

A Novel Method for Visualizing Keywords in Bibliometrics Science

by

Theyab Atallah H Alhwiti

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 5, 2017

Keywords: Data visualization, Exploratory Data Analysis, Clustering Analysis,
Bibliometrics Mapping, Similarity Measures

Copyright 2017 by Theyab Atallah H Alhwiti

Approved by

Fadel Megahed, Chair, Assistant Professor of Industrial and Systems Engineering
Richard Sesek, Assistant Professor of Industrial and Systems Engineering
Robert Thomas, Professor Emeritus of Industrial and Systems Engineering
Cheryl Seals, Associate Professor of Computer Science and Software Engineering

Abstract

In the last two decades bibliometrics science has been evolving to help scientists and researchers to maintain the drastically increasing of the availability of scientific literature. Bibliometrics help to have a broad understanding of an intended scientific field by providing multiple levels of methods and tools to analysis the literature of that scientific field. Unfortunately, the bibliometrics science is a quantitative method that use basic math to assess author outputs by h-index, articles impact by counting citations, or journal by its impact factor. Thus, the objective of this dissertation is to pave a solid path to integrate qualitative method into the emerging field of bibliometrics science. For that, I am going to : (1) apply the bibliometrics methodology on the literature of the statistics field by giving a visual representation of the bibliographic data of the field, that will identify emerging trends and understanding relationships between different developments in the field. Also, (2) study the similarity measures on the co-word analysis then introducing an integrated method between the quantitative and qualitative approaches to calculate the similarity between keywords in co-word analysis. Finally, (3) presenting the important of a better visualizing presentation of keywords on a 2D map by introducing a qualitative method to calculate the weight (strength) of keywords in clustering analysis, within bibliometrics data.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Dr. Fadel Megahed for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. Besides my advisor, I would like to thank my committee members: Dr. Richard Sesek, Dr. Robert Thomas and Dr. Cheryl Seals for their insightful comments. With special thanks to Mohammad Ali Alamdar Yazdi for his kindly help participating in my research. Also, I would like to thank all my co-authors of my journal papers. I would also like to thank all my friends who always supported me Specially Dr. Nader Altheeb, Ahmed and Moath Alhelal. Last but not the least, I would like to thank my family: my respected parents and to my brothers and sister for supporting me spiritually throughout writing this thesis and my life in general. At the end I would like express appreciation to my beloved wife who spent sleepless nights with and was always my support in the moments when there was no one to answer my queries.

Table of Contents

Abstract	ii
Acknowledgments	iii
List of Figures	vi
List of Tables	viii
1 Introduction	1
1.1 Problem Description and Significance	1
1.2 Significance	2
1.3 Research Objective	3
1.4 Dissertation Layout	4
2 Bibliometric Mapping of (ASA) Journals from 1991 to 2016	5
2.1 Introduction	5
2.2 Motivation	5
2.3 Background	6
2.4 Methods	8
2.4.1 Data Extraction and Cleaning	8
2.4.2 Units of Analyses	9
2.4.3 Measures and Similarity Calculation	11
2.4.4 Mapping and Visualization	14
2.4.5 Analysis of Developments Over Time	16
2.5 Results	18
2.5.1 Keyword Level: Term and Temporal Term Maps	18
2.5.2 Publication Level: Cited Publication and Temporal Publication Maps	21
2.5.3 Journal Level: Journal and Temporal Journal Maps	25

2.5.4	Author Level Analysis	26
2.5.5	The Spatiotemporal Analysis	27
2.6	Conclusion and Future Work	29
3	Appropriate Similarity Measure for Co-word “Keywords” Analysis	33
3.1	Abstract	33
3.2	Introduction	33
3.3	Related Work and Background	38
3.4	Method	48
3.5	Empirical Comparison	52
3.6	Case Study	56
3.7	Conclusions And Future Work	62
3.7.1	An Overview of the Impacts and Contributions of this Paper	62
3.7.2	Practical Implications from our Work	63
3.7.3	Limitations and Future Work	64
4	Qualitative Weighted Keywords for Clustering Analysis in Bibliometrics Data.	65
4.1	Abstract	65
4.2	Introduction	65
4.3	Related Work and Background	71
4.4	Method	73
4.5	Empirical Comparison	76
4.6	Case Study	80
4.7	Conclusions And Future Work	86
5	Conclusion and Summary of Dissertation Contributions	87

List of Figures

2.1	Term map with colors indicating ten clusters of terms.	18
2.2	Temporal term map. Color indicates the probability of term use during 1991-2016.	19
2.3	A Boxplot of the term scores for each cluster.	20
2.4	Top 20 keywords in the ASA original keywords, 1991-2016.	21
2.5	Markov Chain Monte Carlo Keynes values.	21
2.6	Cited publication map with colors indicating eight clusters.	21
2.7	Temporal cited publication map, Color indicates the use of citation 1991-2016 .	24
2.8	Citation counts for Huber's paper since 1991. Data retrieved from WOS website.	25
2.9	Cited journal map.	25
2.10	Temporal cited journal map, Color indicates the use of journal during 1991-2016	26
2.11	Author co-authorship.	26
2.12	Author co-citation.	27
2.13	Spatiotemporal Visualization for institutions over the USA, 1991-2016.	28
3.1	Scatter plots obtained for the keywords between each of similarity measures. . .	53
4.1	Timeline of cited articles and citing articles.	75

4.2	Flow diagram of our method to calculate keywords' strength.	76
4.3	Flow chart of our method.	80
4.4	New method keyword's strength versus old method node size.	82
4.5	The node size represented by the frequency of appearance in the data-set.	83
4.6	The node size represented by keyword's strength.	83
4.7	Keyness values of the keyword Variable selection.	83

List of Tables

2.1	Clusters of cited publications.	23
2.2	The top 10 contributing institutions in ASA journals in the USA, 1991-2016. . .	28
3.1	Relations among various direct and indirect similarity measures.	46
3.2	Correlations obtained for Keywords.	53
3.3	Hypothetical scenarios.	54
3.4	List of acronyms and definitions of keywords used in the ASA.	58
3.5	CPCC evaluation results.	60
3.6	Clusters of our Keywords data-set.	61
4.1	Visualization tools used bibliometrics mapping.	70
4.2	Two keywords' example data	78
4.3	Data of the four keywords data	81
4.4	Ranking of four keywords data	82
4.5	The top 20 keywords based on our method.	85

Chapter 1

Introduction

1.1 Problem Description and Significance

The goal of my dissertation is to introduce a new method to assess the present understanding of research compilation, as well as to capture the evolution of scientific literature and the background of a given science, down to the subfield level. In addition, by increasing expediency and efficacy, the new method will be able to assist researchers and scholars in understanding the evolution of their field. What creates this need is the exponential growth of specialized knowledge, which is making it difficult to assess or measure the influence of scientific inquiry (Pinski & Narin, 1976). Although all the literature that has been published in the last couple of decades, and large portions of what have been published before this time frame, has become readily accessible via the Internet, the ability to develop structured overviews, identify highly influential articles, or examine peer reviewed scholastic content has become more difficult and time-consuming, due to the sheer quantity of search engines. Currently, there are several hundred search engines available to explore and research articles, in addition to the typical research methodology of simply probing peer reviewed journals. Some of these search engines are specialized for one or more subfields, but others are more general, like Google Scholar, Web of Science, and PubMed. These platforms are typically very useful in locating individual publications, but they offer limited insights into how the literature is organized. This lack of organization becomes an impediment for researchers, who must investigate a large quantity of publications in order to discern how different streams of literature relate to one another and determine how these streams develop over time. Undoubtedly, obtaining such an overview of the structure of the literature can be an extremely time-consuming process, especially in multidimensional scientific fields such

as statistical research, with publications appearing in different scientific fields. Hence, a deep understanding of the structure of the literature creates a firm foundation for advancing knowledge, consequently, facilitating theory development, closing areas where a plethora of research exists, and uncovering areas where research is needed. Existing methods of evaluating and understanding scientific research based on the citation counts, co-citation counts, or co-occurrence are not reliable. Usually, in academia measuring the quality of scientific research or evaluating scholars themselves is based on the total number of citations, h-index, or the impact factor of the journals he/she has published in. Nevertheless, these measurements are crude. Moreover, in some cases, these cumbersome methods are the means by which a hiring or firing decision is made (Pinski & Narin, 1976). For instance, David Adam describes the absurdity of Finland's government allocating funds to university hospitals by considering the citation counts and the impact factors. This is especially unfortunate since the aforementioned measurements do not account for the quality of the research that is being used to make such critical decisions (Webster & Watson, 2002). Given this, the bibliometrics technique should be incorporated and utilized as both qualitative and quantitative method. My dissertation is divided into three sections: 1) the first task is a deep study and analysis of the statistical field; 2) The second task is proposing a new similarity measure between Keywords in Co-word Analysis; and 3) The third task is to propose a new method for calculating the weight of keywords in clustering analysis in bibliometrics data.

1.2 Significance

The rapid growth of the scientific literature makes it difficult for researchers or scholars to have a comprehensive, and holistic grasp understanding of any given topic of interest. For example, in 2006, 1,346,000 articles were published in 23,750 journals, and the average annual growth of publishing is around 2.5%. These publications have created new subfields of science, which add to the scientific information on a whole (Bjork et al., 2009). Furthermore, in 2009, the STM (International Association of Scientific, Technical Medical Publishers)

report estimated that 1.8 million articles are published in 28,000 journals each year (Ware & Mabe, 2009). Similarly, a total of 50 million articles had been published by the end of 2009 (Jinha, 2010). Given the growing amount of publication, it will be a challenge for researchers, especially new researchers, to be able to capture the evolution of the field or to stay up-to-date. This volume of scientific literature, which is available on the Internet increases the need for new methods and techniques to expedite and facilitate the process of perceiving the overview of a scientific topic, with respect to the huge volume and the visual representation of data. At the present time, the bibliometrics mapping of science is a quantitative method for studying the bibliographic data (titles, keywords, authors, etc.) and visually representing the information. Furthermore, bibliometrics maps are useful for dealing with a large body of literature and have been used in different contexts, such as survey of research literature, government decision making, and scientific publications. Bibliometrics mapping should also include more qualitative evaluation instead of primarily focusing on quantitative assessment. In light of this, I will propose methods that will expand the field of study by employing bibliometrics qualitative data to generate bibliometrics qualitative information. By considering methods analogous to co-occurrence counting, I hope to make a contribution toward the improvement of the new visual representation, which will help researchers and scholars to more fully grasp the breadth and depth of scientific research, within a shorter period of time. Computerized methods and tools, including data mining, statistical visualization techniques and intelligent text mining, have been used effectively to facilitate and maximize the efficacy of assessing the quality of scientific research and individual works in this papers.

1.3 Research Objective

The research objectives are to

- (A) Map and visualize the literature of the statistics field, identifying emerging trends and understanding relationships between different developments in the field.

- (B) Introduce a new method to calculate the similarity between keywords in co-word analysis.
- (C) Introduce a qualitative method to calculate the weight of keywords in clustering analysis, within bibliometrics data.

The proposed methods will infuse qualitative components into bibliometrics analysis, which is, as of now, primarily a quantitative method. The previous methods only utilize quantitative approaches to build clustering, counting the co-occurrence between the keywords. These techniques are summative for the given literature at hand. However, they do not consider other connections between the keywords. This present discourse will focus on developing a new way to calculate the similarity measure between keywords and will visualize the weight of the keywords. The new method will increase the importance of using keywords in scientific literature. The proposed method intend to reignite interest in bibliometrics mapping science.

1.4 Dissertation Layout

This dissertation is organized as follows: chapter 2 is an extensive study of the statistic field, where a deep data mining, visualizations, and bibliometrics method applied to analysis the bibliographic data to revel insightful information and the evolving on the field since 1990s. Chapter 3 provides an ingrained study of similarity measures in the co-word analysis and presenting a new similarity method to the co-occurrence data in the co-word analysis. Chapter 4 proposed a new way to visualize the unit of analysis in the co-word analysis, the proposed method gives distinctly presentations of unit of analysis on the 2D visualization that is more accurate to the real importance of the unit of analysis than the current way. Chapter 5, in this concluding chapter a summarize of the contributions and the possible impact of this dissertation, and discuss of the limitations and the direction of future work.

Chapter 2

Bibliometric Mapping of (ASA) Journals from 1991 to 2016

2.1 Introduction

The size and scope of the literature on statistics can be staggering, making it difficult to identify emerging trends and relationships. Visualization techniques, coupled with statistical and data mining methods, have aided in understanding trends and relationships in several fields, including healthcare and manufacturing research (Han et al., 2011). In this paper, I applied these concepts to the field of statistical sciences. Our data-set is based on bibliographic information, including but not limited to authors, keywords, abstracts, citations, and funding information, extracted from 10,618 papers published in the 17 journals of the American Statistical Association (ASA) in the period of 1991-2016. These bibliographic units of analyses allow us to address the following questions: a) What are the main research fields within statistics (based on a data-driven approach)? b) How do these research fields relate to each other? c) How did these fields develop over the time period of 1991-2016? d) What are the main drivers for these publications? e) What did the top research institutions in the United States contribute to the field of statistics? f) How did these institutional contributions change over the time period of 1991-2016? By analyzing the massive amounts of literature that has been published and cited by ASA papers over the past twenty six years, we can glean various insights into the field via visualization.

2.2 Motivation

The motivation behind this paper lies in a really simple question: How can we capture the evolution of the field of statistics, and its subfields, over the past twenty six years?

Statistics is a field that involves the study of the collection, analysis, interpretation and presentation of data. Therefore, the field of statistics is, by nature, multidisciplinary and fragmented. With the ever-increasing volume and variety of data being collected, the discipline has evolved to the point where identifying the dominant streams of research, much less the incremental contributions within those streams, is becoming more and more difficult. Accordingly, it is challenging to maintain a holistic sense of the field's evolution, and as statistical applications in other fields continue to grow, the challenge will become increasingly difficult. To address these problems, we will use several statistical methods that are commonly adopted in bibliometrics analysis for the purposes of distinguishing the broad streams of research within statistical sciences, and we will highlight the important developments within the subfields since the beginning of the 1990s.

2.3 Background

While the Internet has made statistical literature fully accessible, the ability to develop structured overviews of the literature has become more difficult and time-consuming (Rodrigues et al., 2014; Ji et al., 2016). For example, researchers in statistics typically resort to Google Scholar, Web of Science and/or PubMed as their search engine of choice for locating relevant publications based on a set of keywords. These platforms are typically very useful in locating individual publications, but they offer limited insights into how the literature is organized. Researchers are required to sort through large numbers of publications to understand how different streams of literature relate to each other and how these streams change with time. Obtaining such an overview of the structure of the literature can be an extremely time-consuming process, especially in statistical method research with publications appearing in many different scientific fields. A deep understanding of the structure of the literature “creates a firm foundation for advancing knowledge. It facilitates theory development, closes areas where a plethora of research exists, and uncovers areas where research is needed” (Webster & Watson, 2002). Box & Woodall (2012) provide several examples for innovations in

statistical science research, highlighting the impact of combining ideas from multiple areas of statistics.

Review papers that discuss specific research areas/applications within subfields of statistics can be helpful for understanding the literature. The structure of these reviews can be either based on qualitative and/or quantitative factors. Reviews that are primarily based on qualitative criteria rely heavily on the authors' experience in categorizing, presenting, relating and/or critiquing the different publications that they include in the review. The purpose is often to highlight current developments in a subfield of statistics ((Woodall & Montgomery, 1999, 2014)), an ancillary purpose also being to provide perspectives on the role of statistical methods in a new research area (Nair et al., 2000; Box & Woodall, 2012; Megahed & Jones-Farmer, 2015). While these papers provide insight into a particular topic or research stream, they are not suitable for addressing the motivating question, which is to learn from several thousand published papers. Quantitative factors for analysis are extracted from bibliographic data. Most of the discussion on using such data has been outside the purview of statistics literature, with a few exceptions. Citation counts have been used by Altman & Goodman (1994) to investigate the speed of transfer of new statistical methods into medical literature and to predict which methods would be of importance to future medical research. Stigler (1994) used counts to highlight patterns in citations in the *Journals of Probability and Statistics*. Ryan & Woodall (2005) used citation counts to highlight and summarize the contributions of the top 25 cited papers in statistics. More recently, Baccini et al. (2009) extracted the names of the editors of 79 *Probability and Statistics Journals*, and they used the number of common editors as an input for network analysis on the journals.

