network can be expressed as a directed graph. There have been many statistical models to construct graphs. The oldest and the most famous model is Erdos-Renyi random graph model (Erdös and Rényi (1959), Gilbert (1959)), in which each edge has the same connection probability. However, most of the real-world networks do not follow this model because it lacks an ability to express two important characteristics of many real-world networks: growth and preferential attachment. The first characteristic implies that the real-world network generally evolves over time, and the second characteristic describes the tendency

of a node to connect to other nodes with high number of connections. The article citation networks in the Web of Science also have these two characteristics.

Barabasi (Barabási and Albert (1999)) found that the degree distribution of many networks follow the power law distribution, which mimics the "rich-gets-richer" phenomenon. Barabasi also introduced two models to construct the network that can capture two important characteristics of real-world networks. The first model is Barabasi-Albert model (Barabási and Albert (1999)), which suggests that when a new node with m links is added to a network, the connection probabilities of the existing nodes in that network is decided by the number of nodes connected to these nodes. In this model, the older nodes usually obtain more links than the newer nodes. Barabasi also introduced another model named Bianconi-Barabasi model (Bianconi and Barabási (2001), Barabási et al. (2000)), in which a new term "fitness" was introduced. "Fitness" can be considered as the importance of a node, and the connection probability of each node is proportional to the product of fitness and degree. Caldarelli developed the fitness model that only the "fitness" of each node (Caldarelli et al. (2002)) is considered as the connection probability. The higher the fitness is, the higher probability it receives the connections from new nodes. This idea can be adapted to the article citation network such that the more important the article is, the more citations the article will receive from new articles.

There are several special features in the article citation networks that are worthy to mention. First, articles are added to the citation networks annually via the citation mechanism. Although articles are written and published continuously over the years, scientific journals are published in the form of regular issues several times in a year. It is difficult to find out the accurate publication time of each article in the Web of Science database. Second, it is general for an article to get an increasing number of citations in the first several years after publication, followed by a gradual decrease. Third, there are only a few good articles and most of the rest articles receive few numbers of citations. This phenomenon implies that a citation network has the scale-free-like property.

By incorporating these features, the goal of this chapter is to propose a dynamic model on the structure evolution of a citation network under a specific subject in the Web of Science. In particular, our data comes from the article citation network under the subject of "Statistics & Probability".

4.2 A Brief Introduction to Statistics and Probability Citation Networks in Web of Science

Web of Science database contains 266 different subject fields of articles, such as Engineering, Social science, and many others. In this paper, we focus on the "Statistics

& Probability" article citation network as a demonstration. The articles with the tag "Statistics & Probability" are extracted from the Web of Science database to construct the article citation network. We express this network as a graph and some terminologies are provided below.

An article citation network graph (hereafter we call citation network) is defined as $G = \{V, E\}$, where V is a set of nodes referring as articles and E is a set of directed edges that show citation relations between two articles. Since each node has the attribute of published year, we can divide V into different subsets $\{S^{(y_1)}, ..., S^{(y_n)}\}$ based on the published years, where $S^{(y)}$ is a subset of articles published in year y. Let $S^{(y)} = \{v_1^{(y)}, v_2^{(y)}, ..., v_{N^{(y)}}^{(y)}\}$ and $N^{(y)} = |S^{(y)}|$, where $v_i^{(y)}$ is the i-th article published in the year y. Figure 4.1 shows the structure of the citation network. Assume year y is the year with records in the database. In year $y + s$, $n^{(y+s)}$ articles are published and the citations/edges between articles of year $y + s$ and articles of all years before year $y + s$ (up to year y) appear at the same time.

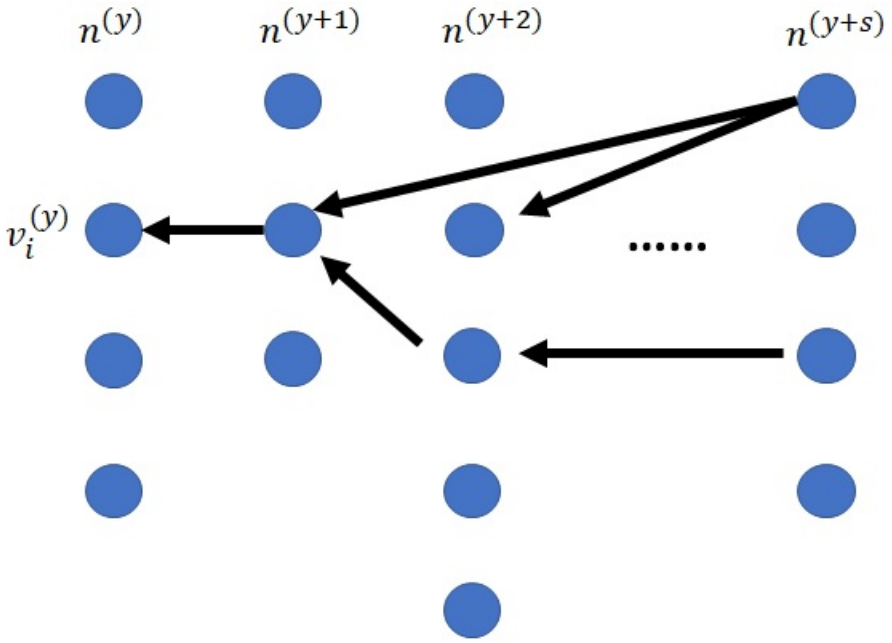

Figure 4.1: Struture of article citation network.

The database we use contains the information from 1981 to 2016. There are several descriptive observations of the citation network which we analyze in this paper.

Figure 4.2 shows an increasing trend in the annual number of published articles from 1981 to 2016. Note that the number of articles published in 2016 is about five times of that in 1981.

It is normal to have citations occurred when an article uses older articles as refer-

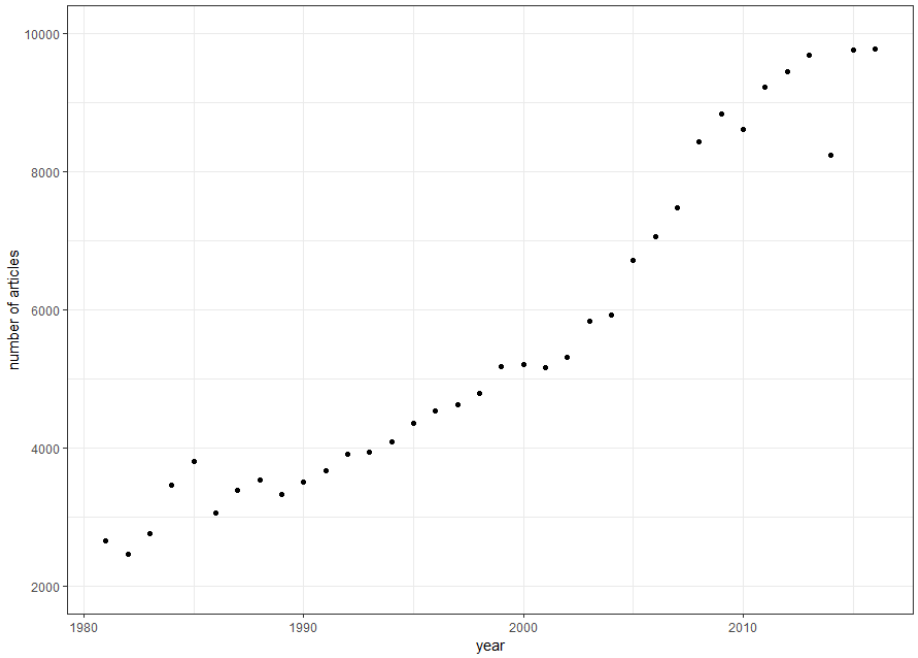

Figure 4.2: Number of article published in each year.

ences, but there is a small amount of citations that an article cites an article published in the later year due to publication issues. Since this scenario is extremely rare, we decide to ignore it in the entire analysis.