We will explore literature published by the American Statistical Association (ASA) by making use of several data-driven bibliographic methods to help understand how the field of statistics has been evolving since 1991. By utilizing a number of quantitative measures, this study employs bibliographic methods, to delineate the main fields of research within statistics, to understand their evolution over time, to identify the key drivers for the evolution,

and to provide visualizations of the findings for dissemination to the statistics community. The details of the methods are provided in Section 2.4.

2.4 Methods

As explained in Börner et al. (2003), the process for visualizing/analyzing the literature can be divided into the following sequential steps: (1) data extraction, (2) definition of unit of analysis, (3) selection of measures, and (4) data visualization and analytics. It should be noted that these steps rely heavily on methods from the fields of text mining, bibliometrics and information visualization (Rodrigues et al., 2014). Text mining provides natural language processing methods that allow for extracting the bibliographic data from the publications. The bibliographic literature presents the background needed for identifying the measures and for representing the relationships between them using graphs and networks. Finally, the information visualization literature provides the details needed for mapping and visualizing the data.

In this section, we describe how the data has been collected and preprocessed; then we discuss the units of analysis and how they help us answer the motivating questions, followed by an overview of the measures used for each of the selected units of analyses. Lastly, we discuss the methods for analyzing statistics literature at the term (keyword), document, and journal level (Note that the journal level includes proceedings, and/or book titles), these methods also examine authors, and research addresses to study the spatiotemporal, aspect of the literature, and in the last subsection, we explain how we extend the analysis for each of the levels to reflect patterns over time.

2.4.1 Data Extraction and Cleaning

This study uses data from the Web of Science (WoS) database. The choice of WoS was based on four factors. First, WoS provides the “world's largest collection of research data, books, journals, proceedings, publications and patents” (*WEB OF SCIENCE @ONLINE*,

2016). Second, the 17 journals of the American Statistical Association, ASA, are covered on the Web of Science Database (see <http://www.amstat.org/publications/journals.cfm>). Third, WoS has been used in previous papers in the bibliographic analysis of statistical (Ryan & Woodall, 2005). The fourth, and the most important reason, is that WoS provides access to a *full record* of the 10,618 papers. The *full record* includes, but is not limited to, the *abstract, authors' names, keywords, references cited, publication date, research address, and publications citing this record*. It should be noted that obtaining all this information may not be possible with other databases. For example, PubMed does not provide reference data, which is crucial to our analysis.

The bibliographic analyses in the subsequent steps can be as good as the data on which they are based (Cobo et al., 2011). This concept is well understood in statistical practice. While data quality is multidimensional (Wang & Strong, 1996; Jones-Farmer et al., 2014), we focus on data consistency. Ballou & Pazer (1985) posited that consistency occurs when the “representation of the data value is the same in all cases.” An example of an inconsistent representation would be the use of the terms “control-chart”, “control chart”, and “control charts” as keywords by three different journal papers. When preprocessing the data, these three representations should be merged/combined. This could be easily done using software packages that include standard text-mining procedures. We do not consider other important data quality dimensions, such as accuracy and completeness, because these are difficult to measure with bibliographic data. For example, one cannot assess whether a set of keywords provides the most complete/accurate representation of the paper or clarifies that papers written by “John Doe” are in fact written by four different authors.

2.4.2 Units of Analyses

The second step in our approach is to select the units of analysis. In this study, we have chosen five units of analyses: original keywords, documents (articles, books, publications or manuscripts), journals, authors, and research address. Each unit depicts different facets

of the statistics domain and facilitates different types of analysis. The examination of the keywords can deepen one's understanding of the cognitive structure of a field (Bhattacharya & Basu, 1998; He, 1999; Cahlik, 2000). This is typically done through co-word analysis, which is a content analysis technique that uses patterns of co-occurrence of pairs of words, or noun phrases, to identify the relationships between ideas within the subjects presented in the papers containing the keywords (He, 1999). Indexes based on the frequency of co-occurrence of the terms are then used to measure the strength of relationships between items. Based on these indexes, the keywords are clustered in groups and displayed in network maps. Therefore, the examination of the keywords allows us to identify the research fields within statistics and quantify the amount of overlap/separation between these fields. When we examine the keywords over time, we can understand how the research areas are evolving and building on materials from other subdomains in statistics.

Documents are the most common unit of analysis used in mapping and visualizing a knowledge domain (Börner et al., 2003). Maps based on documents have been used for a variety of purposes, including domain analysis (Rodrigues et al., 2014) and assessing research performance (Bornmann et al., 2014). The approach for analyzing the documents is based on shared citations (see details in 2.4.3). These pairings of shared citations are clustered and are displayed in network maps as with the keywords. The difference, however, is that the pool of documents is not only based on the 10,618 publications, it also includes all cited documents (journal papers, conference papers, books, and the like) within these publications as obtained from the full records. Accordingly, such an analysis can help assess how influential documents drive statistical research. Since this is based on network analysis, one can learn how a certain influential paper drives research in other sub-disciplines and how this impact changes over time. By focusing on the interconnection between published papers and their citations. This present investigation can be seen as an extension to the work of (Ryan & Woodall, 2005).

A map of journals can be used to obtain a macro view of science (Bassecoulard & Zitt, 1999), showing the relative positions and relationships between major disciplines. Journal

maps are also used on a more focused scale to show fine distinctions within a discipline (Ding et al., 2000). In this study, we aim to achieve the latter objective by constructing maps of how the different statistical journals focus on different areas of research. It should be noted that the sample of the journals is populated from the 17 journals of the ASA, as well as journals extracted from the cited references.

At the authorial level, we aimed to study the relevance among authors of the ASA papers. There are two kinds of relations among authors, direct and indirect. First, the direct relationship is represented by the co-authorship between two authors (Newman, 2004). Two authors are directly related if they collaborated on one or more papers. Second, the indirect relationship is represented by author co-citation (White & Griffith, 1981). Two authors are indirectly related if other papers have cited them together.

Finally, we aim to study the evolution of ASA research over time and space. Spatiotemporal data is divided into two kinds of data: the spatial data, which is synonymous with geographic data, and the temporal data, which comprises the time period of the data. In our study, spatial data was defined by an institution's name and location, while temporal data was defined by the year of publication.

2.4.3 Measures and Similarity Calculation

Measures for bibliographic analysis have been defined by White & McCain (1997, p. 103), Börner et al. (2003, p. 191-193). In this study, the threshold for choosing the minimum number of co- occurrences or the unit of analysis appearances is based on two criteria. The first criterion involves the data size and ease of visualization. For our analysis at the term level, we use the co-occurrence of keywords as the measure for their relatedness. This is calculated by counting the number of times a pair of keywords occur together in a document. A large value for the number of co-occurrences indicates a strong relationship between the pair, and vice versa. In an attempt to find the top 10 most used keywords in ASA papers

from 1991 to 2016, we divided the time period of analysis into five periods 1991-1995, 1996-2000, 2001-2005, 2006-2010 and 2011-2016. We counted keywords occurrences in a period and then ranked them based on their frequency.

To perform the analysis at the publication level, we assessed the relatedness of publications based on direct and indirect citation relations. Rodrigues et al. (2014, p. 2) define these relations as follows: “Two publications have a direct citation relation if one publication cites the other, and they have an indirect citation relation if they both cite the same publication (bibliographic coupling) or are both cited by the same publication (co-citation).” The weight of bibliographic coupling and co-citation relations is equal. Similar to Rodrigues et al. (2014), an artificial citation from each publication to itself is created. The use of the artificial citation allows direct citations to have a larger effect on the relatedness of the documents than indirect citation. This is because “a direct citation between two publications counts as both a bibliographic coupling relation and a co-citation relation” (Rodrigues et al., 2014, p. 2). Note that the usage of citations as a measure has the following consequences: a) the number of documents cited will be much larger than the original 10,618 publications; b) both the journal lists and the publication dates will differ from our original data-set; and c) the cited documents will no longer be limited to journal publications, since some of the citations will be conference proceedings, books, and the like. Here, we do not limit our analysis to the 1991-2016 period to allow us to understand and visualize the foundational papers for the ASA publications that appeared in the 1991-2016 time period.

For the analysis at the journal level, we use the same citation measures used for the publication analysis, with the exception that our unit of analysis is now a journal instead of a cited publication. Furthermore, for our analysis at the authorial level, we use the co-occurrence of collaborations between two authors as a measure of their relatedness within the author co-authorship analysis; we likewise use the co-occurrence of other papers citing them together as a measure of their relatedness at the author co-citation analysis. Co-authorship is a great measure of a collaboration between researchers in the same field (Melin & Persson,

1996). Maps based on co-authorship sought to study the collaboration between scientists in ASA papers, which were published within the 1991-2016 time frame. Researchers are connected if they collaborate on one or more documents. Also the document is considered co-authored if it has more than one author (Melin & Persson, 1996). The strength of collaboration between authors depends on the number of works they have produced together. In this study, we aimed to find the relationship between two authors, and we grouped them in clusters in order to make it easy to track their work and production. Similarly, co-citation measures the relatedness of two authors' work; when other papers cite one or more of their papers, we can say that the two authors are co-cited.

To perform a spatiotemporal analysis, we collect information from papers across time and space (Katz, 1994). For each paper, we count the total number of citations against the year and the research institution. After assessing the relatedness of each of the aforementioned units of analysis, we use a clustering approach to identify the related keywords, publications, journals, and authors. The methodology for clustering is based on the approach of (Waltman et al., 2010).

Usually in the bibliometrics area a combination of multidimensional scaling (MDS) and hierarchical clustering is used to map and cluster the items (unit of analysis). However, visualization of similarities (VOS) mapping and clustering technique uses an approach that differs from MDS (van Eck et al., 2010). In VOS, the distance between two items (unit of analysis) represents the relatedness or similarity between them, VOS determines the similarity between the units of analysis by transforming the co-occurrence frequencies using the similarity measure. S_{ij} represents the similarity between items i and j . The VOS uses the association strength of items to measure the similarity between items i and j . The association strength (S_{ij}) equation is given by:

$$S_{ij} = \frac{C_{ij}}{C_i C_j} \quad (2.1)$$

Then VOS locates items on the map by minimizing

$$V(x_1, \dots, x_n) = \sum_{i < j} S_{ij} \|X_i - X_j\|^2 \quad (2.2)$$

Subject to:

$$\frac{2}{n(n-1)} \sum_{i < j} \|X_i - X_j\| = 1 \quad (2.3)$$

Basically the idea behind the VOS approach is to minimize the weighted sum of the squared distance between all unit of analysis pairs.

2.4.4 Mapping and Visualization

Clustering provides a breakdown of statistics literature into a number of domains. Throughout this paper, we use the terms *domains*, *fields*, *areas* and *disciplines* interchangeably. At the keyword level, we construct a *term map* for all keywords that occurred in 10 or more publications. The relatedness of the terms and the grouping of terms into different clusters are based on the methodology described in 2.4.3. Terms are located in a 2D space, and the distance between the terms is a function of their relatedness. Additionally, the colors that we use for the term indicates the cluster that the term belongs to. The locations of the keywords on the map are determined using the VOS mapping technique (van Eck et al., 2010), and the software VOSviewer is then used to visualize the map (van Eck & Waltman, 2009). The threshold for the number of keywords to be included has been chosen to facilitate the analysis and make the visualization more informative (the larger the number of terms, the more the overlap between nodes since the 2D space for projection is limited). Also, at the keywords level, we have charted the top 10 most used keywords for each period (1991-1995, 1996-2000, 2001-2005, 2006-2010, and 2011-2016). We used the keywords frequency and keyness. Keyness represents the importance of keywords by using the log-likelihood calculation methods Biber et al. (2007) between the two unequaled size sets.

At the publication level, we construct a citation map to obtain an overview of statistics literature. Similar to the term map, a citation map provides a 2D representation of the literature, where cited documents are located based on the VOS mapping technique. The shorter the distance between the cited publications, the stronger the relationship, that is the combined effect of direct and indirect citations. We also construct a citation cluster map to provide a more high-level overview of statistics literature. Both maps are based on the 412 most frequently cited publications in statistics. Within these publications, the least cited and most cited papers have been cited 35 and 411 times, respectively. (Recall that these citations are not limited to ASA journals and/or the 1991-2016 period of analysis, for the reasons explained in Section 2.4.3)

Using the citations within the 10,618 publications in the ASA journals, we develop a *journals map* and a *journals cluster map* for visualizing the relatedness of journals, proceedings, and/or books that have been cited in these publications. Hereafter, we use the term *period:journal* to refer to any of these three types of sources cited. Both maps are based on the 707 most frequently cited journals in statistics during 1991-2016. The least cited and most cited journals have been cited 30 and 19,997 times, respectively.

At the authorial level, we have two maps. The first map is the author co-authorship map. We construct a co-authorship map to obtain an overview of the collaboration between authors in statistics literature. Similar to the previous maps, the co-authorship map provides a 2D representation of the researchers and their location, based on the VOS mapping technique. The co-authorship map is based on the 214 most produced authors in statistics. Among these authors, the least productive author produced 10 papers, while the most productive author produced 64 papers.

The second map is the author co-citation map, which features 2D representations of the resemblance of authors' work. The map is based on the 795 most cited authors in statistics. Within these authors, the least cited and the most cited authors have been cited 50 and 1,255 times, respectively.

Also, in bibliometrics area, there have been several studies that seek to study the diffusion of research, co-citation, and spatiotemporal change among research institutions (Börner et al., 2006; Katz, 1994). We perform a spatiotemporal analysis on the ASA data to identify the contribution and the production of research among research institutions in the world. The spatiotemporal analysis represents the time of publication and the location of the research institution. The spatiotemporal map gives an overview of the contributions of research institutions in statistics literature from 1991-2016. We value the contribution by the number of citations the institutions have received. The map is based on the 1417 most contributing research institutions in the world. Within these institutions, the least cited institution received 0 citations, and the most cited institution received 17,867 citations. When a research institution has contributed but its work has not been cited, we considered its contribution to be zero. This type of analysis provides an even higher level of analysis from that of the publication level. In fact, our intention is to provide multi-levels of analysis from the very specific (term-level) to the very high-level (journal), alongside the spatiotemporal analysis to capture the overview of statistics literature.

2.4.5 Analysis of Developments Over Time

To identify how the field of statistics evolved over the aforementioned time period, crates for each of our units of analysis are calculated for five time periods, 1991-1995, 1996-2000, 2001-2005, 2006-2010 and 2011-2016. At the term level, we use the following *term score* as a measure of whether the keyword has been used more frequently in the earlier or later time period:

$$ts_i = \frac{\sum_{j=1}^{26} c_j \times m_{i,j}}{m_i} \quad (2.4)$$

ts_i is the *term score* for a given keyword i and j represents the year (1 for the year 1991 and 26 for the year 2016). The indicator variable C_j is set to 0 for 1991-1995, 0.5 for 1996-2000, 1 for 2001-2005, 1.5 for 2006-2010 and 2 for the last six years 2011-2016. $m_{i,j}$ is the number of publications where a keyword i occurs in year j . m_i represents the total number of times a keyword i occurs in our data-set. Based on this score, it can be easily seen that ts_i can take any value in the interval $[0, 2]$. A score of 0 indicates that all occurrences of this keyword happened in the 1991-1995 period. Similarly, a value of 2 indicates that all occurrences happened in the 2011-2016 period, and a value of 1 indicates that the occurrences have been distributed equally among the five time periods. Based on these three example scores and/or equation 2.4, one can easily deduce that the term score represents the percentage of occurrences in which a particular keyword has been used during the time period 2011-2016. We use these scores to develop a *temporal term* map. It should be noted that this analysis is somewhat similar to the analysis performed by (Rodrigues et al., 2014). We use a different scoring coefficient, however, to facilitate the interpretation. At the publication level, we calculate the citation rates for the five time periods. For each cluster of publications and time period, we calculate the *citation cluster score*, as follows:

$$cs_k = \frac{\sum_{j=1}^{26} c_j \times n_{k,j}}{n_k} \quad (2.5)$$

cs_k is the *citation score* for publication cluster k , j represents the year, and C_j is the indicator function used in equation 2.4. n_k represents the total number of publications cited in cluster k , and $n_{k,j}$ is the number of publications cited in year j in cluster k . The interpretation of the *citation cluster score* is similar to that of the *term score*, meaning, 0 indicates that all publications within that cluster have been cited during 1991-1995. We present the results for this in a *temporal cited publication cluster map* and in a tabulated format.

For the journal level, we repeat the analysis (of the publications) but with the *cited journal* as the unit of analysis. We calculate the *cited journal score* (js_k) using the same logic as that of the citation cluster score. For the sake of conciseness, we do not provide the

equation for how we calculate the s_k . The results are presented in a *temporal (cited) journal map*.

2.5 Results

2.5.1 Keyword Level: Term and Temporal Term Maps

There are 576 out of 17,769 original keywords that have met the criterion of appearing in at least 10 publications. The term map, shown in Figure 2.1, depicts ten clusters of terms that are used in statistics literature. The clusters cover the following topics: (1) Reliability (red); (2) Density Estimation (chocolate); (3) Computer Experiments and Geostatistics (green); (4) Expectation Maximization (purple); (5) data quality (blue); (6) Time Series (yellow); (7) Design of Experiment (pink); (8) Data Science (brown); (9) Statistical Process Control (orange); and (10) Dimension Reduction (dark green). The number of terms per cluster is 102, 88, 78, 70, 58, 58, 54, 46, 12 and 10, respectively. In addition, the size of the text is an indication of the relative frequency of use within the 10,618 ASA journal papers. The most frequently used terms in Figure 2.1 are Markov Chain Monte Carlo, R, and kernel with 440, 349, and 280 papers listing them as keywords, respectively. Note that the spatial orientation is a function of the relatedness of any two terms. Therefore, e.g., Nonparametric Regression and Nonparametric Density Estimation are very close even though they belong to separate clusters. It should also be noted that this figure does not depict all 576 keywords since there are some terms that spatially overlap. For example, the term Gibbs Sampler does not appear in the figure since it overlaps with the term Markov Chain Monte Carlo. In the supplementary materials, we provide a table containing all keywords and their clusters for the readers' reference.

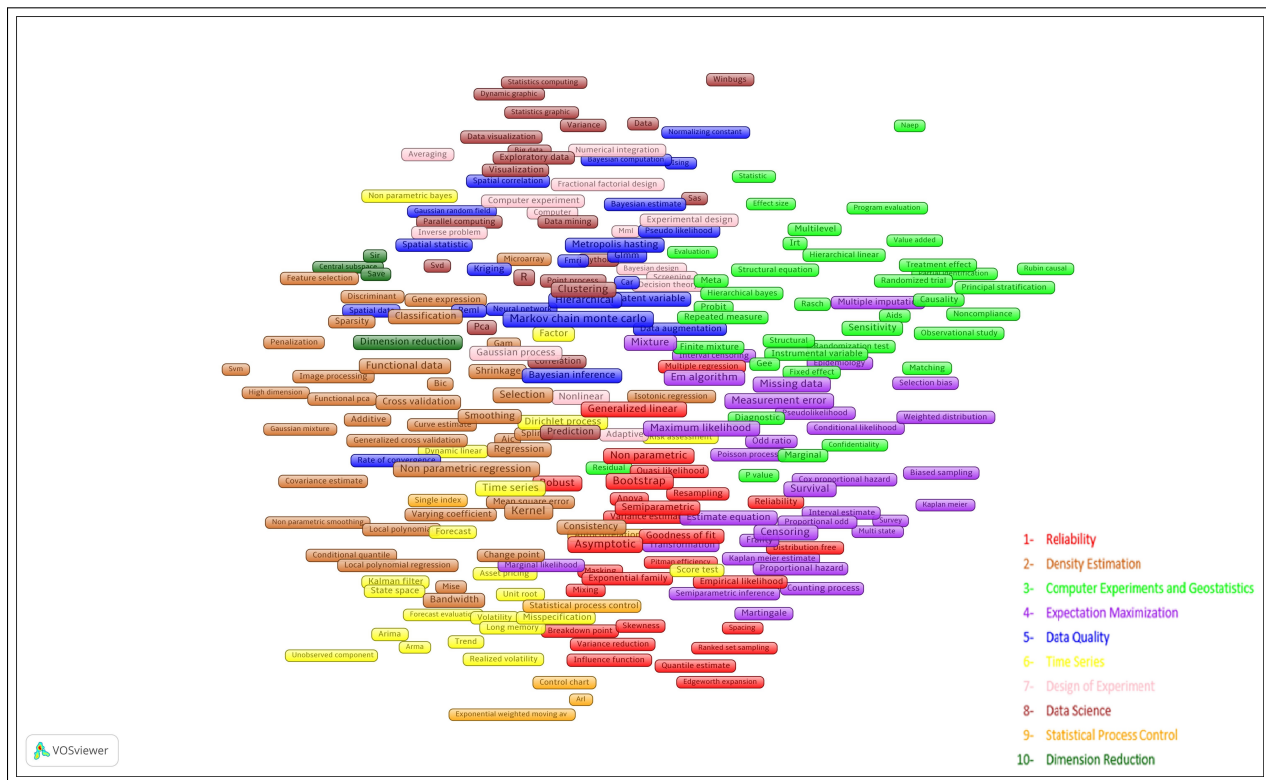


Figure 2.1: Term map with colors indicating ten clusters of terms.

Figure 2.2 shows the same map as in Figure 2.1, with the exception that the color of the term now indicates the probability of a term's usage during 1991-2016. This figure shows an increasing rate in the use (or design of statistical modules) in *R* ($ts = 1.85$), *Matlab* ($ts = 1.81$), *Graphical User Interface* ($ts = 1.54$), *Lasso* ($ts = 1.72$), *Adaptive Lasso* ($ts = 1.85$), and *Oracle Property* ($ts = 1.82$). The first three keywords reflect the increasing role of software and data visualization in the field; for example, the ASA has started sponsoring the *Journal of Statistical Software* as of 2006. While Lasso has been introduced by (Tibshirani, 1996), the majority of Lasso citations have occurred since 2006. A quick Google Scholar search indicates that 72% of *Lasso* citations happened during the 2006-2016 time period while 31% of the citations happened during 1991-2005. Additionally, (Zou, 2006) wrote his seminal JASA paper “The Adaptive Lasso and Its Oracle Properties” in 2006. It is therefore understandable why *Adaptive Lasso*, and the *Oracle Property* have only appeared after 2006. On the contrary, publications involving the keywords *Maximum Likelihood* ($ts =$

1.00) and *Hierarchical Bayes* ($ts = 0.6$) have been decreasing in ASA journals. We believe that the rationale for the decrease is not due to an actual decrease in these research areas but is due to the field's utilization other terms instead. For example, Hierarchical Bayes has been replaced by *Bayesian Hierarchical Model* ($ts = 1.52$). Publications on more well-established topics such as *MCMC*, *Shrinkage*, and *Statistical Process Control* are distributed fairly equally among the four time periods. We provide a table containing all the ts scores for all 576 keywords in the supplementary document.

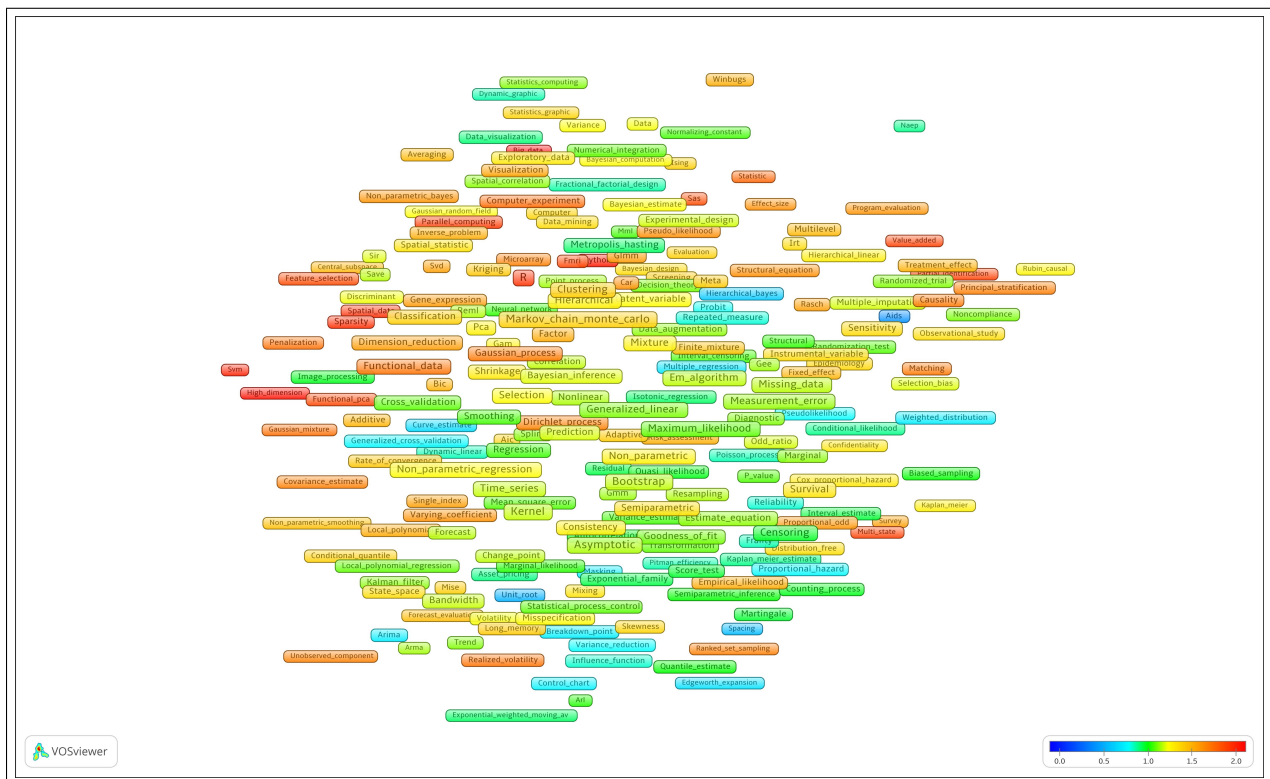


Figure 2.2: Temporal term map. Color indicates the probability of term use during 1991-2016.

To obtain a better understanding of how a certain group of keywords has evolved, we provide a boxplot of the *term scores* for each of the 10 clusters in Figure 2.3. The mean *term scores* for each of the ten clusters are as follows: 1.080, 1.382, 1.246, 1.118, 1.287, 1.202, 1.260, 1.400, 1.190, and 1.170. Recall that each of these scores indicates the percentage of usage during 1991-2016. A one-factor ANOVA on the clusters suggest that the null hypothesis of

equal treatment means should be rejected ($p = 0.000$). Accordingly, we perform Tukey's honest significant difference (HSD) test to consider all possible pairwise differences of means at the same time. The Tukey's pairwise comparisons grouped the clusters into four groups, which are groups A, B, C and D. Group A contains clusters 2,3,5,7,8,9, and 10. Group B contains clusters 3,5,6,7,9, and 10. Group C contains clusters 3,4,6,7,9, and 10. Group D contains clusters 1,4,6,9, and 10. The results indicate that the clusters that do not share a letter are significantly different. For example, clusters 2, and 8 are significantly different from clusters 1 and 4.

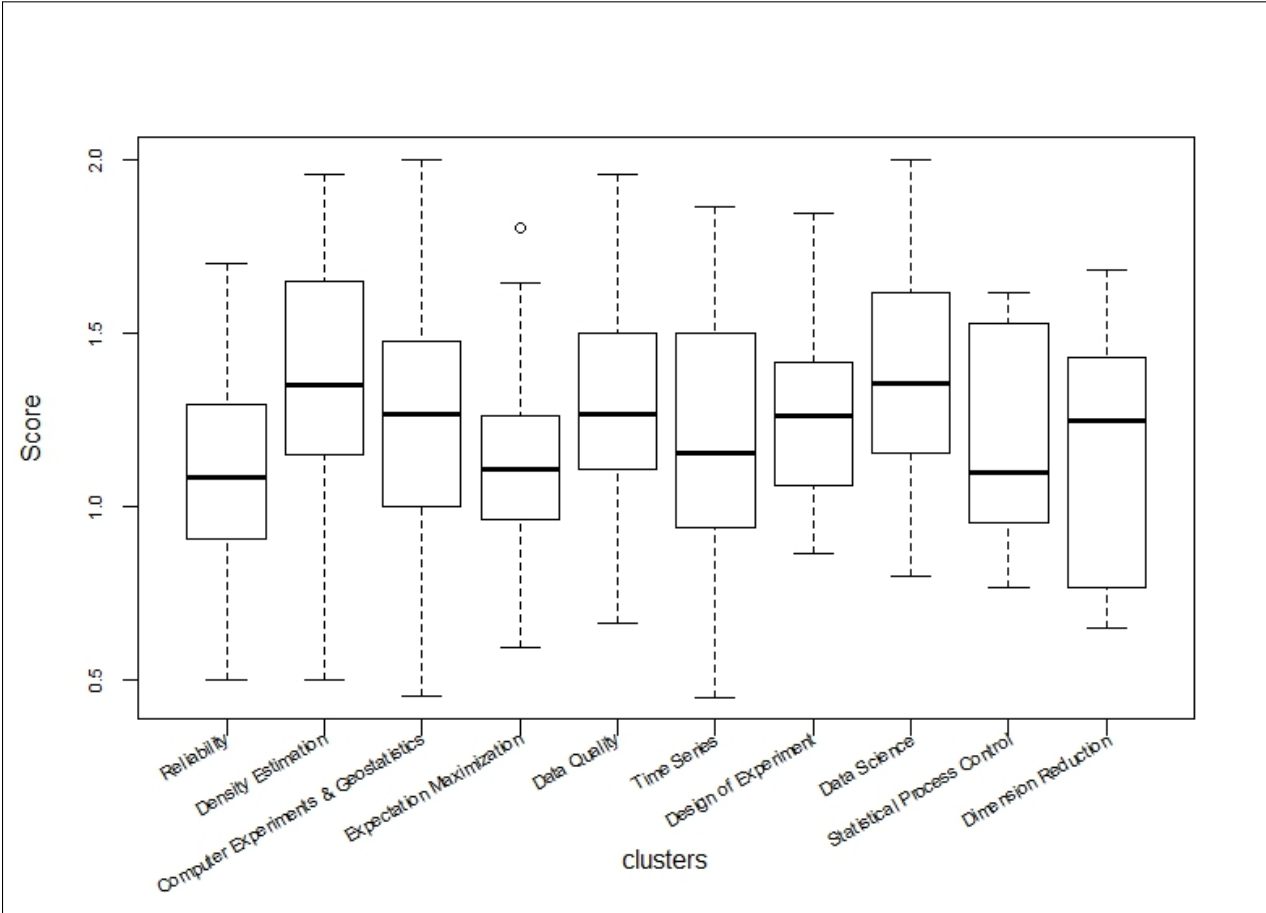


Figure 2.3: A Boxplot of the term scores for each cluster.