Next, we check the in-degree distribution of all articles. Degree is the number of connections a node has in the network. In a directed network, such as the citation network, there are two types of degrees. In-degree indicates the number of citations an article received in the citation networks, while out-degree indicates the number of references an article referred. This paper investigates the in-degree of an article as it implies the citation attractiveness of an article from its later articles. Figure 4.3 shows the scale-free property of the citation network. The pattern is similar to the Web hyper-link distribution (Albert et al. (1999)).

There is a special characteristic of citation network that gradually changes over the 30-years period. Let $d_i^{(y_1,y_2)}$ be the number of citations that an article $v_i^{(y_1)}$, published in y_1, accumulated from y_1 to y_2, where $y_2 = y_1 + k$ for $k = 0, \ldots, 10$. Mathematically, we write

$$d_i^{(y_1,y_2)} = \#\{v_i^{(y_1)} \leftarrow S^{(y_2)}\},$$

where # indicates the number of the elements of set. The proportion of citations is defined as

$$\hat{c}^{(y_1,y_2)} = \frac{\sum_{i=1}^{N^{(y_1)}} d_i^{(y_1,y_2)}}{N^{(y_1)} N^{(y_2)}}. \tag{4.1}$$

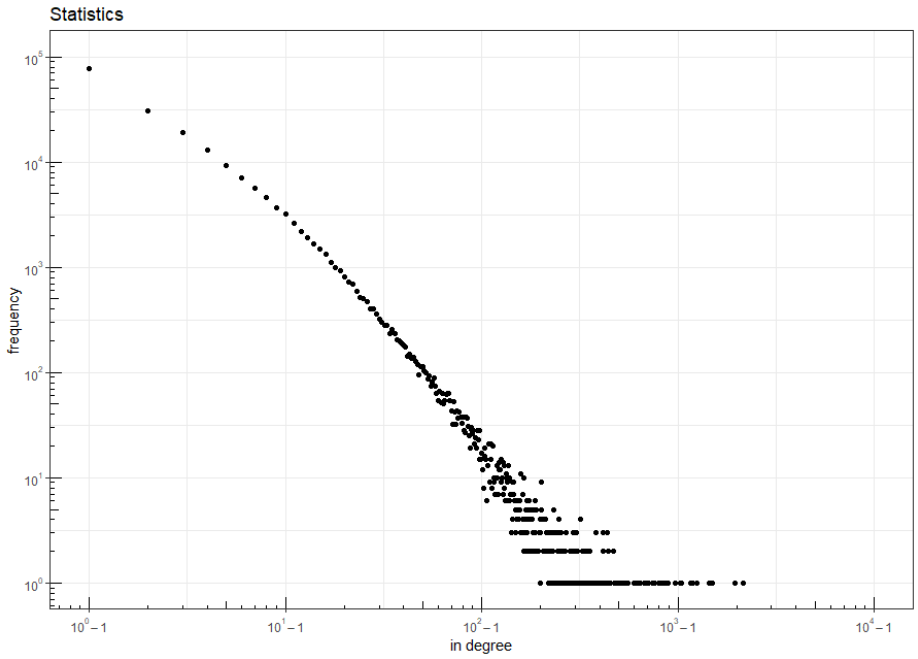

Figure 4.3: In-degree distribution of 36 years article citation network

where the number of citations occurred is divided by the possible number of edges. Figure 4.4 shows the plot of the citation proportion changes over 10 years from different publication years. There are two observations. First, most of the articles receive their highest number of citations in its third or fourth year after they published, and the number of citations will decrease after it reached the peak. Second, by comparing different starting years, the citation proportion generally decreases when the time progresses, and it becomes relatively stable in recent years.

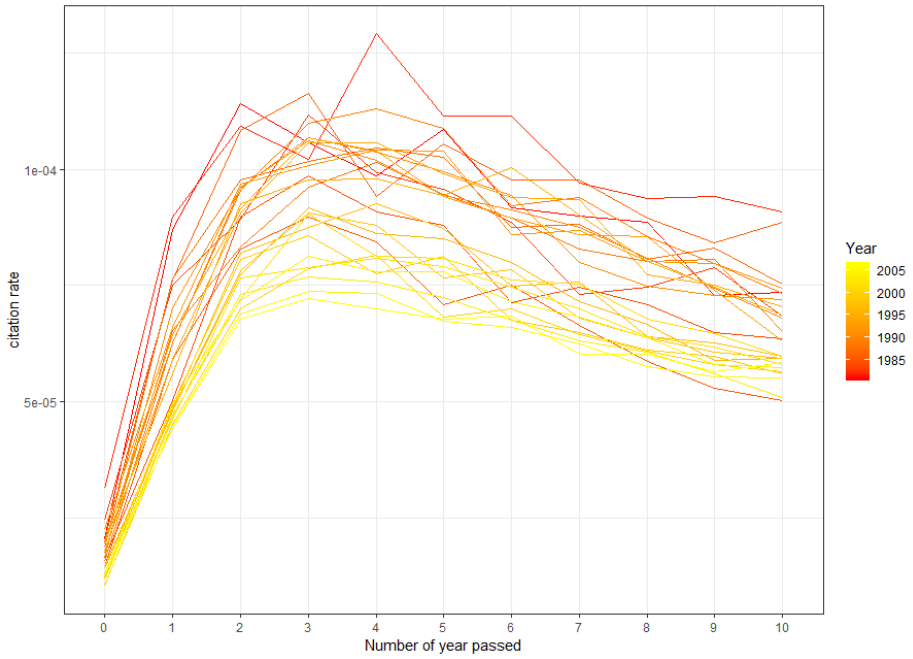

Figure 4.4: Citation rate

4.3 A Generative Model for WoS Citation Network

We propose a statistical model to describe the dynamic and structure of the citation network. We define the probability that the i_1-th article in year y_1 (denoted as $v_{i_1}^{(y_1)}$) received citations from the i_2-th article in the year y_2 (denoted as $v_{i_2}^{(y_2)}$) by

$$Pr\{v_{i_1}^{(y_1)} \leftarrow v_{i_2}^{(y_2)}\} = a^{(y_1,y_2)}\eta_{i_1}^{(y_1)}c^{(y_1,y_2)} \tag{4.2}$$

where $\eta_{i_1}^{(y_1)}$ shows the "importance" of the article $v_{i_1}^{(y_1)}$, $c^{(y_1,y_2)}$ shows the cited rate of the article in the year y_1 from articles in the year y_2 and $a^{(y_1,y_2)}$ is an standardization constant. We describe how to obtain these parameters below.

For $c^{(y_1,y_2)}$, we assume that $y_2 = y_1 + k$, $c^{(k)} = g(k|\alpha,\beta,\gamma,\delta)$, where g is a scaled gamma probability density function defined by

$$g(x|\alpha,\beta,\gamma,\delta) = \delta\frac{\beta^\alpha}{\Gamma(\alpha)}(x+\gamma)^{\alpha-1}e^{-\beta(x+\gamma)}. \tag{4.3}$$

$c^{(y_1,y_2)}$ seems to depend on both y_1 and y_2 in Figure 4.4, but in fact, the curve shapes

are similar in the sense that $c^{(y_1,y_1+k)}$ increases when $k = 0, 1, 2, 3$, and decreases when $k = 5, 6, ..., 10$. Thus, we simplify that $c^{(y_1,y_1+k)}$ depends on k only and not on y_1.

We assume that the importance η_i of articles $v_i^{(y_1)}$ in the network follows tapered Pareto distribution (Kagan and Schoenberg (2001)). The tapered Pareto distribution has a cumulative distribution function

$$F(x) = 1 - (\frac{\kappa}{x})^{\nu} e^{-\frac{x-\kappa}{\theta}} \tag{4.4}$$

and density

$$f(x) = (\frac{\nu}{x} + \frac{1}{\theta})(\frac{\kappa}{x})^{\nu} e^{-\frac{x-\kappa}{\theta}} \tag{4.5}$$

for $x \geq \kappa$, where κ is the user-defined lower truncation point, ν is the shape parameter for the function decay, and θ is a parameter regarding where the exponential taper to zero. In our assumption, parameters of tapered Pareto distribution do not depend on years.