In Figure 2.4, we have charted the top 20 most used keywords for each time period: 1991-1995, 1996-2000, 2001-2005, 2006-2010 and 2011-2016. The data come from the original keywords in ASA journals. We have deleted words such as “methods”, “techniques”

and “models” because they do not provide any additional information regarding the most important words. The first column shows the time period, and the second column shows the top 10 words that appeared in the original keywords. The third column shows the maximum frequency for each keyword. The fourth column is the chart for the maximum frequency for each keyword. The fifth column represents the “keyness,” which indicates the importance of the keyword in the period when compared to the whole set of the keywords, and it is particularly useful when the two sets are unequal in size, like in this case where we compare the frequency of the word in the period and the whole set from 1991-2016. As shown in the chart below, the keyness value for each word will differ based on its appearance in the time period. For example, the keyword “Markov Chain Monte Carlo” is the top keyword in the period 2001-2005 with a keyness of 11.83, but in 2006-2010 it is also first. However, its keyness is 8.19. This indicates that the keyword “Markov Chain Monte Carlo” was more important in the period 2001-2005 than the period 2006-2010. In period 2011-2016 the “Markov Chain Monte Carlo” is 2.11. By looking at the MCMC keyness value we can state that the 2001-2005 period was the most effective era for research on the MCMC within ASA literature, we extract the chart for “Markov Chain Monte Carlo” in 2.5 . The sixth column shows the chart for keyness for all the keywords. Then, the last column shows the term frequency for the keywords in each period. For interactive visualization please see <http://www.viziolation.com/asa.html>













Period	Keywords	Keyness	Maximum Frequency	By Year
1996-2000	markov chain monte carlo	1.11% 	0.0100 	
2001-2005	markov chain monte carlo	11.83% 	0.0100 	
2006-2010	markov chain monte carlo	8.19% 	0.0100 	
2011-2016	markov chain monte carlo	2.12% 	0.0100 	

Figure 2.5: Markov Chain Monte Carlo Keyness values.

2.5.2 Publication Level: Cited Publication and Temporal Publication Maps

The total number of publications that have been cited by the 10,618 ASA papers is 136,762 individual publications. Since we focused on publications that have been cited at least 35 times, 412 documents meet this criterion. We use these documents to create the *cited publication map* in Figure 2.6. The map indicates the relationship between highly cited publications and shows how these publications cluster together. These publications are grouped in 7 clusters, as indicated by the different colors depicted in Figure 2.6. As explained in the *terms map*, we cannot show the entire 412 cited publications. It should also be noted that, as expected, some of these publications outdate the 1991-2016 time period. These are either influential books or seminal papers, like P.J. Huber's “*Robust Estimation of a Location Parameter*,” published in the *Annals of Mathematical Statistics* in 1964.

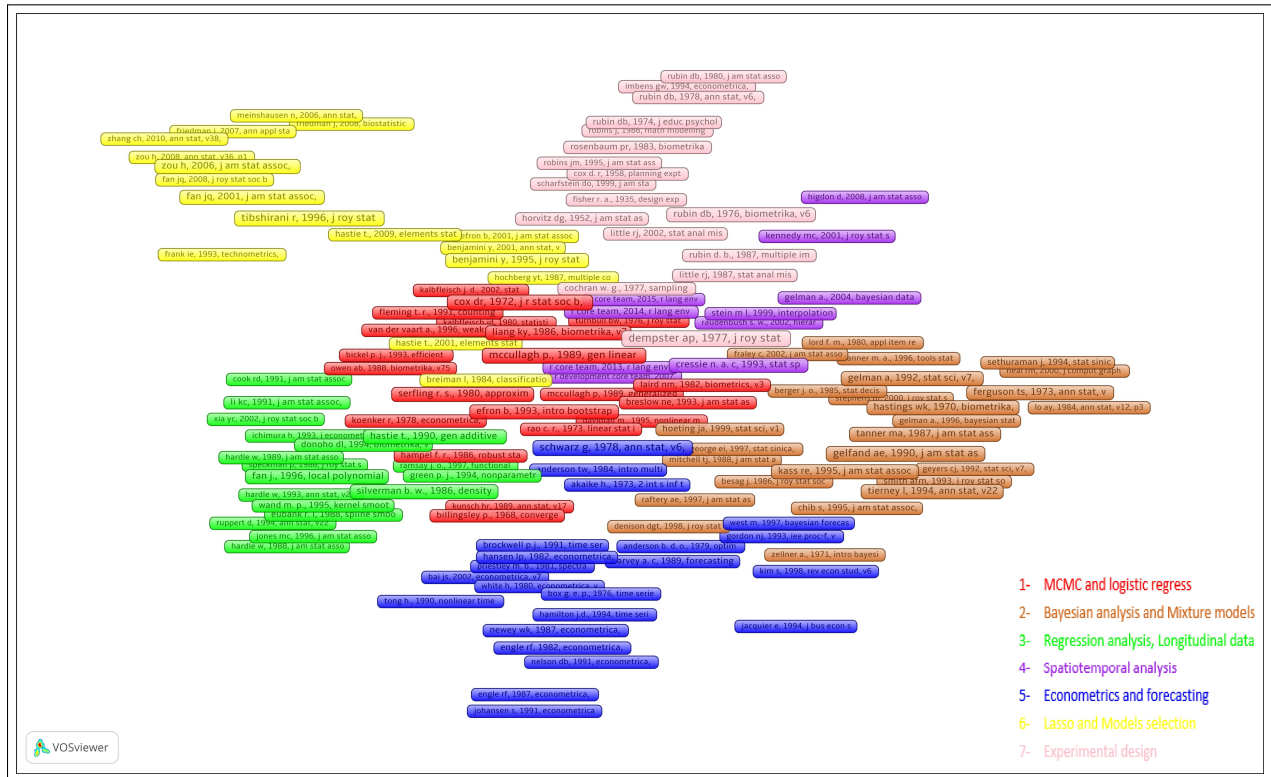


Figure 2.6: Cited publication map with colors indicating eight clusters.

We have examined these clusters manually to assign an appropriate label to each one of them. The labels and descriptions of the content of these clusters are provided in Table 2.1. The reader should note that some of the topics overlap among the clusters. This is not surprising, as we have depicted this in Figure 2.5. We also provide the mean *cited publication score* for each of the clusters in Table 2.1. Conclusions similar to the temporal analysis of the keywords can be made, where there are an increasing number of citations for Lasso, and statistical graphics/R. In general, the citation rate for most clusters is increasing. This is probably due to the increase in the number of journals.

Table 2.1: Clusters of cited publications.

cluster	Main Topics	Number of publication in cluster	Number of citations					sc_k
			1991-1995	1996-2000	2001-2005	2006-2010	2011-2016	
1	MCMC and logistic regress	94	892	1344	1340	1313	1259	1.06
2	Bayesian analysis and Mixture models	90	515	1295	1618	1399	1416	1.15
3	Regression analysis and Longitudinal data	75	377	813	1036	1198	1165	1.22
4	Spatiotemporal analysis	45	35	88	241	897	1529	1.68
5	Econometrics and forecasting	39	382	548	484	511	600	1.08
6	Lasso and Models selection	37	26	125	221	711	1588	1.7
7	Experimental design	33	238	406	512	510	732	1.23

In Table 2.1 , we have highlighted how the clusters of cited publications vary over the five time periods of the analysis. Figure 2.7 provides an in-depth look at how some of the most cited publications vary over time using the *citation score* of equation 2.5. One can observe two expected results: a) publications published after 2009 will have a *citation score* of 2, and b) publications published in the first period will have higher values of their *citation score*.

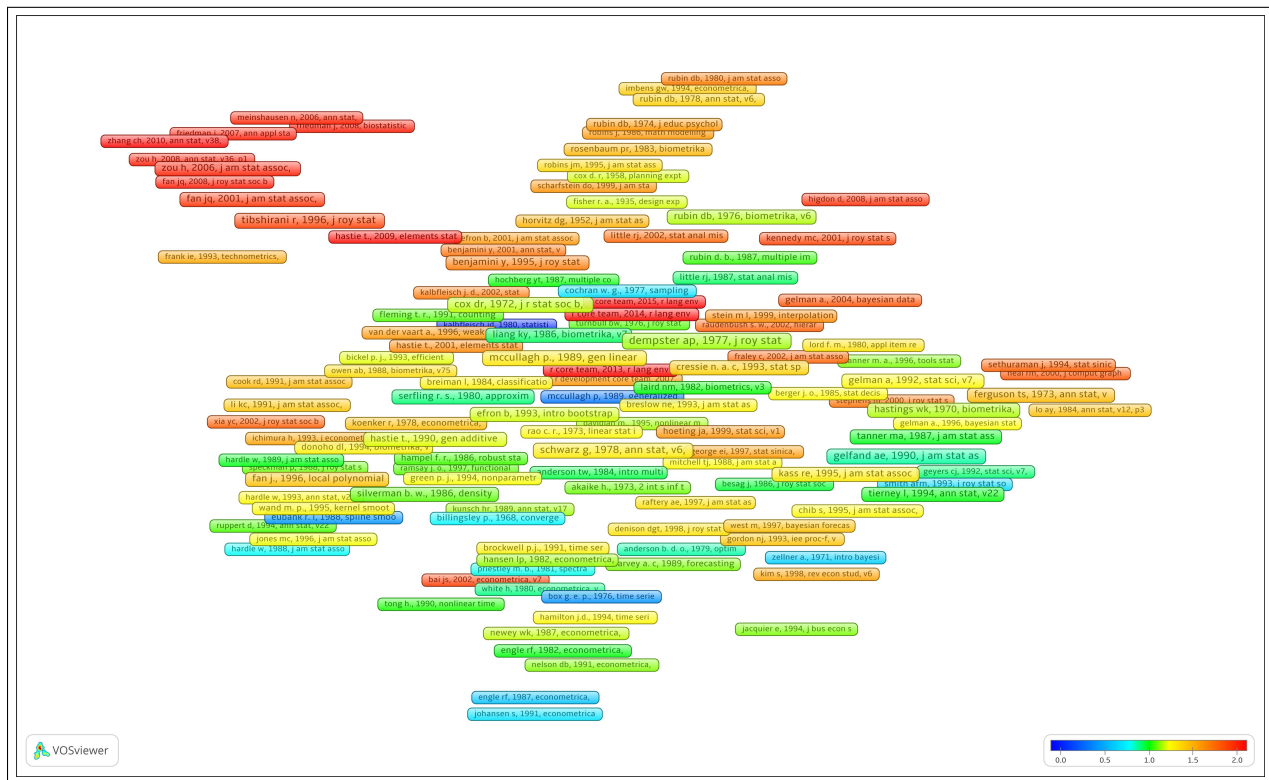


Figure 2.7: Temporal cited publication map, Color indicates the use of citation 1991-2016

More interesting observations can be made by investigating some of the earlier citations. There are several papers whose citation rate has not decreased in the past twenty six years, even though they have been published several decades ago. These papers are primarily colored in yellow and green. Huber's 1964 paper serving as an example here. Based on our data-set, Huber's paper has been cited two times during the time period of 1991-1995, cited 6 times during the time period 1996-2000, cited 6 times during the time period 2001- 2005, cited 9 times during the time period 2006-2010, and cited 16 times during the time period

2011-2016. It should be noted that these counts are only based on the citations within the original 10,618 publications, and their citations, namely, the 136,762 documents that we have investigated here. These results can offer only a sample of how this paper has been cited. By further investigating this result, we can confirm that the observation made from the graph reflects a growing interest in this paper by our community. More specifically, the WoS indicates that this paper has been cited by 1833 publications. The top six years in terms of the number of records cited are: 2016, 2015, 2014, 2013, 2011, and 2012, respectively. These six years account for 644 of the citations as shown in Figure 2.8. Therefore, 46.33 % of the citations for the Huber's paper have occurred in the 2011-2016 time period. The usefulness of Figure 2.7 lies in its ability to offer statisticians an opportunity for generating hypotheses about the influence of certain publications. These hypotheses can then be rejected (or not) based on a more detailed investigation.

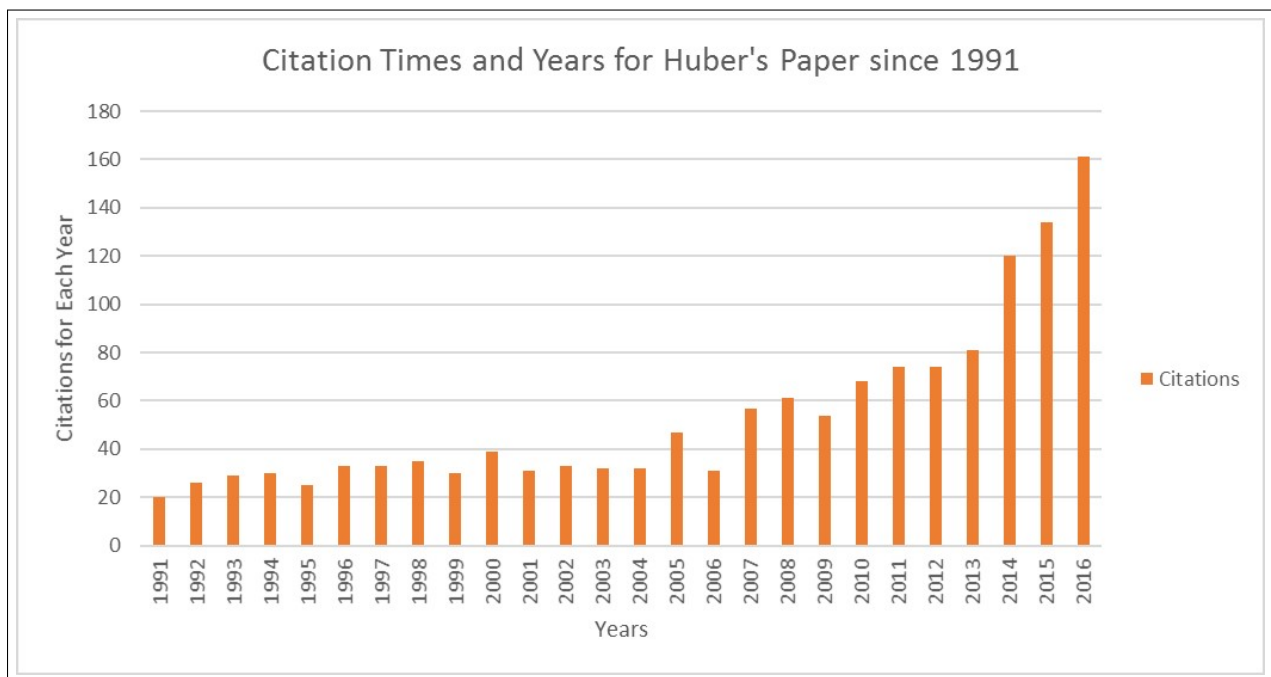


Figure 2.8: Citation counts for Huber's paper since 1991. Data retrieved from WOS website.

2.5.3 Journal Level: Journal and Temporal Journal Maps

Based on the mapping and clustering approach of van Eck et al. (2010), we obtain six clusters of cited journals. Figure 2.9 shows the six clusters. It can be easily seen that the clusters cover the following topics: (1) computer science (red); (2) medical with a focus on surveying and design of experiments (brown); (3) applied statistics and statistical quality control (green); (4) econometrics and time-series analysis (purple); (5) theoretical statistics (blue); and (6) data mining (yellow). The number of journals per cluster is: 240, 158, 128, 95, 84, and 2, respectively. The top 10 cited journals are *JASA*, *Annals of Statistics*, *Biometrika*, *Biometrics*, *Econometrica*, *Technometrics*, *JRSS Series B met*, *Journal of Econometrics*, *JRSS Series B*, and *Statistics in Medicine*. While our original data-set is only based on the ASA journals, only two ASA journals appeared in the top 10 cited sources. Accordingly, we believe that the effect of the original sample on our analysis is somewhat limited and that conclusions from this figure may be generalized to the field of *statistics*.