As the magnitudes of η_i is not decided clearly, we use $a^{(y_1,y_2)}$ to make the product of these factors as a proper probability.

4.4 Parameter Estimation for the Generative Model

Since there are many parameters in the statistical model, it is difficult to estimate all parameters simultaneously using the maximum likelihood method. Instead, we estimate these model parameters via a 3-stage approach.

First, we assume that the expectation number of articles follow the sigmoid function:

$$y = a + \frac{b-a}{1 + e^{-c(x-d)}}. \tag{4.6}$$

because it gradually increases over the years and becomes stable in the recent years. The parameters are estimated via a least square method package in R called nls and they are $a = 2983.8, b = 10919.0, g = 0.1757, h = 2005.3$. Figure 4.5 shows that the estimated number of articles from the sigmoid function roughly matches the values from the true data.

Second, we estimate $c^{(k)}$ via $\hat{c}^{(y_1,y_2)}$. As we assume before that $c^{(k)}$ can be fitted by scaled gamma probability distribution function, we use the least square method and minimize the term below

$$\sum_{y_1=1981}^{2006} \sum_{k=0}^{10} \{\hat{c}^{(y_1,y_1+k)} - g(k|\alpha, \beta, \gamma, \delta)\}^2,$$

Then the optimized parameters are obtained: $\alpha = 1.7586, \beta = 0.1894, \gamma = 0.1633, \delta =$

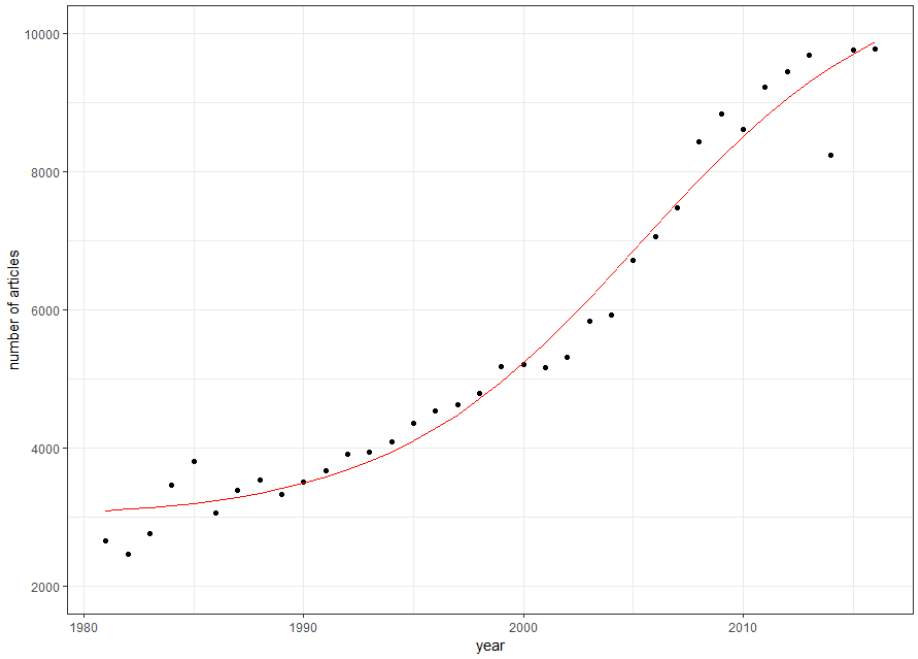

Figure 4.5: Fitted sigmoid function

0.0012

Third, we estimate the parameter of the importance via the R package PtProcess. However, the importance of an article cannot be directly measured. Therefore, we consider the in-degree of an article as a surrogate quantity for the importance of an article. It is based on a simple idea that the number of citations received by an article is related to its quality and importance. The better quality and the more important an article is, it is very likely that the higher the number of citations from later articles, the larger the number of in-degree. Thus, we investigate in the in-degree distribution in order to understand the underlying distribution of importance.

In addition, we observe that the characteristic of the citation network change over years, but the change is slow, so we assume that the characteristic remains unchanged during ten consecutive years. Thus, we extract the in-degree of articles published the first year from each 10-years network. For example, we extract the in-degree of articles published in 20000 received from a network that consists of articles being published during 2000-2009. We then fit the in-degree of each year with tapered Pareto function. Figure 4.6 shows the in-degree distribution of 1981 and 1982 from 1981-1991 network and 1981-1992 network. The red line is the fitted tapered Pareto model.

We obtain different parameters for the in-degree distribution of each year using its next 10-years information and we list them in Table 4.1. The range of ν is around [0.68, 1.31] and the range of θ is around [30.96, 56.49] shown in the table.

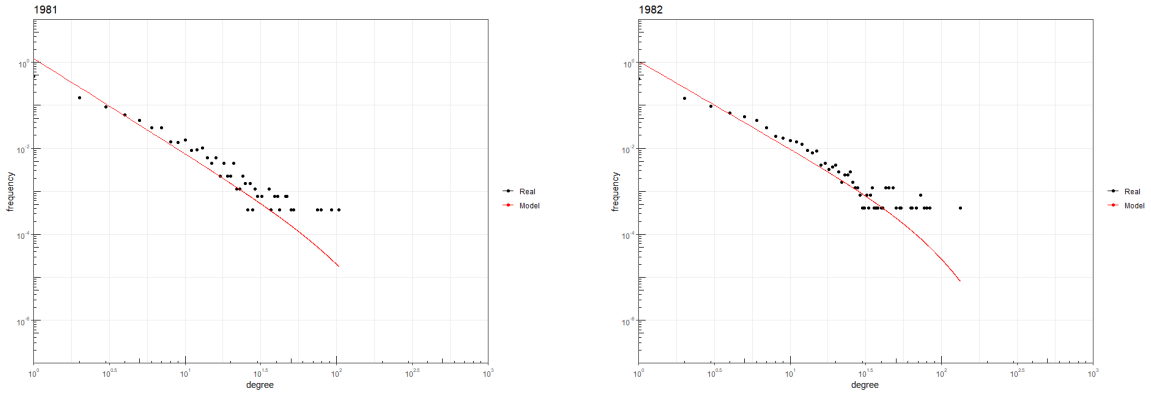

Figure 4.6: 1981 and 1982 degree distribution fit with tapered Pareto function

As we mentioned before, the in-degree is only a surrogate version of the importance of an article. It is just an approximation to the parameter of importance, and we do not have an accurate parameter estimates at this stage. We provide an improved parameter estimates using simulations in the next section.

Table 4.1: The parameter of the in-degree distribution.

	ν	θ
1981	1.14	55.90
1982	1.00	51.22
1983	1.14	50.07
1984	1.31	55.41
1985	1.36	52.89
1986	0.99	52.03
1987	1.06	54.15
1988	1.04	46.34
1989	0.92	48.46
1990	0.99	56.49
1991	0.94	47.93
1992	0.90	52.38
1993	0.88	43.25
1994	0.89	49.71
1995	0.91	55.47
1996	0.91	46.93
1997	0.86	46.29
1998	0.86	43.98
1999	0.83	43.70
2000	0.82	44.70
2001	0.76	44.92
2002	0.76	46.98
2003	0.73	44.19
2004	0.68	38.81
2005	0.67	30.96
2006	0.68	38.56

4.5 Simulation and Model Improvement

We improve our model by minimizing the dissimilarity between the original network and the network simulated from our model. This approach is based on a fact that a randomly generated graph based on a proper model should be very similar to the original graph. In this sense, we need to first define the measure of graph dissimilarity. There are many criteria to characterize a network, but the simplest and the most important one is to compare the degree distribution. Since the importance of an article is of our interest, we follow the study in the previous section and use the in-degree distribution as the important property of the graph. On top of the dissimilarity measure of the in-degree between two networks, we decide to use distance-2 in-degree distribution as an additional information of the dissimilarity.

Distance-2 in-degree of a node describes the number of nodes that the target node indirectly connects through its neighboring nodes. Figure 4.1 is a toy example to show the node with distance-2 in-degree. Let $v_i^{(y)}$ be the article of interest, the in-degree of $v_i^{(y)}$ is 1 as it only received one direct citation from an article, denoted as $v^{(y+1)}$, published in later years. Then $v^{(y+1)}$ received two direct citations from articles published in its later years, and we denote them as $v^{(y+2)}$ and $v^{(y+s)}$ respectively. Since $v^{(y+2)}$ and $v^{(y+s)}$ do not have direct connections to $v_i^{(y)}$, their shortest paths to $v_i^{(y)}$ are formed through $v^{(y+1)}$. This means the distance-2 in-degree of $v_i^{(y)}$ is 2.

Then we improve the model via simulations as follows. (1) We generate the number of nodes in each year $N^{(y_i)}$ by using the expected number of nodes from the model estimated in Figure 4.2; (2) For each node, we generate the importance η_i based on the tapered Pareto model with one parameter set. (3) Based on the probability of a citation occurred $Pr(v_{i_1}^{(y_1)} \leftarrow v_{i_2}^{(y_2)})$ in Equation (4.2), the citation network is constructed. Note that $a^{(y_1,y_2)}$ is defined to produce an appropriate number of edges, i.e. $a^{(y_1,y_2)} = a^{(y_1)} = \frac{N^{(y_1)}}{\sum_{i=1}^{N^{(y_1)}} \eta_i^{(y_1)}}$. We repeat the network generations 30 times to reduce the uncertainty and inconsistency of degree distributions. The average number of degrees is then used to compare with the original graph.

It is well-known that the network degree follows power-law-type distribution, which means most of the nodes have no or small number of connections while some hubs exist. These hubs can be considered as the outliers since their degrees are not trivial to predict. Therefore, we neglect the information from the hubs and the independent nodes and we build our improved model via the information from the remaining nodes, which are represented as the middle part (majority) in the degree distribution. We minimize the Kullback-Leibler divergence of distance 1 as graph dissimilarity.

Regarding the parameter selection, we use the yearly parameters estimated in section 4 as the candidates in the parameter set, then we have 10 parameter sets

for a 10-years graphs. We choose the one that can minimize the Kullback-Leibler divergence as the characteristic parameter of a network. Following this approach, the best parameter of every network can be found. We check the degree distribution of the estimated model with the original network. The distance-2 degree distribution is also used as the additional comparison.

The simulation results are shown in Figure 4.7 - 4.23 and here are some observations when time goes forward. First, the degree distribution of the simulated network move to the right of the real network. Second, the distance-2 degree distributions get a better fit in later years than those in earlier years. This makes us believe that the structure of the network evolves through time. Figure 4.24 shows both the number of article and citations show an increasing trend over the years. The proportion of citations and articles is clearly different between earlier and later years. Figure 4.25 shows the average number of references an article used within 10-year spans, so the first value labeled 1990 stand for the number of references of an article published in 1990 used during 1981-1990 divided by the total number of articles published in 1990.This figure shows that the average number generally increases, indicating that the recent articles tend to use more recently published articles as the reference. It may be an indication that the technology advancement is fast and thus the recent articles tend to reference newer articles than older ones.

In general, we can follow the same spirit to consider distance-3 in-degree distribution, or any in-degree distribution of further distances. However, these long-distance in-degree distribution usually requires heavy computation. Since they do not provide any exceptionally new information when compared to distance-2 in-degree distribution, we choose not to conduct these computations. In addition, we choose the parameters estimated from 10-years distance-1 in-degree distribution, because we try to perform numerical optimization to minimize dissimilarity defined by Kullback-Leibler divergence. The values of these parameters are generally no big difference from those obtained using distance-1 in-degree distribution of 11 years or longer.