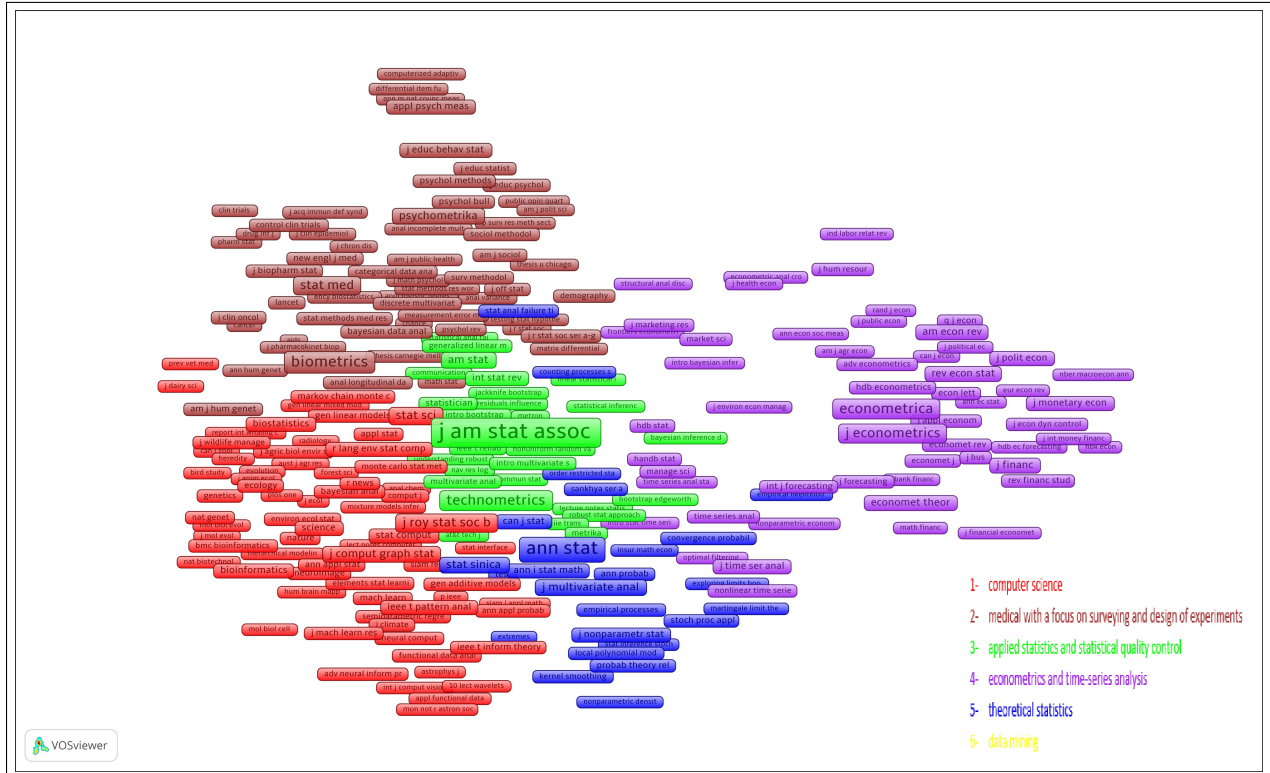


Figure 2.9: Cited journal map.

To examine the indirect relationship between two authors, we performed the author co-citation analysis (ACA). ACA shows the relationship strength between two authors by counting the times the two authors have been cited together by other papers (White & McCain, 1998). The total number of authors who have been cited is 54,200. We focused on authors who have been cited at least 50 times, we chose threshold of 50 to expedite visualization. We had 795 authors clustered in 6 clusters, shown in Figure 2.12. The number of authors per cluster is: 175,155,136,135,106, and 88, respectively.

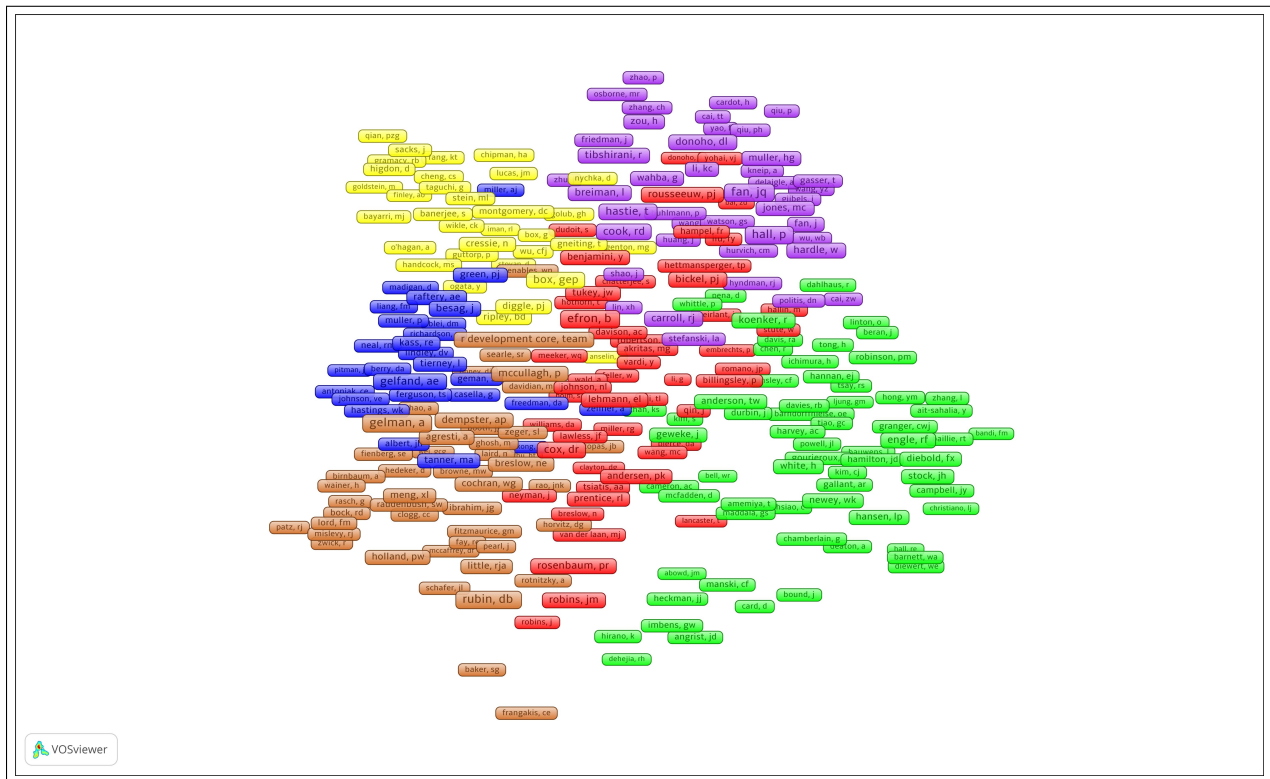


Figure 2.12: Author co-citation.

By means of these the two analyses, we can figure the direct and the indirect relationships between two authors. For example, Hall, p and Fan, jq have a strong direct relationship because they share the same cluster in the co-authorship map, and they share the same cluster in the co-citation map. Also, we can say that Efron, b and Rosenbaum, pr have an indirect relationship since they share the same cluster in the co-citation map, but they do

not have a direct relationship since they fall into two different clusters within the realm of co-authorship.

2.5.5 The Spatiotemporal Analysis

ASA publications from 1991 to 2016 were analyzed to identify the spatiotemporal changes and the most contributing institutions among the research institutions in the World. The data-set included 1417 research institutions in 10,618 papers, papers that have been produced by 11,995 authors. In this paper, institutions include universities, colleges, research labs, corporation labs, and all other academic institutions. We preprocessed the data to clean for our analysis. We included non-US institutions from our analysis to better visualization. Furthermore, we examined each author and research institution listed on each paper, and if the authors of a paper are from the same institution, then we count the citations for their institution. However, if the paper was co-produced by more than one institution, then we count the citations for each institution. We decided to merge each institution that has more than one campus or different departments into one entity. The map below aims to identify the contribution of research institutions to ASA journals in the world, throughout the 1991-2016 time period. We retrieved the zip codes and countries' names, the publishing institutions' names for each paper from the research address and the year from the publication year. The collaborations among authors from different institutions have led to assign the total time citation for the publication for each institution appeared in the research address field. There are 1417 research institutions in our data-set. The size of the circle represents the total time cited. The highest total time cited is 17,867 citations for Harvard University. The color of the circle represents the publication's year, from white (1991) to dark red (2016) as shown in Figure 2.13.

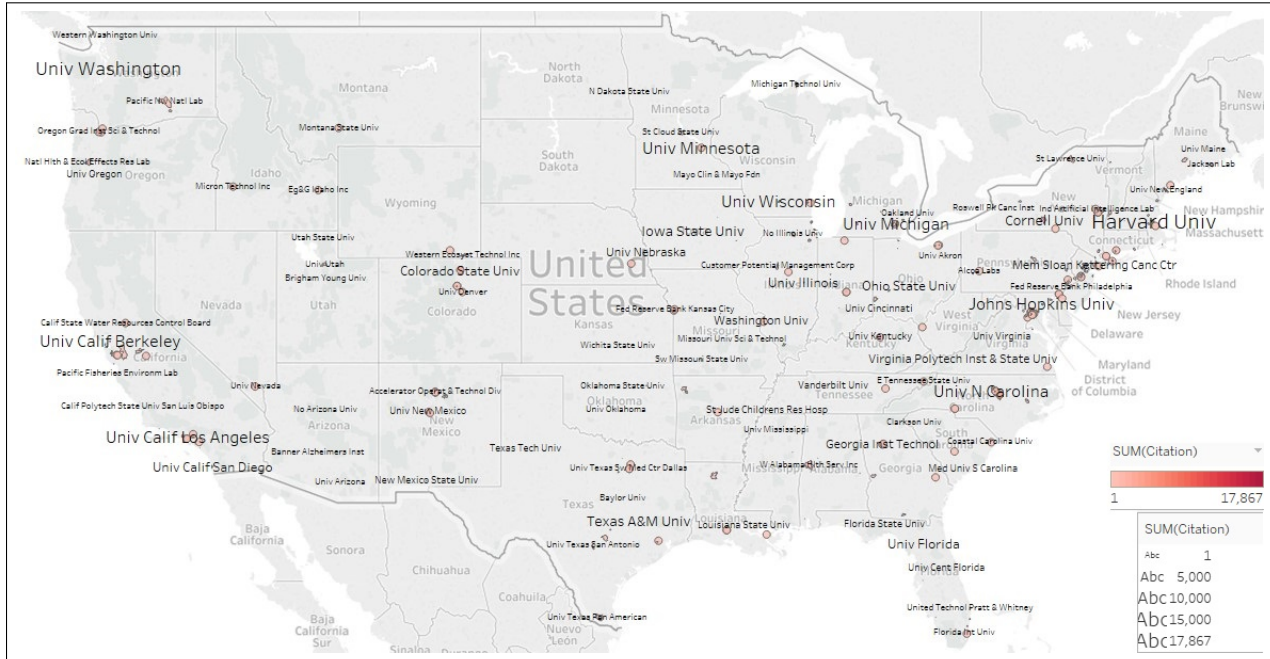


Figure 2.13: Spatiotemporal Visualization for institutions over the USA, 1991-2016.

Table 2.2 below shows the top 10 contributing research institutions. The first column shows the name of the research institutions. The second column shows the number of contributing for each institution in the data-set. The third column represents the number of citations received by each institution. In the supplementary materials, we provide an interactive visualization for our spatiotemporal analysis.

Table 2.2: The top 10 contributing institutions in ASA journals in the USA, 1991-2016.

Name of institution	of contributing for institution	of citations for institution
Harvard University	364	17,867
University of Washington	229	11,728
Stanford University	193	11,690
University of Wisconsin	219	8,184
University of NC	330	7,712
UC Berkeley	149	7,163
Duke University	261	7,060
University of Michigan	255	7,005
Johns Hopkins University	162	6,814
University of Minnesota	223	6,382

2.6 Conclusion and Future Work

The objective of this paper was to track the evolution of statistics literature over the past 26 years. We attempted to accomplish this objective by examining a sample of statistics literature that contained keywords extracted from 10,618 papers published in ASA journals. Also, we sought to achieve this goal by studying the cited references, journals within them, relationships between authors, and the spatiotemporal element within statistics literature. We also divided the time period of analysis into five periods 1991-1995, 1996-2000, 2001-2005, 2006-2010 and 2011-2016 in an attempt to identify emerging trends within the field. The methods employed in the analysis originated from the fields of bibliometrics (selection of the units of analysis), text mining (data extraction and processing), and information visualization (mapping and clustering). These methods allowed us to address the following six sub-questions that are related to the evolution of statistics:

- What are the main research fields within statistics (based on a data-driven approach)?
- How do these research fields relate to on another?
- How do these fields develop over the time period from 1991-2016?
- What are the main drivers for these publications?
- What did the top research institutions in the world contribute to the field of statistics?
- How did these institutions' contributions change over the 1991-2016 time period?

Therefore, our approach builds upon the data-driven approaches of Ryan & Woodall (2005) and Baccini et al. (2009) by using additional quantitative measures (based on keywords, cited references, cited journals, and authors) to study statistics literature. We also provide visualizations that depict results and can be used to generate additional hypotheses for further analyses.

From our keyword analysis, based on the 10,618 ASA papers, we have identified 10 major research streams within statistics. The streams cover the topics of: (1) reliability; (2) Density Estimation; (3) Computer Experiments And Geostatistics; (4) Expectation Maximization; (5) data quality; (6) Time Series; (7) Design Of Experiment; (8) Data Science; (9) Statistical Process Control; and (10) Dimension Reduction. The research within the streams has been steady, or increasing, as indicated by the calculation of the average *term score* for each of the clusters.

By investigating the cited references within these papers, we obtain a more detailed understanding of the publications driving the research in the 10 term clusters. Our results indicate that the cited references can be divided into seven groups. It was interesting to observe how Huber's 1964 paper has received considerable attention over the last six years, compared to its earlier years. This is different from how earlier works in statistics are cited over time. As Ryan & Woodall (2005, p. 462) note, "As a method becomes a generally accepted part of statistics, e.g., the *one-sample t-test*, the citation rate of the paper in which

the method was initially proposed decreases.” We believe that observations about anomalies in citation counts may not be possible without the bibliometrics approach that we have followed (under the assumption that one does not have prior knowledge for what to look for).

Based on the 136,762 citations, we have shown that the most cited journals can be grouped into seven clusters. These included two clusters on each of theoretical statistics, biological/ medical statistics, applied statistics, and data mining. From the temporal analysis of these journals, we have learned that there is an increasing amount of statistical research being conducted in machine learning. We have also shown how such analysis can be useful in generating insights into the relative impact of certain journals in a specific field.

By applying the co-authorship analysis and the author co-citation analysis on our data, we have the ability to distinguish the direct and indirect relationships among authors. The direct relationship among authors is clustered into 15 clusters of authors, while the indirect relationship is clustered into 6 clusters. By investigating the direct and the indirect relationships among the authors, we can track an author's research interest. Also, by examining the contribution of research institutions over the 1991-2016 time period, we can identify the spatiotemporal changes and the most contributing institutions among the research institutions in the World.

A number of limitations in our analysis need to be reiterated. First, the results of our analysis may be impacted by our sample of publications that we selected. The use of alternative criteria might have led to a different view on statistics literature. Since most review papers are inherently subjective, we do not consider this present discourse noticeably dissimilar from others of its kind. M.L. (2003) stated that there is no universally accepted definition of statistics, and therefore, any paper synthesizing the literature on statistics can vary based on any given perspective. There are several technical limitations that should be kept in mind with any bibliometrics analysis (Rodrigues et al., 2014). First, the maps used are restricted to a 2D space, which means that may not always be able to depict the

relatedness of the different units of analysis in the most accurate way. This information loss is inherent in any dimension reduction technique. The use of clustering restricts the terms, publications, journals, and/or authors to only one cluster. This makes it difficult to properly represent units that relate to multiple topics, especially if the distances between the units are not taken into account. We have also used somewhat arbitrary thresholds for the keywords (at least 10 occurrences), cited publications (at least 35 citations), journals (at least 30 citations), author co-authorship (at least 10 publications), and author co-citation (at least 50 co-cited). The purpose of these thresholds was to provide some measure of importance and relevance to the units of analyses. The choice of these thresholds may change the perspective of the literature. We have performed several trials to gage the impact of the threshold on the clustering. The results of these trials, which are not included in the paper, seem to indicate that the choice of threshold has limited effect within the vicinity of the values selected for the threshold. We have not conducted testing with significantly different threshold numbers since we think that these threshold values are reasonable. That being said, readers can perform a more extensive sensitivity analysis using the links presented for the interactive figures.

In summary, this paper presented an approach to analyze large amounts of bibliometrics data in the field of *statistics*. Visualizing this data with bibliographic tools, like VOSviewer, allows one to obtain a high-level view of the structure of our literature. This approach could be extended to a particular subfield of statistics. For example, it would be interesting to see how the observations from this data-driven approach will differ from the recent Woodall & Montgomery (2014) paper on the research issues and directions within *statistical process control*. Additionally, this method of analyzing the literature can help identify emerging trends in the literature. An example of this was highlighted through the growing amount of citations for Huber's 1964 paper. We hope that the analysis presented in this paper revivify additional discussion on the history of statistics and its evolution.

We provide an Excel file that contains the results of the analyses for the keywords, cited publications, journals, authors, and research address. In addition, the excel file contains all the information needed to recreate all the figures in this manuscript using different software. To download the file, please direct your browser to our website <http://www.viziolation.com/>.

Chapter 3

Appropriate Similarity Measure for Co-word “Keywords” Analysis

3.1 Abstract

The volume of scientific literature, most of which is available on the Internet, has increased in the last couple of decades, elevating, in turn, the need for new methods and techniques to expedite and facilitate the process of perceiving an overview of a scientific topic, specifically with respect to volume magnitude and visual data representation. However, as scientific literature continues its near exponential growth, methods for increasing the efficacy of data analytics are needed, warranting this present discourse. Hence, the aim of this paper is to introduce a hybrid approach for calculating the similarity measures for keywords in a co-word analysis. The proposed approach integrated the co-citations of keywords' articles within the co-word analysis of keywords to get accurate similarity metrics between keywords. This leads to better synergy between quantitative and qualitative approaches in bibliometrics research. To accomplish such an undertaking, we studied literature from the American Statistical Association (ASA) for the period of 1991-2016. Our result shows that our new method, compared to current similarity measures, has improved the ability of the clustering method to assign keywords into the right cluster by an average of 50%.

3.2 Introduction

In the past few decades, the amount of scientific literature has increased drastically, and given the evolution of the Internet and communications, much of this information is just a mouse-click away. In 2006 alone, there were 1,346,000 articles published in 23,750 journals, and in 2009, the number raised to 1.8 million articles in 28,000 journals, according to STM

(International Association of Scientific, Technical Medical Publishers) (Ware & Mabe, 2015, 2009). Also, they reported that in 2014 there were 2.5 million articles published in 28,100 journals. The average annual growth in publishing and journals is around 2.5% and 3%, respectively (Ware & Mabe, 2015, 2009). Adding to scientific knowledge on a whole, this amount of scientific literature has even created new subfields within (Jinha, 2010). Given this, new scholars and researchers find keeping up with their respective fields a challenging task. This issue raised the need for methods, tools and techniques to expedite and facilitate the process of comprehensively perceiving a given field of study. The need for analyzing and visualizing the huge amount of scientific literature data led to the creation of bibliometrics science.

Applying bibliometrics analysis on any scientific field is for equipping scholars or researchers with a coherent knowledge of the intended field, within a reasonable time frame. At the same time, an ancillary goal is to provide depth perception for the connections between units of analysis (e.g. authors, keywords, documents, and so forth). Once this is achieved, it will be more manageable for scholars to have a more accurate sense of what the new hot topics are, especially with respect to the new subfields within a given field. As a result, this paper aims to help the researcher recognize the development of a given field by using the co-word analysis approach to cluster keywords better.

The main focus of bibliometrics science involves the collection of data to present information in an insightful way. Bibliometrics tools and methodologies transform bibliographic data into formats that are more intellectually digestible. Moreover, retrieved from scientific publishing products, bibliographic data is the metadata of documents or fields; examples of metadata include keywords, words, journals, documents, authors, citations, or spatiotemporal elements, all of which are used as units of analysis (items or objects) to construct maps and to visualize the intended field. Usually, bibliographic data come in the form of innumerable records, so this type of data needs to be processed before visualizations is optimal; consequently, this discipline is heavily dependent on data mining and data visualization.

By way of overview, Börner et al. (2003) used the following sequential steps to online the procedure behind Knowledge Domain Visualization, which is also interchangeable with “bibliometrics” or “scientometrics.” (1) data extraction, (2) definition of unit of analysis, (3) selection of measures, (4) calculation of similarity between units. and (5) data visualization and analytics. Data extraction is the first step of constructing a visual representation of the intended subject and comprises the collection of bibliographic, which is highly contingent upon two criteria. The item of the subject that one wants to be study, and the relation between the items that will be mapped and visualized at the end. Retrieving the bibliographic data will be through different methods from data sources that have full record of bibliographic data or part of it, such as web of science WOS, Google Scholar, PubMed, and the like. Bibliographic literature presents the background needed for identifying the measures and for representing the relationships between them using graphs and networks.

Next, the step of defining the unit of analysis consists of two parts. The first part is to choose the unit of analysis that the one wants to study and present in a visualized map. The unit of analysis is divided into five categories. The first is the journal level where the goal is to study the relationship between journals. Next is the document level, containing articles, books, and other sorts of publications; this level is helpful when one wants to study the relationship between publications by subject or area. Third, the word level, includes a keywords list, title, abstract, or full body text as a unit of analysis. We believe that each keyword in a keywords list should be treated as one unit because the main goal here is to study the idea, not just words. Nonetheless, in the other units words will be used as the unit of analysis. For example, in chapter 2, we choose keywords to study the development of subjects and new trends in the statistics field. Fourth, there is the author level and it involves the use of authors as unit of analysis, like in chapter 2 where we investigate the direct and indirect relationship between authors. Finally, the spatiotemporal level, as the name suggests, pertains of space and time, more specifically, the study of institutions, country, and evaluation overtime.

The second part of defining the unit of analysis is preparing the units for analysis, a process that consists of cleaning and making data ready for analysis; it is this preparation phase that raises the efficiency and effectiveness of any applied analysis. This step relies heavily on methods from text mining (Rodrigues et al., 2014). Text mining employs natural language processing (NLP) methods to extract bibliographic data from the publications and preprocesses the unit of analysis. In this paper, we used NLP to preprocess keywords, our unit of analysis. We used the stemming technique to convert keywords into their basic forms (Francis & Flynn, 2010).

The third step of bibliometrics is the selection of measures. The main goal of this step is to measure relatedness between the unit of analysis. To know if your units are related or not, one should apply an appropriate measure. This explains why there are several measures in bibliometrics, each one possessing various degree of strength and weakness with respect to the particular unit of analysis being studied; however, regarding citation relation, there are three conventional approaches. The first and most used approach for citation relation is co-citation, which measures the number of times two items are cited together, and is conventionally used at the document, journal, author, and Spatiotemporal levels (Small, 1973). This technique centers upon the indirect relationship between the items of study. Next is bibliographic coupling, an approach that involves gauging the number of times two items are cited by the first and the second items (Kessler, 1962), for example, the number of references listed in two articles' references lists. (Note, this strategy can be used for the same unit of analysis as co-citation.) The third approach is direct citation, which is the number of citations received from one item to another. This approach can be used at the journal, author, and spatiotemporal levels.

The fourth step of bibliometrics visualization is the calculation of unit similarity. A similarity measure is one of the essential steps of visualizing the bibliometrics outcomes, especially since it reflects the relation between two items. For that reason, choosing an appropriate similarity measure usually results in an accurate visualization for the data-set

(Huang, 2008). The relation between two items is represented graphically in a 2D map where the distance between the items reflect their similarity. The smaller the distance indicated, the stronger the relation between the items. In this paper, the items correspond with our original keywords. However, the proposed approach can be extended to other units of analysis as well.

The fifth step of bibliometrics is data visualization and analytics. The main goal of this step is to represent the output of data visually to the user. Data visualization, be it interactive or non-interactive, is a representation of information on a graphical platform that helps the user to examine, explore, discover, comprehend , and analyze a large amount of data in a short period of time (Börner et al., 2003; Khan & Khan, 2011). This step includes grouping similar items into clusters, laying out the items on a 2D space to represent the relationship between them. Present vast amount of data visually makes the information of inquiry easier to understand aids in the process of analyzing it effectively.

There are different ways to analyze the data, but clustering, by far, is most preferred especially in conjunction with co-word analysis. In addition to the efficiency in which it reveals the quality of keyword similarity, the clustering analysis method is a crucial and helpful technique that automatically arranges items into considerably coherent clusters (Jain et al., 1999). In this paper, we use hierarchical clustering method, since it is an unsupervised technique. In co-word analysis, clustering is based on counting keyword co-occurrences and measuring relatedness between the keywords.

All of the above summarizes the typical process for producing helpful data visualization. Moreover, the product of this process provides researchers, both novice and veteran alike, with insights into various scientific fields and helps them consider the quality of the research in a holistic way. However, there is still room for improvement, which is why quantitative and qualitative methods should be used together in bibliometrics science. Such a hybridity could aid in the assessment of research quality or influence.

This paper focuses on keywords as the unit of analysis. After applying and understanding the field of bibliometrics, one may begin to see how the keywords analysis may be the best way to gain general insight into any field. This is especially the case since authors carefully choose their keywords for their research, studying the keywords gives a clear idea about the field's flow, particularly when compared to other units of analysis (Whittaker, 1989). Also, our method is ideal for keywords since it is possible to count the co-occurrence and find the co-citation of their articles.

Given that keywords is the unit of analysis in this paper and that the number of co-occurrences between keywords is a central element within our method, we had to make use of a quantitative similarity measure. Nonetheless, such an intrinsically quantitative measure, though convenient and objective in nature, is not always reliable. Hence, by thoughtfully integrating such a quantitative measure into a qualitative method, one could improve and give more reliability to a co-word analysis, which is why this present discourse focuses on the effect of the co-citation between their articles on the relatedness between the keywords. This paper accomplishes the aforementioned task by providing a comprehensive literature review; a description of the proposed method; a case study; and, lastly, a brief segment for concluding remarks.

3.3 Related Work and Background

In bibliometrics science there has been many attempts to integrate quantitative and qualitative methods to create innovative approaches to outperform old approaches. The main purposes of these new approaches are to figure out new ways to look into bibliographic data to get more information or to figure out designs that better detect inner relation. The hybridization of co-word analysis and citation in bibliometrics science has not been studied in depth due to the sheer immensity of the data involved (Zitt et al., 2011).

To the extent of our knowledge this is the first work that introduces a new similarity measure for keywords and co-citations. However, there are a few hybrid works that have been

done in the bibliometrics research. For example, one of the first attempts to integrate the quantitative and qualitative method in bibliometrics was introduced by Braam et al. (1991) when they introduced their approach to co-cluster documents by combining co-citation and word analysis. Then Hammouda & Kamel (2002) proposed a new approach to improve web document clustering by combining semi-structures inherent in web documents, document index graphs, and phrase-based similarity measures. Also, Cao & Gao (2005) they tried to improve the categorization accuracy of papers by integrating document content with a citation structure. Chim & Deng (2008) introduced the concept of using a suffix tree to cluster new documents based on phrase-based document similarity. Another researcher proposed a new approach to automatically construct a term taxonomy and the relation of terms based on weighted keywords co-occurrence (Li et al., 2015). Also, one team introduced a hybrid approach to measure the similarity level between documents (Heidarian & Dinneen, 2016).

In light of other attempts at hybridization, we have pushed forward in the spirit of innovation, understanding the increasing need to keep up with the rapid growth of scientific literature. As has been said, the volume of scientific literature, most of which is available on the Internet, has increased in the last couple of decades, elevating, in turn, the need for new methods and techniques to expedite and facilitate the process of perceiving an overview of a scientific topic, specifically with respect to volume magnitude and visual data representation. At the present time, the bibliometrics mapping of science is a quantitative method for studying bibliographic data (e.g. titles, keywords, authors, and so on) and visually representing the information. Furthermore, when it comes to dealing with a large body of literature, bibliometrics maps are exceptionally useful and have been used in different contexts like research literature surveys, government decision-making, and scientific publications.

The aforementioned considerations are precisely why we saw an opportunity to improve visualization via co-word analysis. However, before delving more into the hybrid method, let's backtrack a bit to paint the co-word analysis method with broad strokes.

Since its inception, co-word analysis has been used in many bibliometrics and information retrieval studies to assess and illustrate a given field. The co-word analysis has been used in different scientific fields: Hoonlor et al. (2013) used the co-word analysis to study the trends in computer science research, and Liu et al. (2014) used it to study the field of human-computer interaction, which entails the interchange between academic and technological research in the polymer chemistry area (Callon et al., 1991). Additionally, Ding et al. (2001) applied this approach to map the intellectual structure of the information retrieval field for the period of 1987-1997. Coulter et al. (1998) applied it to software engineering. In the environmental acidification area, Law et al. (1988) applied this approach as well. Co-word analysis has also been helpful in Condensed Matter Physics Bhattacharya & Basu (1998) and the An & Wu (2011) stem cells field.

Integral to co-word analysis is the notion of co-occurrence. Two keywords co-occurring in the same paper indicates the relationship between them. The higher the number of the co-occurrence, the greater the relationship between them. Usually, the strength of the association between the two keywords is decided by the times they appear together within the same paper, but this does not reflect their inner relation to the field. In other words, two keywords are connected more if their articles are related too. In other words, the co-occurrence of items measures the relatedness between two items by the number of times two items appeared together.

Co-occurrence is mostly used at word level, which is usually called co-word. The first proposition for a co-word analysis was made by Callon et al. (1986) in the book *Mapping the Dynamics of Science and Technology* in 1986. Since the appearance of this book, the co-word analysis technique has become an important method in bibliometrics science. The co-word process analyzes scientific literature based on item co-occurrence, where it studies relationships and reveals term trends and patterns in scientific papers (Callon et al., 1991); additionally, it detects patterns of keywords, words, and noun phrases as they co-occur in a corpus of texts, keywords lists, or titles. In addition to examining the co-occurrence of

terms as the indicator of the similarity between any two, the co-word analysis identifies the relationships between items within the subject areas presented in the texts being investigated (He, 1999).

Clustering is also an important process for showing co-word relatedness, where similar keywords form in the same group, the dissimilar keyword is in a different group, and the distance between them represents their relatedness. Usually, the similarity measure is bounded between $[0, 1]$, where 0 means that the two keywords are disjointed, revealing their distinction from each other. 1 means the two keywords are the same. Put differently, co-word analysis represents the relation between keywords in a map by decreasing distance between them (Whittaker, 1989). In distance-based mapping, the distance between two keywords represents their similarity, and the single numeric value, i.e. the “similarity value,” is affected by two factors: one, the properties of the keywords and, two, the measure itself (Huang, 2008).

There are two main steps within the co-word analysis (van Eck et al., 2005). The first step is the calculation of the similarities between keywords, and the second step is mapping keywords, which will be utilized more in Chapter 4. In the first step, one usually calculates the co-occurrence for each pair of keywords. Subsequently, one must store the co-occurrence in a co-occurrence frequency matrix and then convert it to a similarity matrix.

The co-occurrence frequencies matrix does not reflect the similarity between keywords (Waltman & Eck, 2007). Therefore, the co-occurrence frequencies matrix should be converted to the similarity matrix to show the similarity between keywords. The co-occurrence data are used in bibliometrics science to construct co-word maps (van Eck & Waltman, 2009), and the main purpose of the similarity measure is to transform the co-occurrence frequency into a similarity matrix for keywords by normalizing the co-occurrence frequency and presenting it as a relation of similarity between keywords. There are two approaches for transforming a co-occurrence frequency matrix into a similarity matrix. The direct similarity measure is the

first approach; this measure involves the normalization of the co-occurrence by using similarity techniques such as the inclusion index (Rip & Courtial, 1984), the Jaccard coefficient (Peters & van Raan, 1993), the cosine similarity (Larsen & Aone, 1999) or the association strength (Van Eck & Waltman, 2007). The second one consists of the indirect similarity measures, mainly used in co-citation data in the bibliometrics field (McCain, 1990).