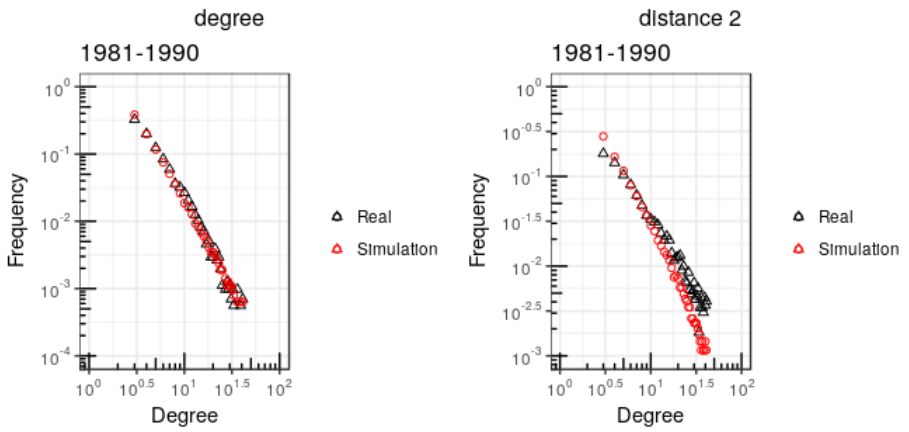

Figure 4.7: 1981-1990 network

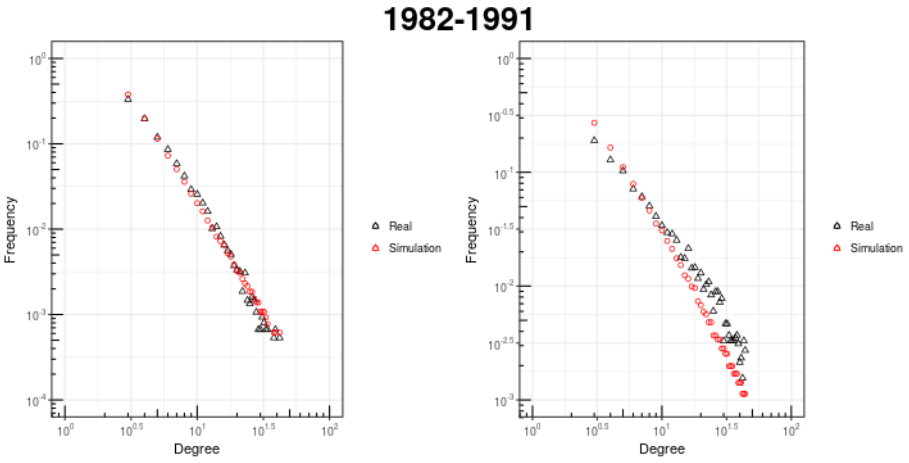

Figure 4.8: 1982-1991 network

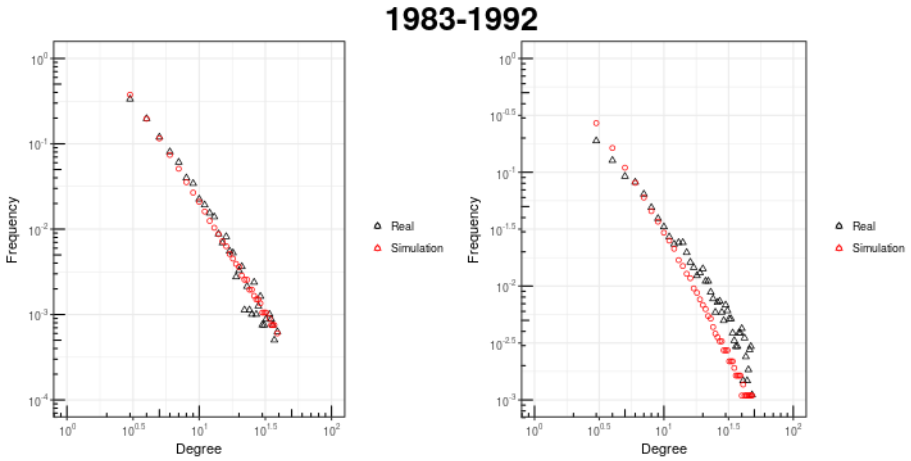

Figure 4.9: 1983-1992 network

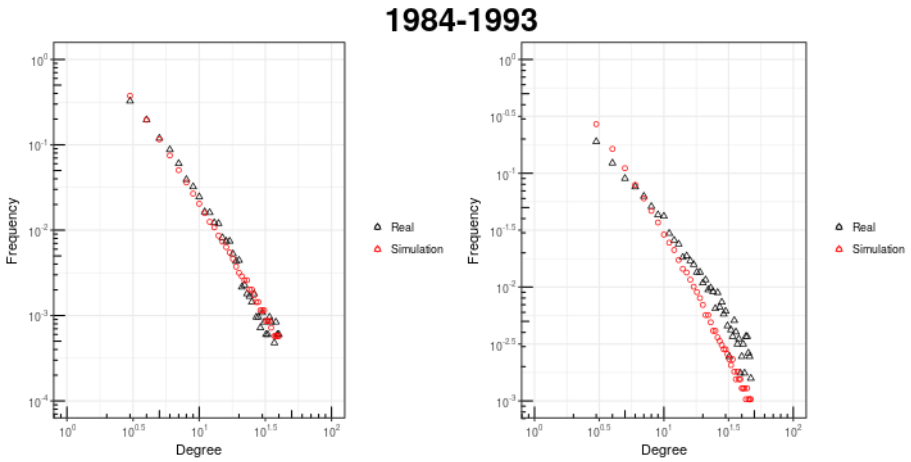

Figure 4.10: 1984-1993 network

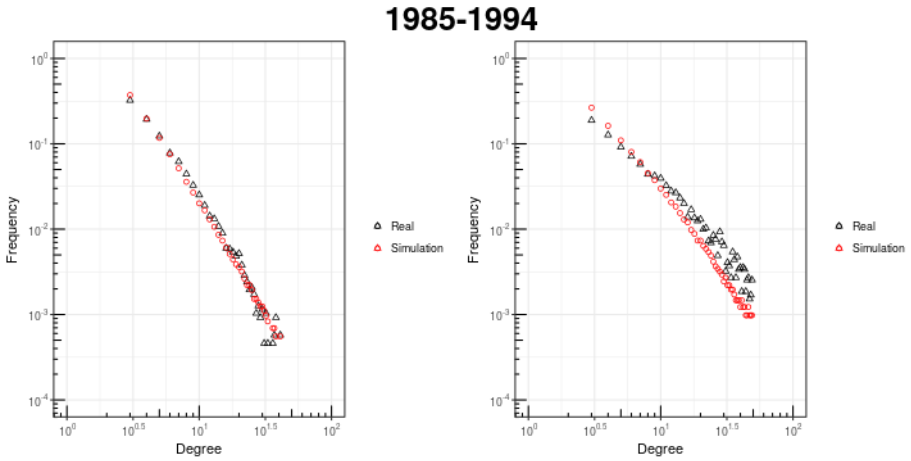

Figure 4.11: 1985-1994 network

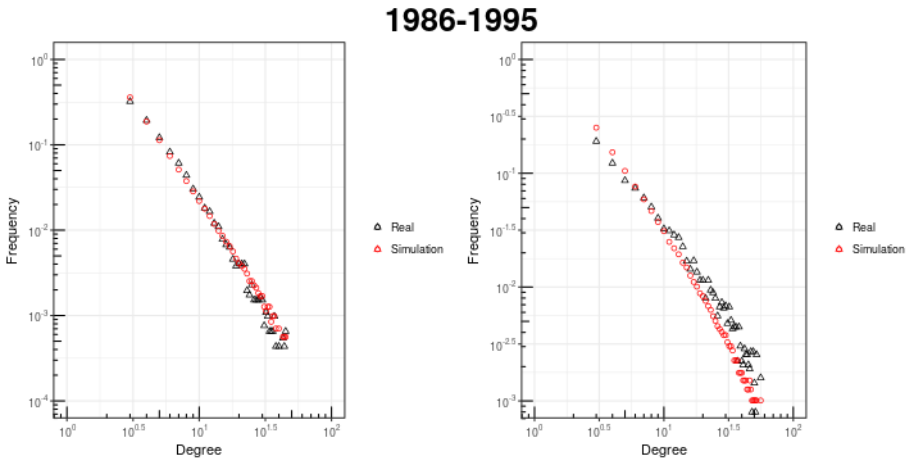

Figure 4.12: 1986-1995 network

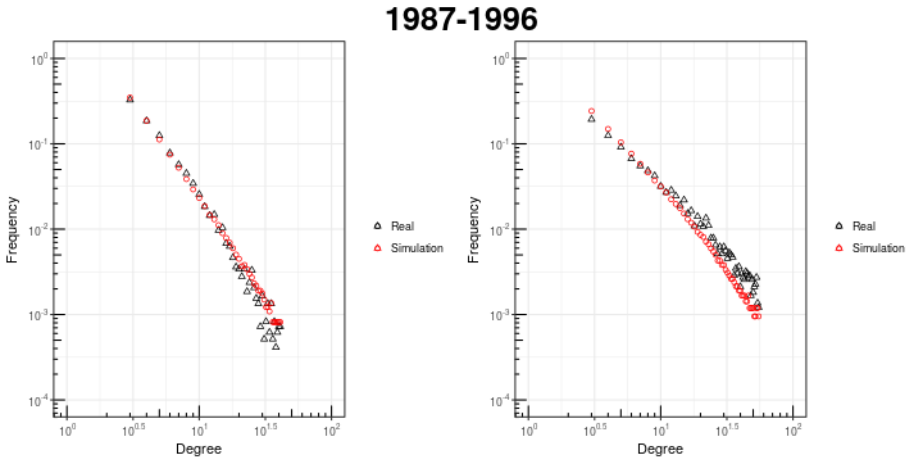

Figure 4.13: 1987-1996 network

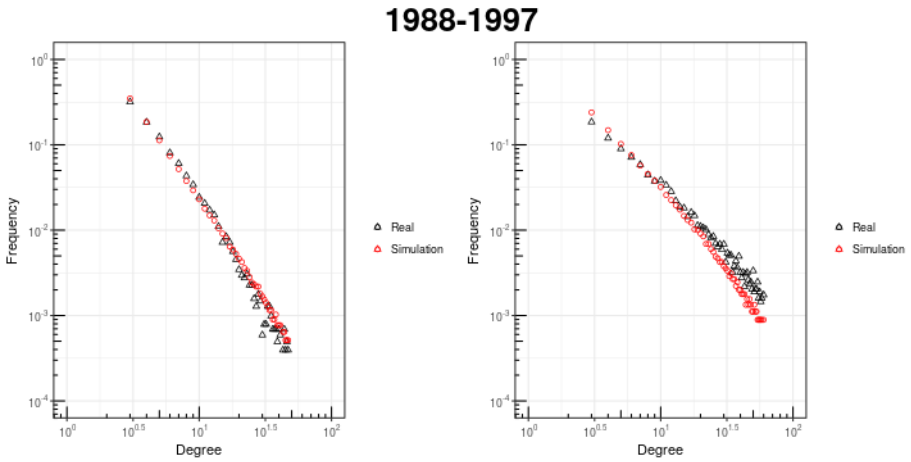

Figure 4.14: 1988-1997 network

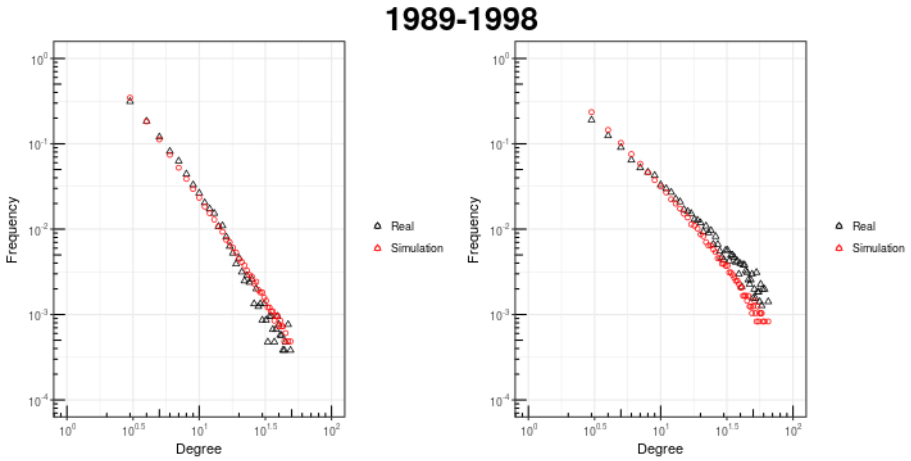

Figure 4.15: 1989-1998 network

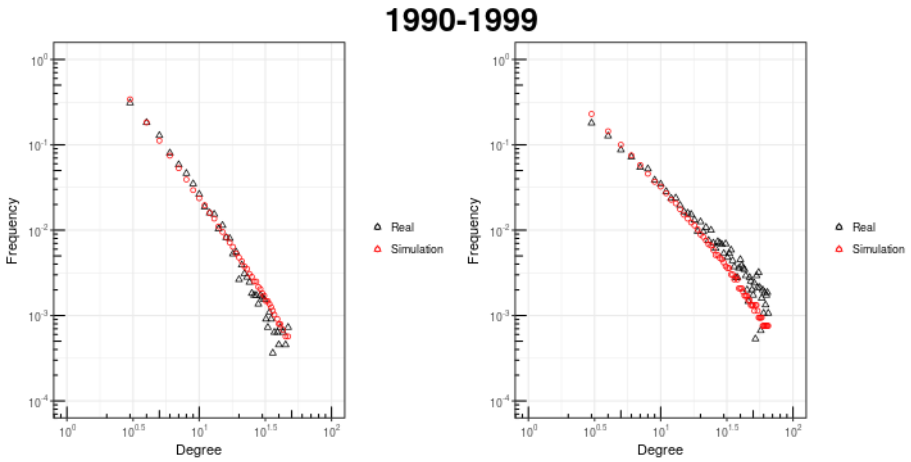

Figure 4.16: 1990-1999 network

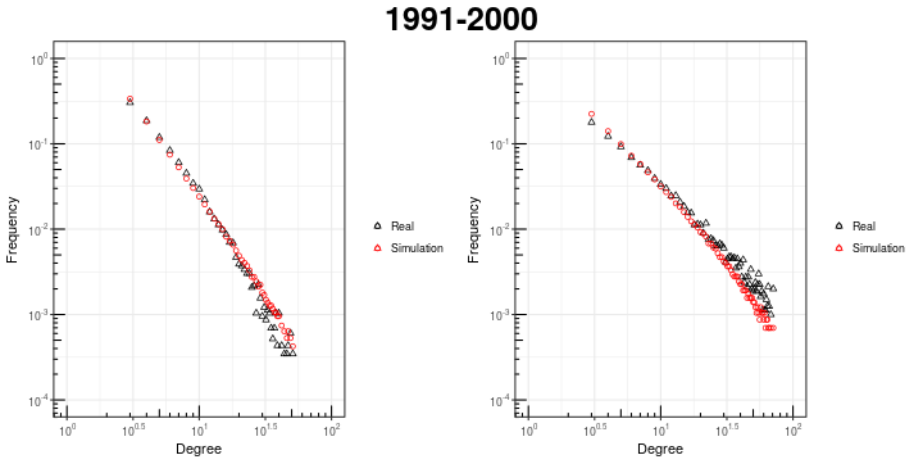

Figure 4.17: 1991-2000 network

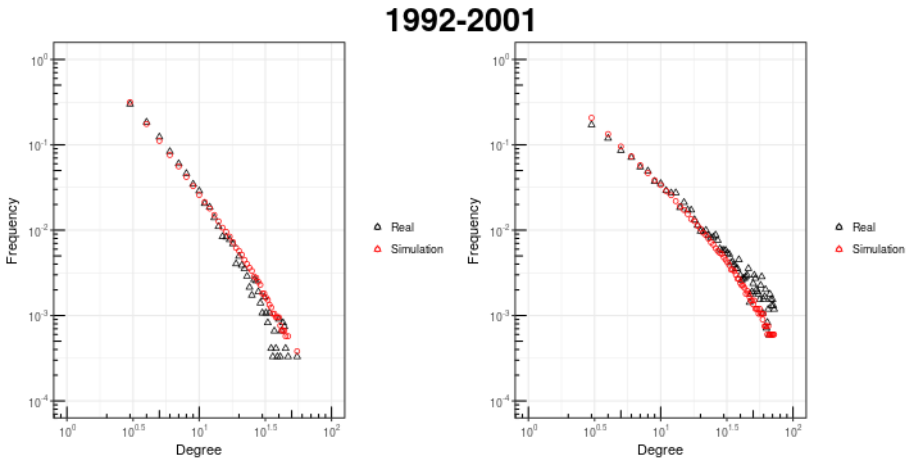

Figure 4.18: 1992-2001 network

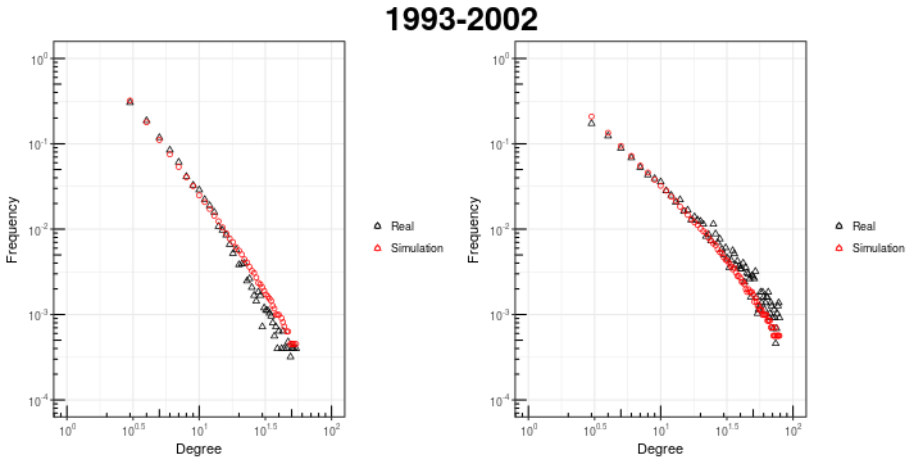

Figure 4.19: 1993-2002 network

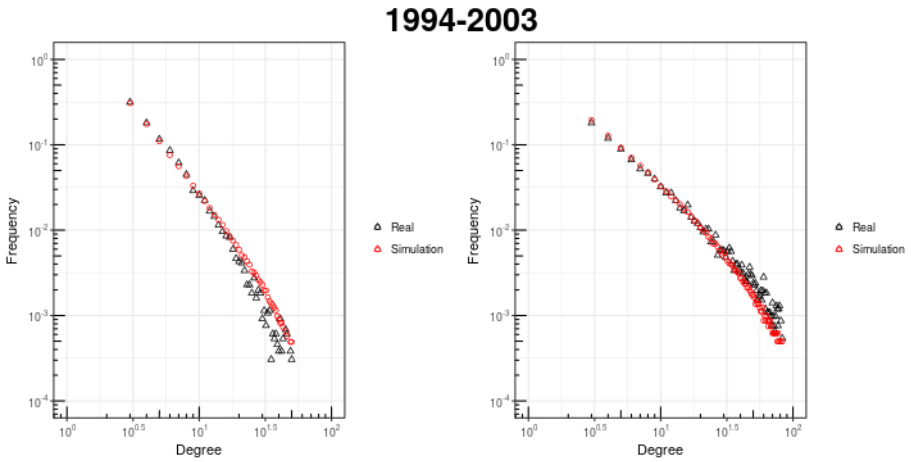

Figure 4.20: 1994-2003 network

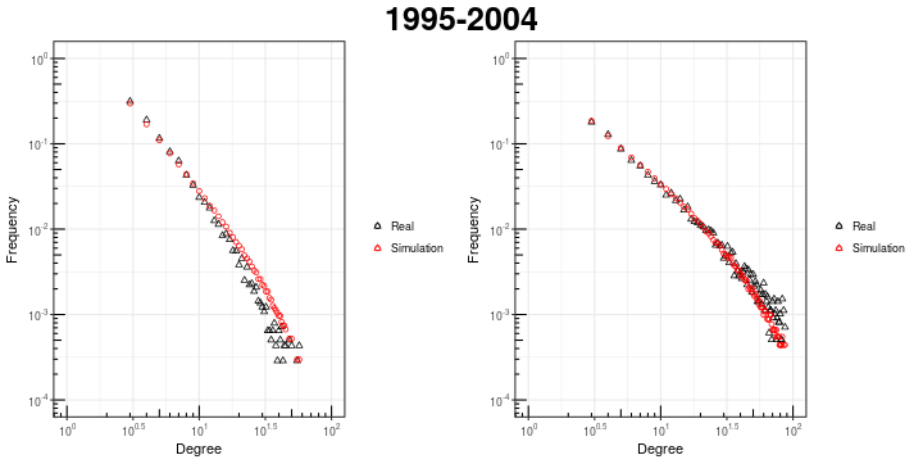

Figure 4.21: 1995-2004 network

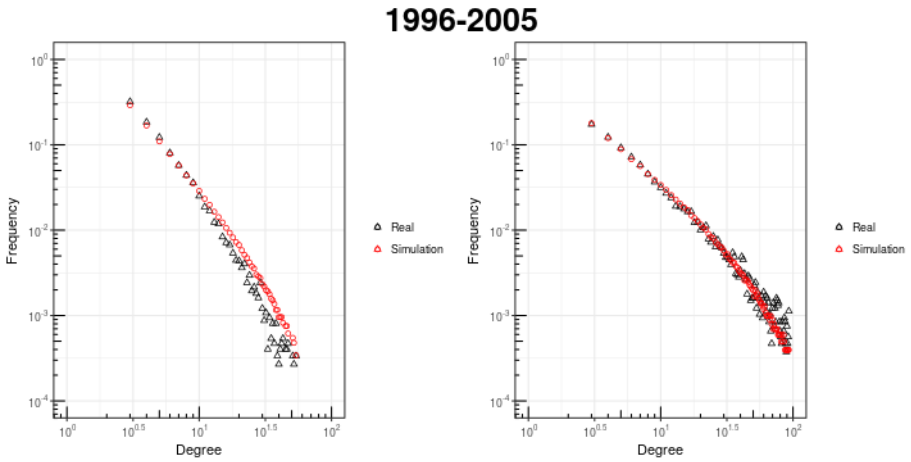

Figure 4.22: 1996-2005 network

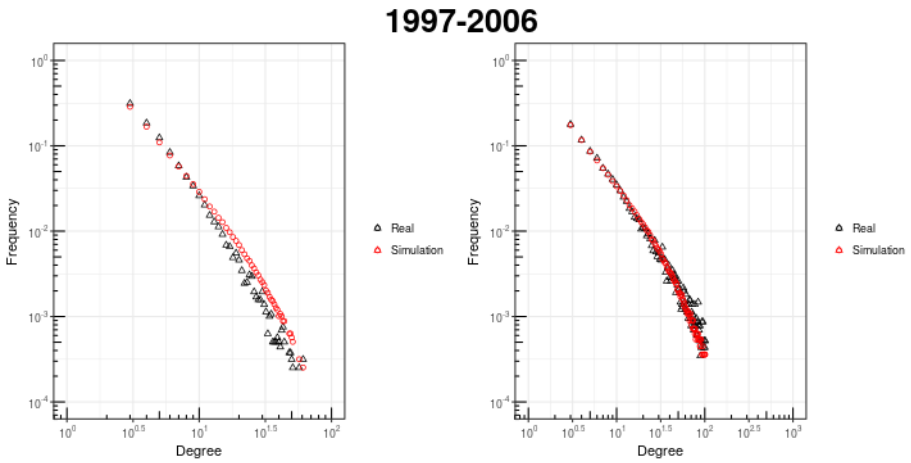

Figure 4.23: 1997-2006 network

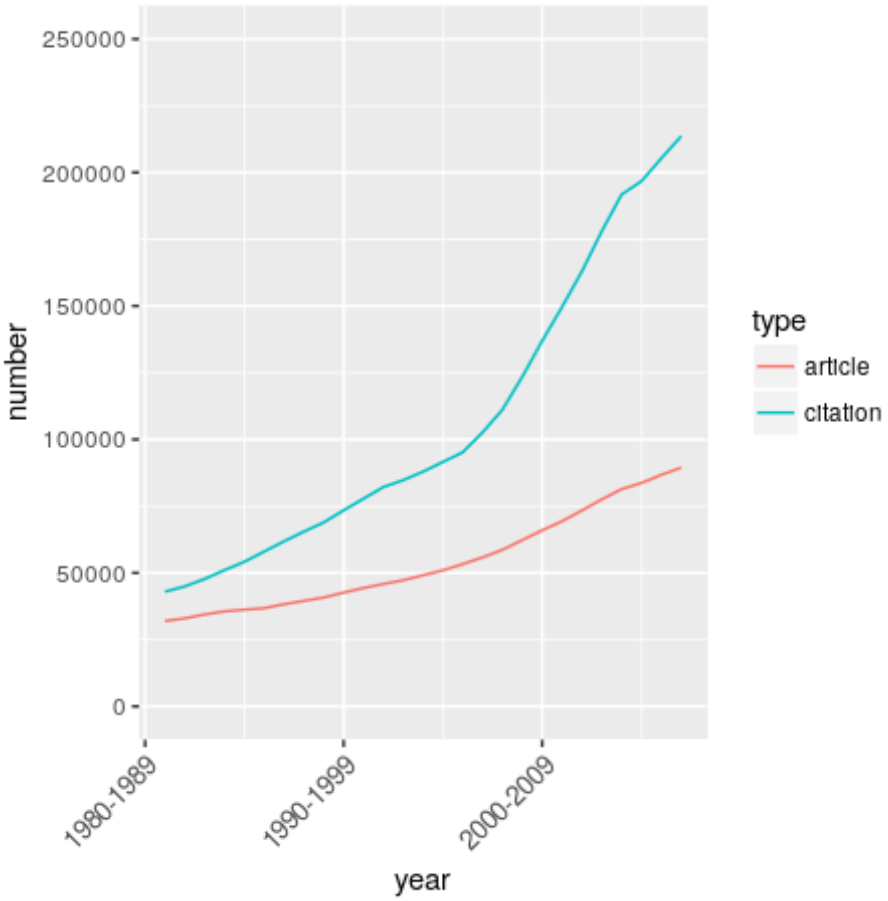

Figure 4.24: Number of articles and citations in each year

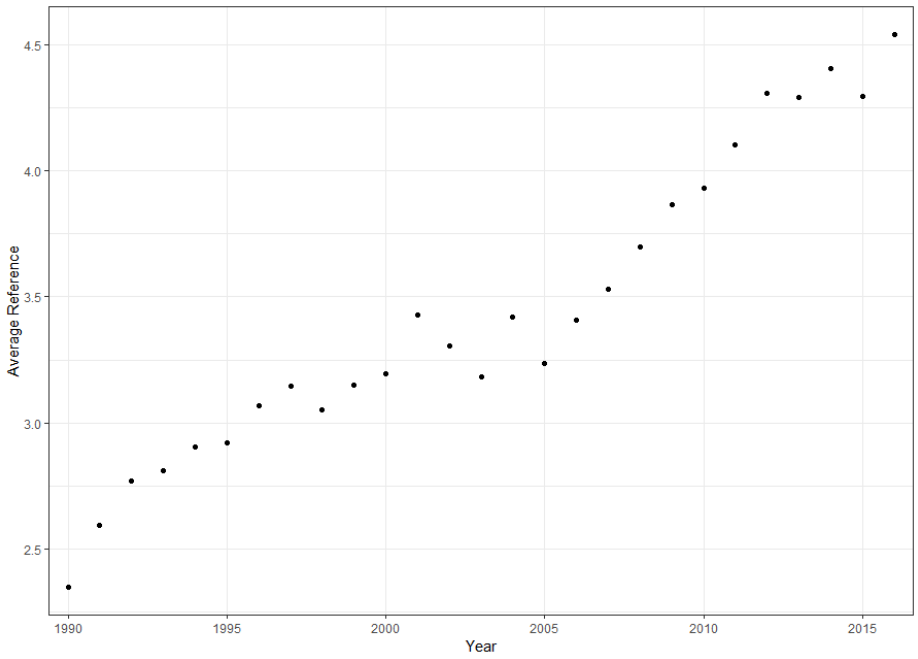

Figure 4.25: Average number of reference the article used within 10 years

Chapter 5

Concluding Remarks

In this thesis, we comprehensively study the structure and property of the citation networks obtained from one of the largest scientometric databases in the world called the Web of Science (WoS). In the first chapter, we first review the steps of the pioneers in the researches of network centrality and research metrics. Then the next three chapters provide several key contributions summarized below.

In the second chapter, we define six new ratio-type measurements to describe the properties of a general network, and we apply them to describe the citation relationships. They not only state the number of citations in the same field, but also indicate the cross-disciplinary citations from one field to another. We study the cross-disciplinary citations from all fields in WoS to statistics, showing which fields cite statistics articles a lot while other fields seldomly cite statistics articles. In addition, we also illustrate some phenomena described by these measurements. For example, if a subject has high average cited rate and low cited article rate, it may indicate the presence of a hub in a network.