To determine the similarity between two keywords, the direct similarity measure adjusts the co-occurrence of the two keywords to the total number of occurrences for each of the two keywords. Because the similarity value between two keywords is bounded between $[0,1]$, it should be noted that one of the most important properties of direct similarity is that it measures all non-integer values of similarity (Eck & Waltman, 2009). The co-occurrence frequency matrix is symmetrical and proximal, which means that numbers in the matrix could be similar or dissimilar between items (Kruskal & Wish, 1978; Cox & Cox, 2000). In this paper, we focus on converting the keywords co-occurrence frequency matrix into a similarity matrix, excluding the use of all measures that are not suitable for the similarity of the co-occurrence matrix. The measure should be suitable for our data and must meet two criteria. First, they must be suitable for co-occurrence data. Second, they must be suitable for a similar symmetrical matrix. Given this, we do not consider indirect similarity measures because they are not suitable for co-word analysis (McCain, 1990). Another reason for this decision is that (Eck & Waltman, 2009) stated that an indirect similarity approach is not ideal for co-occurrence data because it compares the co-occurrence profile of the object to determine the similarity between the objects. Nonetheless, from a statistical perspective, an indirect similarity measure has been described as an unconventional approach by (Schneider & Borlund, 2007), the claim being that it is more appropriate for co-citation data. Since our focus is on keyword co-occurrence, it is more compatible to use the direct similarity measures.

In bibliometrics there are several limitations presented regarding similarity measures for co-occurrence data or co-word analysis. In such a multidisciplinary field, like statistics,

words and terms could have different meanings in various statistical subfields (Yang et al., 2017; Peters & van Raan, 1993; Vaughan & You, 2010). In our proposed method, this problem will be avoided since we do not use just the number of co-occurrence between two keywords; we also use their articles' co-citation to measure their relatedness. If two articles are co-cited together, they fall in the same subfield, and therefore their keywords have a similar meaning or intention. Plus, the keywords that occur more frequently will be related with so many other keywords, but since old similarity measures are quantitative, a purely quantitative approach will not be able to differentiate between if the two keywords are in relevant or irrelevant subfields (Zhang et al., 2014; Peat & Willett, 1991; Yang et al., 2017). The current similarity measures only take into account the number of keywords co-occurrences as a measure between each pair of keywords, without considering the relation of their articles. Such a tunnel-vision focus misses an important relation between keywords. If two articles are co-cited, they are related, and they fall into the same subfield; therefore their keywords are more related than if they co-occur.

As mentioned earlier, there are different types of direct similarity measures, and the four measures that are most well-known among similarity measures in the literature of co-word analysis are: the association strength, the cosine, the inclusion index, and the Jaccard index. First, the association strength is used for normalizing the co-occurrence frequency matrix, which is also known as the proximity index Rip & Courtial (1984) or the probabilistic affinity index (Zitt et al., 2000). The association strength is proportional between the co-occurrence of the two keywords and the appearance of each keyword. One of the main drawbacks in the association similarity is that it will not reach full similarity, namely, the similarity between two keywords cannot be 1 unless both keywords occur one time and each co-occur in that appearance. Such a scenario seems unrealistic. (See Table 3.1 for more related measures to the association strength.)

Second, the cosine, also synonymous with different bibliometrics terms, was introduced in 1986 in Salton's book *Introduction to Modern Information Retrieval* (Salton & McGill,

1986), and since then it has become the most used similarity measure in bibliometrics science (Eck & Waltman, 2009). In addition, representing the ratio between the co-occurrence of the two keywords and the average appearance of each keyword, the cosine is also known as the equivalence index (Callon et al., 1991; Kostoff et al., 1999) or Salton's measure (Luukkonen et al., 1993; Glänzel, 2001). The cosine similarity measure outperforms other similarity measures in bibliometrics (Nelson et al., 2004). However Eck & Waltman (2009) stated that the cosine similarity measure is not appropriate for normalizing co-occurrence data. Also, Heidarian & Dinneen (2016) have discussed in detail the drawbacks of cosine similarity and concluded that cosine similarity is more appropriate for measuring the difference between two items instead of similarity.

Third, the inclusion index, often referred to as the overlap measure (Salton & McGill, 1986; Jones & Furnas, 1987), has been used in not a few studies (Rip & Courtial, 1984; Kostoff et al., 2001). Finally, the Jaccard index (Small, 1973) is one of the most used similarity indices in co-word analysis and is applicable in co-citation analysis. The Jaccard index represents the ratio between the keywords' co-occurring and the number of times one of the keywords occurs the least. Eck & Waltman (2009) stated that the Jaccard index is not appropriate for normalizing co-occurrence data in co-word analysis.

Worth mentioning is that the previous similarity measures are employed outside the bibliometrics fields, some of them lacking particular designations. for example, in non-bibliometrics fields the Ochiai coefficient replaces the cosine; the Simpson coefficient replaces the inclusion index; and lastly, the Dice coefficient usurps the role of the Jaccard (Cox & Cox, 2000). The cosine similarity measure is used in co-citation (Anderberg, 1973), documentdocument similarity (Ahlgren & Jarneving, 2008; Ahlgren & Colliander, 2009; Salton & Buckley, 1988), between two articles (Baeza-Yates et al., 1999). Having been theoretically evaluated in bibliometrics literature, the four popular direct similarity measures have been categorized into two classes (Egghe & Michel, 2002, 2003; Egghe & Rousseau, 2006; Baulieu,

1989, 1997; Janson & Vegelius, 1981). The classes are set-theoretic and probabilistic similarity measures. In short, the cosine, inclusion index, and the Jaccard index fall into the set-theoretic similarity measures class, but the association similarity measure belongs to the probabilistic similarity measures class.

Other practices include indirect similarity measures like the Pearson correlation coefficients (McCain, 1990) and the chi-squared distance, notwithstanding that these measures do not have all the notable theoretical properties for co-occurrence data (Ahlgren et al., 2003). Also, we exclude the Euclidean distance matrix because it is used for calculating a dissimilarity matrix (Leydesdorff & Vaughan, 2006). Also, Heidarian & Dinneen (2016) have explained in detail the drawback of Euclidean distance. Also, there are more indirect similarity measures used in bibliometrics science, including the Bhattacharyya distance (Lin, 1991), the indirect cosine (Ahlgren et al., 2003), and the Jensen-Shannon distance (Shannon, 1948). Table 3.1 shows the direct and indirect similarity measures with their alternative names and monotonically related measures. Also, it shows where the similarity measure has been used in the literature of similarity measures. The last column shows the parameters that used in each method.

Table 3.1: Relations among various direct and indirect similarity measures.

	References	Alternative names	References	Monotonically related measures	References	Parameters
Direct Similarity Measures						
Association strength	(Van Eck & Waltman, 2007)	Probabilistic affinity index	(Zitt et al., 2000)	(Pointwise) Mutual information	(Church & Hanks, 1990)	Keywords count
		Proximity index	(Rip & Courtial, 1984)			Keywords count
		Pseudo-cosine	(Jones & Furnas, 1987)			Keywords count
Cosine	(Larsen & Aone, 1999)	Ochiai coefficient	(Sokal et al., 1963)	Equivalence index	(Callon et al., 1991)	Keywords count
		Salton's index/measure	(Luukkonen et al., 1993)			Keywords count
Inclusion index	(Rip & Courtial, 1984)	Overlap measure	(Salton & McGill, 1986)			Keywords count
		Simpson coefficient	(Hubalek, 1982)			Keywords count
Jaccard index	(Small, 1973)			Dice coefficient	(Salton & McGill, 1986)	Keywords count
Indirect Similarity Measures						
Bhattacharyya distance	(Lin, 1991)					Keywords count
Jensen-Shannon distance	(Shannon, 1948)					Keywords count
Cosine	(Ahlgren et al., 2003)					Keywords count
Pearson correlation Coefficients	(McCain, 1990)					Keywords count
The chi squared distance	(Ahlgren et al., 2003)					Keywords count
New Method						Keywords count and Articles' co-citations

In the literature of co-word analysis, scholars have used the title, the abstract, the keywords list, and the full text body of the article as the units of analysis to complete the investigation. Articles' titles tend to be informative for the reader, telling him or her what to expect; however, titles do not provide researchers with an exhaustive overview of the field. Usually, authors try to write provocative titles to lure readers to either download or to increase the reading of his paper (Whittaker, 1989), which Whittaker (1989) called the “audience effect.” Also, he added another reason to not consider the title as a good unit of analysis for co-word analysis, which he called the non-standard titles; this is where the author uses words, non-phrases, or rhetorical questions to catch the eye of the reader. One more shortcoming of using titles is that at the end it would have words instead of ideas or concepts (Whittaker, 1989); moreover, this point applies to the abstract and full text, as well. In addition, in most of the journals the number of characters for the title is limited (Gbur Jr & Trumbo, 1995). Although it does provide a noticeable insight into that particular article, the abstract is a brief summary of an article and its findings, neither of which helps the researcher to catch the main theme of the field. Moreover, the full text body sometimes contains different words and phrases that may not be helpful for a researcher.

The keywords list is a group of words, phrases, concepts, or ideas that are used in a published scientific work, enabling a reader to gain insight into an article quickly. The main purpose of keywords is to give the scientific article's reader a bird's eye view of the article's main theme, which is why the typical scientific author carefully chooses technical keywords that best fit the article (Whittaker, 1989). In 1975, the *Journal of Applied Behavior* became the first scholarly journal to use the keywords list (Hartley & Kostoff, 2003). Gbur Jr & Trumbo (1995) described the keywords list as a focused mini-abstract of the article. Also, he suggested an ingenious procedure for methodically creating a keywords list (Gbur Jr & Trumbo, 1995). Keywords that are listed by authors are more informative and accurate for researchers, making for an optimal online search experience (Gil-Leiva & Alonso-Arroyo, 2007).

Given the enormity of the available database, Hazewinkel (1999) has argued the importance of having controlled keywords to make it easier for average users and new scholars to get the desired information easily (Hazewinkel, 1999). According to Whittaker (1989), the relationship between keywords are more salient when more authors use similar keywords together in their publication, eventually leading to a new direction or subfield in that scientific area. The newest scientific databases provide another keywords list that has been picked by others. For example, WOS calls such a list “New ISI keywords.” Many researchers have argued for the reality of a so-called “indexer effect.” For instance, Whittaker (1989) interviewed authors of published scientific papers and discovered that these authors do not deem the new keywords list an accurate characterization of their work (Whittaker, 1989). Also, Leydesdorff (1987) states that when someone who has no expertise in a field selects the so-called new keywords, that is a good indication of the inaccuracy of the selection (Leydesdorff, 1987). Whittaker (1989) reported that the cohesion of clusters in keywords is higher than title words, and analyzing keywords gives an excellent output analysis of the intended field than titles (Whittaker, 1989). Moreover, Hartley & Kostoff (2003) have listed a number of advantages of keywords. For example, they believe that the keywords list allows the reader to find the relevant papers that s/he is looking for. Also, it is a terrific way to start researching the topic of interest online. In addition, the keywords list helps the publishing institution and researchers to group and classify topics within a scientific field. Also, Hartley & Kostoff (2003) reported that 75% of the journals in the statistics field use keywords, which is the highest percentage between the fields he reported.

3.4 Method

The procedure for keywords analysis is as follows:

1. Select the threshold for the minimum appearance in the data-set, S_n .
2. Select the minimum co-occurrence between two keywords, C_{ij} .

3. Store the co-occurrence of keywords into a co-occurrence frequency matrix, C .
4. Normalize the co-occurrence frequency matrix into the similarity matrix, A .
5. Cluster.
6. Map.

As we discussed earlier, choosing the right similarity measure to normalize the co-occurrence frequency matrix is a crucial step in the co-word analysis, that is, it is critical to use the right similarity measure. The first dilemma involves determining the appropriate similarity measure to use to normalize the data. Basically, we need to choose between the indirect and direct similarity measures. As mentioned above, the indirect similarity measures are not ideal for co-occurrence data, but they are more appropriate for co-citation data. On the other hand, to normalize the co-occurrence data, applying the direct similarity measures to correct the data for differences in the number of co-occurring keywords is more appropriate for co-occurrence data (van Eck & Waltman, 2009).

Two keywords co-occur if they appear together in the same keywords list. Let S_i and S_j represent the occurrence frequency of the keywords i and j , respectively. Let C_{ij} represent the co-occurrence of keywords i and j . The keywords co-occurrence matrix is denoted by C for the keywords $1, \dots, n$, where n indicates the number of keywords. The similarity between two keywords is represented by A . Since the co-occurrence is calculated based on the frequency of two keywords appearing together, then we can posit that C is a non-negative integer matrix ($n * n$). In co-word analysis literature, there are conditions that must be met by the co-occurrence of two keywords C_{ij} . These are illustrated below:

$$C_{ij} \geq 0, \text{ for } i, j = 1, \dots, n \quad (3.1)$$

$$C_{ii} = 0, \text{ for } i = 1, \dots, n \quad (3.2)$$

$$C_{ij} = C_{ji}, \text{ for } i, j = 1, \dots, n \quad (3.3)$$

$$\sum_{j=1}^n C_{ij} > 0, \text{ for } i = 1, \dots, n \quad (3.4)$$

After presenting the mathematical notation of the co-occurrence frequency matrix, then the next step is to normalize the co-occurrence matrix and convert it to a similarity matrix. As we mentioned above the four most used direct similarity measures in bibliometrics science are defined below. The association strength, the cosine, the inclusion index, and the Jaccard index, are equations, 3.5, 3.6, 3.7, and 3.8, respectively.

$$S_a = \frac{C_{ij}}{S_i S_j} \quad (3.5)$$

$$S_c = \frac{C_{ij}}{\sqrt{S_i S_j}} \quad (3.6)$$

$$S_i = \frac{C_{ij}}{\min(S_i, S_j)} \quad (3.7)$$

$$S_J = \frac{C_{ij}}{S_i + S_j - C_{ij}} \quad (3.8)$$

Under the assumption that occurrences of keywords i and j are statistically independent; we can postulate that the association strength represents the proportional to the ratio between the co-occurring of keywords i and j and the expected number of co-occurrences of keywords i and j (van Eck & Waltman, 2009). The cosine equals the ratio between the number of co-occurrence of keywords i and j and the average appearing of keywords i and j . The inclusion index equals the frequency of the more frequent keyword (Rip & Courtial,

1984). The Jaccard index represents the co-occurring of keywords i and j over the union of the keywords i and j (Hamers et al., 1989).

Now, after explaining the well-known method, we are going to introduce our new method, the method defined in equation 3.12 below. In this equation, S_{nm} denotes the strength similarity between keywords i and j , as C_{ij} denotes the number of times keywords i and j appeared together. S_i denotes the number of times keyword i appeared in the data-set. S_j denotes the number of times keyword j appeared in the data-set. S_t denotes the number of times the articles that included keywords i and j have been co-cited. O_n represents the articles that contain the keywords list of either i or j keywords. Notably, conditions 3.1, 3.2, 3.3, and 3.4 are satisfied in the new method. Besides that, 3.9, 3.10 and 3.11 are special conditions for the new method. Equation 3.9 states that if the keyword i and j articles are co-cited together, then the maximum number S_t can be is the minimum between S_i and S_j , and if they are not co-cited then S_t will be zero. S_t equals zero when O_i and O_j have not interacted, see condition 3.10.

$$S_t \leq \{\min (S_i , S_j) , 0\} , \text{ for } i, j = 1, \dots, n \quad (3.9)$$

$$S_t = 0, \text{ for } O_i \ominus O_j \quad (3.10)$$

In situations when the two keywords' articles have never been co-cited together, the value of S_t will be zero, and to prevent the equation from being undefined, we used e as a constant in condition 3.11

$$\frac{e^{S_t}}{e^{S_t}} = 1, \forall S_t \quad (3.11)$$

Here, $e^0 = 1$, then $\frac{e^{S_t}}{e^{S_t}} = 1$, a genetic equation that can be applied to any data-set. The new method equation 3.12 is

$$S_{nm} = \frac{C_{ij} \left(\frac{e^{S_t}}{e^{S_t}} + S_t \right)}{S_j S_i + S_t} \quad (3.12)$$

Noteworthy is that S_t , the effect of keywords' article co-citation, has not been used in any other known direct or indirect similarity measures for the co-occurrence data.