In the third chapter, we introduce a new quantity called article network influence to measure the importance of an article when a citation network is given. Through the article citations, the influence of an article is propagated not only to its followers but also the followers' followers and so on. We define the quantity called influence range to describe this hierarchical relationship. We apply this new quantity to study the citation network under the topic of statistics and probability during 1981-2016, and we provide tables of the top 20 influential articles during every 10 years. There are some fundamental differences between several traditional measures to network influence, and we discuss these differences in details. Unlike other methods which only use subset of the network, we consider the whole network for evaluating the article performance. Furthermore, instead of evaluating the importance of the article, we are evaluating the propagation/diffusion ability of the article. It is the main reason that we consider the whole network instead of subset of the network.

Note that we only consider the articles which are published in the journals. We

do not include conference paper or working paper in our analysis because of the lack of data. Furthermore, the data we can access is from 1981 to 2016, we are not able to use the citation data after 2016 so our result may not reflect the current situation. For example, Annals of the Institute of Statistical Mathematics (AISM) receives much attention these days which indicates the performance of AISM is improving. However, since we do not have the current citation data, AISM did not show up in the top 10 statistics journal in the year 2016.

In the fourth chapter, we analyze the structure of the article citation network. We consider a probability-based evolution mechanism to describe the formation of an article citation network. Based on this mechanism, we propose a generative model for the generation of the statistics citation network obtained from the WoS. In specific, in order to obtain a key component called the importance of an article, we first study several important properties like in-degree distribution and citation rate. Then by assuming that this importance follows tapered Pareto distribution, we form our generative model and it is further improved by simulation studies.

The studies in this thesis open a wide door for statisticians to participate in the research of scientometrics and the development of new research metrics. There are many potential future works derived from this thesis. We list some of them below as examples.

The study of the second chapter is actually a subset study on the whole cross-disciplinary citation analysis in the WoS database. Note that this chapter only considers the cross-disciplinary citations of statistics articles from all other fields, but in fact, similar studies can be done on any selected two fields in WoS. Thus, the main contribution of the second chapter is to propose six measurements for network descriptions. It is of great interest to develop a real-time automatic system that a user can enter two subjects in WoS and a report of cross-disciplinary citations is returned.

The study of the third chapter on article citation network is actually the simplest case among all the networks in institutional research such as author network or institutional network, because there are several special properties of article citation networks that greatly reduce its computational complexity. For examples, the binary edges in article citation network reduce the calculation associated to adjacency matrix. The citation networks in other layers, like authors, institutes, journals, or even countries, cannot be simplified like before. It is of great interest to generalize the network influence to accommodate these generalized scenarios in other layers of citation networks.

The study of the fourth chapter on the generative model of citation network is innovative in the current framework of network analysis. However, we mainly focus on the article citation network in statistics, and it would be interested to know if other subjects have different parameter for the generative model. Also, the distribution of ANI can be calculated from the simulated network to check the similarity with the

original network. Furthermore, there are still several potential improvements about this model. For example, this generative model is still based on the importance of an article to determine the connection probability from other articles of later years, which is still a variant of preferential attachment. It is questionable whether this "rich-gets-richer" phenomenon is enough to describe such a complicated mechanism of network evolution. Moreover, the importance of an article is an unobserved quantity, so there are several statistical tools that can be applied to unmask this quantity and it will be left in future projects.

References

Albert, R., Jeong, H., and Barabsi, A.-L. (1999). Diameter of the world-wide web. *Nature*, 401:130–131.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Barabási, A.-L., Albert, R., Jeong, H., and Bianconi, G. (2000). Power-law distribution of the world wide web. *Science*, 287:2115a.

Baton, J. and Bruggen, R. V. (2017). Learning neo4j 3.x (2nd edition).

Bavelas, A. (1950). Communication patterns in taskoriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Bianconi, G. and Barabási, A.-L. (2001). Competition and multiscaling in evolving networks. *Europhysics Letters*, 54(4):436.

Blondel, V., Guillaume, J.-L., M Hendrickx, J., and Jungers, R. (2008). Distance distribution in random graphs and application to network exploration. *Physical Review. E*, 76:066101.

Bogu, M. and Pastor-Satorras, R. (2003). Class of correlated random networks with hidden variables. *Physical Review. E*, 68:036112.

Caldarelli, G., Capocci, A., Rios, P. D. L., and noz, M. A. M. (2002). Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, 89(25):258702.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Elsevier (2018). Research intelligence: Research metrics guidebook.

Erdös, P. and Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41.

Fronczak, A., Fronczak, P., and Hołyst, J. A. (2004). Average path length in random networks. *Physical Review. E*, 70:056110.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gilbert, E. N. (1959). Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144.

Gualdi, S., Medo, M., and Zhang, Y.-C. (2011). Influence, originality and similarity in directed acyclic graphs. *Europhys Letters*, 96(1):18004.

Kagan, Y. Y. and Schoenberg, F. P. (2001). Estimation of the upper cutoff parameter for the tapered pareto distribution. *Journal of Applied Probability*, 38A:168–185.

Mariani, M. S., Medo, M., and Zhang, Y.-C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, 10(4):1207 – 1223.

Newman, M. E. J. (2008). The first-mover advantage in scientific publication. *Europhysics Letters*, 86(6).

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

Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 25(4):348 – 349.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Wang, T.-C. and Phoa, F. K. H. (2016). Focus statistics for testing network centrality on uncorrelated random graphs. *Network Science*, 4(4):460473.