Our threshold is 10 appearances for keywords in the data-set, our keywords data-set consisting of the co-occurrence frequencies of 485 keywords in ASA in the period of 1991-2016. The total number of articles used in this study is 8,758 articles. The 485 keywords appeared in 8,319 articles, occurring a total of 14,781 times. The number of co-occurrences between the keywords is 11,369.

The next step is creating the co-occurrence matrix of all keywords by computing the frequency of two keywords that appear together in the same article. Then, we count the co-citations of articles for each couple of keywords. Then we store them into a keywords co-cited articles matrix. We had two symmetrical matrices, the first one based on the word co-occurrence and the second one based on the co-citation between articles of the two keywords. Next, we performed one of the most important steps in the co-word analysis, we converted the two matrices into a similarity matrix by applying our new similarity method. Then we used the clustering method to illustrate our new method. However, the clustering method will not be described here, since the main focus of this paper is the similarity measures.

3.5 Empirical Comparison

Usually, in bibliometrics science an empirical comparison is used to show the difference between methods or relatedness, and to present new methods as well. We conducted an empirical comparison that is similar to what Leydesdorff (2008); Eck & Waltman (2009) have done in their works; nevertheless, we add to it our proposed new method. First, we calculated the similarity matrix for each of the four well-known direct similarity measures, which are the association strength, the cosine, the inclusion index, the Jaccard index, and

our new method. Next, we created a matrix scatter plot to compare the similarity measures, intending to elucidate the relationship between them. Since the co-occurrence matrix has zero co-occurring values between some keywords, we had zero-values in the similarity matrix as well for all five similarity measures. Hence, we removed all zero-values for the similarity measures to avoid false relatively high level of correlations(Eck & Waltman, 2009), explaining why we calculated the non-zero values only. See the matrix scatter plot below, Figure 3.1

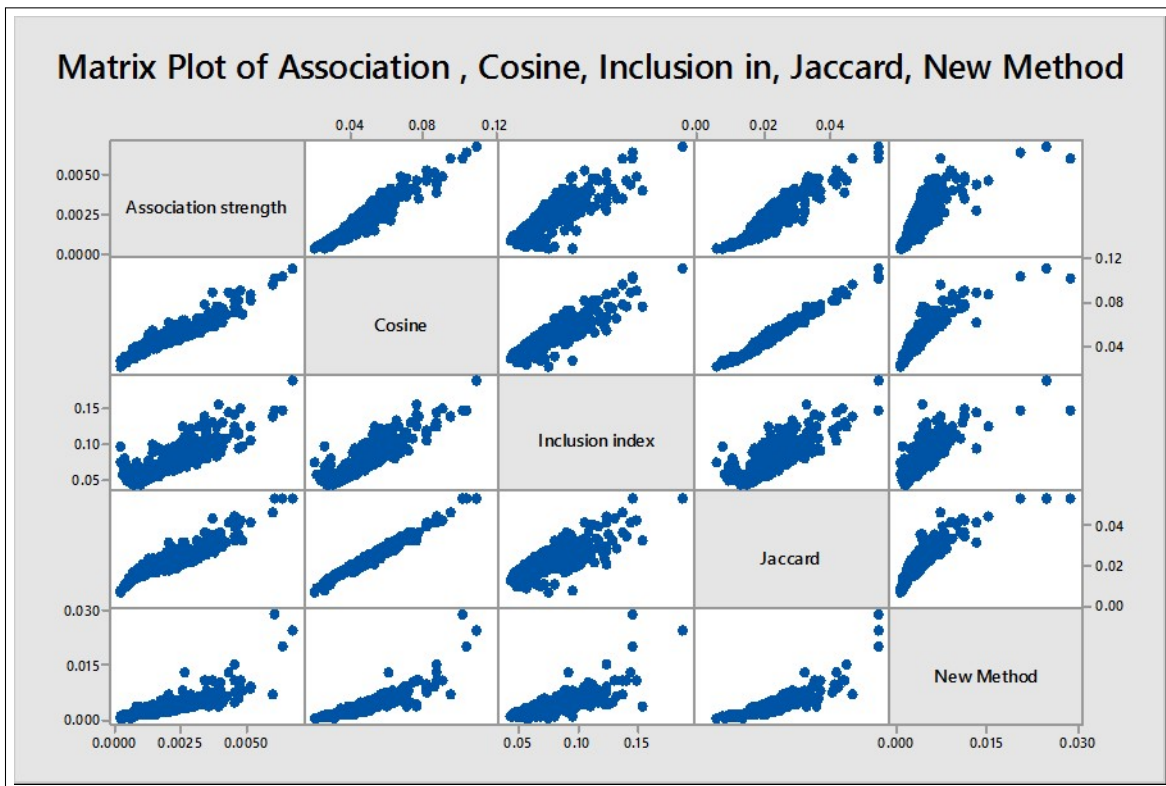


Figure 3.1: Scatter plots obtained for the keywords between each of similarity measures.

Then, we calculated the Pearson correlation to show the degree of linearity between each two-similarity measure; after that, we calculated the Spearman correlation to show the monotonical relatedness. Table 3.2 shows the values for the Pearson correlation, which is in the upper right part, and the Spearman correlation, the lower left part.

In Figure 3.1, the matrix scatter plots show that there is a substantial difference between our new method and the other direct similarity measures. The relation between the new method and the other similarity measures is weak. The matrix scatter plot shows that

Table 3.2: Correlations obtained for Keywords.

	Association	Cosine	Inclusion	Jaccard	New Method
Association		0.955	0.859	0.928	0.814
Cosine	0.961		0.878	0.986	0.869
Inclusion	0.877	0.888		0.798	0.778
Jaccard	0.93	0.983	0.804		0.872
New Method	0.917	0.945	0.857	0.923	

there is a strong relation between the cosine and Jaccard index. This result is supported by earlier studies (Leydesdorff, 2008; Egghe, 2009; Hamers et al., 1989). Also, there is a strong relationship between the association strength and the cosine. Beside these two relations, the rest of the relations between the other similarity measures are weak. Furthermore, if we look at the scatter plot figure, we can conclude that more of the low value of the new method corresponds with a high value of the other similarity measures. This finding is supported by the Spearman correlation, which is high between the new method and the conventional ones. Additionally, looking at the correlations in Table 3.2 between the new method and association strength, we see that the cosine, inclusion, and Jaccard are 0.814, 0.869, 0.778, and 0.872, respectively. Our result is similar to that of Eck & Waltman (2009) but is different from the result reported by Leydesdorff (2008). We can state that in practical applications the similarity measures will have different outputs.

For deep insight into our new method, especially in comparison with the other direct methods, we relied upon hypothetical scenarios, which have been used in the Eck & Waltman (2009) works . At this time, we are going to introduce more scenarios than their work to illustrate our idea. The eight scenarios are displayed in Table 3.3. These eight scenarios represent a co-word analysis between keyword i and keyword j , where m is the total number of documents that the keywords appear in, C_{ij} is the co-occurrence number between the two keywords, and S_t represents the number of times the two keywords have been co-cited together in different articles.

Table 3.3: Hypothetical scenarios.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8
m	1000	1000	1000	1000	1000	1000	1000	1000
S_i	40	40	40	40	400	400	400	400
S_j	40	40	40	20	400	400	400	200
C_{ij}	8	8	8	8	160	160	160	160
S_t	0	8	4	4	0	160	80	80
Association strength	0.005	0.005	0.005	0.01	0.001	0.001	0.001	0.002
Cosine	0.2	0.2	0.2	0.282	0.4	0.4	0.4	0.566
Inclusion index	0.2	0.2	0.2	0.4	0.4	0.4	0.4	0.8
Jaccard index	0.111	0.111	0.111	0.154	0.25	0.25	0.25	0.364
New Method	0.005	0.045	0.025	0.050	0.001	0.161	0.081	0.162

Basically, the eight scenarios fall into two parts. Scenarios 1,2,3, and 4 represent the first part, and scenarios 5,6,7, and 8 represent the second part. In the first and second parts, m stays at 1000 articles in the eight scenarios, while S_i occurs 40 times in the first part and 400 in the second part. Also S_j remains the same for the first three scenarios of each part; however, we divide it by two in the fourth scenario. C_{ij} occurs 8 times for the first part and 160 times in the second. We have varied the values of S_t to illustrate our new method better. Below, we are going to explain each scenario and compare them together.

In scenario 1, there are 1000 documents that represent the data-set, and keywords i and j appeared in 40 documents, meaning each one appears in 4% of the documents and they co-occur 8 times. As we all know the keywords are statistically independent (Eck & Waltman, 2009). The new method and the association strength will have the same value when S_t equals zero. If we compare scenarios 1 and 5, where S_t is equal to zero, we can note that the cosine and inclusion index have the same similarity value (which is 0.2) for scenario 1 and (0.4) for scenario 5. Also, the Jaccard has a lower value (0.11) for scenario 1 and (0.25) for scenario 5. On the other hand, the association strength and new method have the same value for scenario 1 (which is 0.005) and a lower value for scenario 5 (which is 0.001). The reason why the new method and association strength have higher similarity values in

the fifth scenario than in the first one is due to the inner workings of scenario 5, wherein each keyword occurs 40% throughout all articles. $40\% \times 40\% = 16\%$, and the co-occurrences between them happened exactly 16%. On the other hand, in scenario 1 each keyword occurs 4% throughout all articles. $4\% \times 4\% = 0.16\%$, but the co-occurrence is 0.8%. This means it co-occurs five times more than in scenario 5. The new method similarity value for scenario 5 is 0.001 and is 0.005 for scenario 1, five times greater than that of scenario 5.

Now, let's compare scenarios 1,2,3, and 4. The m value is constant at 1000 articles, and S_i occurs 40 times for the four scenarios. Moreover S_j has 40 occurrences for the first three scenarios and 20 occurrences for the fourth scenario, while C_{ij} only has 8 co-occurrences for the four first scenarios. S_t , as we explained earlier, is the times the keywords' articles have been co-cited. S_t is zero in the first scenario, which is similar to the association strength; 8 in the second scenario; and 4 within the last two scenarios. The similarity value for scenario 1 is 0.005, and since the S_t is equal to 0, it has no influence in the similarity between keywords i and j . However, in scenario 2 the S_t equals 8, which means that 100% of the co-occurrences have been co-cited and that increased the similarity value between keywords i and j from 0.005 to 0.045, which is 8 times more. Nonetheless, if we compare scenario 2 to 3, the S_t in scenario 3 is equal to 4, which is half the value of the S_t in scenario 2; hence, the similarity value in scenario 3 is two times less than the one in scenario 1. In scenario 4 the S_j is half of the one in scenario 3, but they are equal to C_{ij} and S_t . In contrast, scenario 4 is more than scenarios 3 and 2 because 40% of the S_j in scenario 4 have co-occurred with S_i , and 50% of them have been co-cited. From Table 3.3, we can state that

$$S_{nm} \leq S_a \leq S_J \leq S_c \leq S_I \quad (3.13)$$

3.6 Case Study

To perform our new method and co-word analysis, we selected the most frequent keywords, in order to reduce the data dispersion, and we chose our threshold to be at least

10 appearances. We obtained 485 keywords. We chose to have the minimum co-occurrence between any two keywords to be at least one time.

In the past couple of decades, more and more new subfields have appeared in the statistical science field. This phenomenon raised the need to develop methods to study the bibliometrics of statistical science. In this study, the co-word analysis was used to investigate the knowledge structure of the statistics field. However, we found that the old similarity measures used in the co-words analysis is a quantitative similarity measure between the keywords, prompting us to introduce our new method, which integrates quantitative and qualitative components to improve the similarity between keywords. Our result showed that the new method has the ability to put similar keywords into the same cluster better than the old direct similarity measures.

We will use a data-set from the field of statistical science as a case study to illustrate our method; this data-set was compiled from American Statistical Association (ASA) journals, between 1991 and 2016. The data were retrieved from the Web of Science(WOS) (*WEB OF SCIENCE @ONLINE*, 2016). In the search engine, we searched for all 17 of the ASA journals by typing the name of the journals. *“Journal of the American Statistical Association,” “Journal of Statistical Software,” “Journal of Computational and Graphical Statistics,” “Journal of Business & Economic Statistics,” “Technometrics,” “Journal of Educational and Behavioral Statistics,” “Journal of Nonparametric Statistics,” “Journal of Agricultural, Biological, and Environmental Statistics,” “Journal of Quantitative Analysis in Sports,” “Journal of Survey Statistics and Methodology,” “Statistics Surveys,” “The American Statistician,” “Journal of Statistics Education,” “SIAM/ASA Journal on Uncertainty Quantification,” “Statistical Analysis and Data Mining: The ASA Data Science Journal,” “Statistics and Public Policy,” and “Statistics in Biopharmaceutical Research.”*

We retrieved a total of 10,618 records, all of which were found on the WOS. Next, keeping only the articles, the essence of any scientific field, we excluded several record types: proceedings paper, biographical-item, book review, correction, editorial material, letter, meeting

abstract, note, reprint, review, and software review. After omitting the unneeded records, we were left with 9,028 articles, only 8,758 having keywords. The total number of keywords are 44,147, and the average number of keywords for each article is around five, which is reasonable since most of the journals ask the authors to provide from three to five keywords when they submit articles. From there, we cleaned and preprocessed the keywords.

The cleaning and preprocessing of the 17,766 unique keywords of the data-set involved several steps. First, we wrote an R language code to execute the data cleaning and preprocessing; we used text mining techniques and packages, such as `tm` (Feinerer, 2017) and `wordcloud` (Fellows et al., 2012) . The process went as follows: first, we went through the keywords list to correct the spelling mistakes and change plural words to singular words. We used the text mining technique stemming to change, for example, words like “charts” to “chart” and “abilities” to “ability.” Also, to preserve the grammatical and statistical simplicity of a given word, we changed words like “adaptation” to “adaptive.” With respect to verb tenses, gerunds, and infinitives, we selected the most elementary form, e.g. we changed “fitting” to “fit.” Along the way, we also removed all unneeded signs and symbols entailing the omission of all double spaces to consolidate keywords with the same apparent meaning, like “R Package,” “R program,” “R programming,” “R software,” or “R language” to “R”. In spite of our efficiency, we faced one problem with popular acronyms or abbreviations for keywords, most likely the result of authors' submitting keywords lists in different styles. For instance, “Akaike information criterion (AIC),” “akaike AIC,” “akaike criteria,” “Akaike criterion,” “Akaike information criterion,” “alternate conditioning expectation (ACE),” and “alternative Conditional Expectation (ACE) Algorithm” all have the same apparent meaning, so we decided to signify them with the AIC acronym. Below, Table 3.4 shows a list of acronyms and definitions for the most used keywords in the ASA data-set. After that we checked the outcome manually. The processes led to the final number of 44,147 keywords, and from there, we extracted 17,766 unique keywords.

Table 3.4: List of acronyms and definitions of keywords used in the ASA.

Acronym	Definition	Acronym	Definition
AOD	Aerosol Optical Depth	IRT	Item Response Theory
AIC	Akaike Information Criterion	LASSO	Least Absolute Shrinkage And Selection Operator
ACE	Alternate Conditioning Expectation	LDA	Linear Discriminant Analysis
ANCOVA	Analysis Of Covariance	LMM	Linear Mixed Effect Model
ANOVA	Analysis Of Variance	MML	Marginal Maximum Likelihood
AUC	Area Under The Curve	MRF	Markov Random Field
ARL	Average Run Length	MISE	Mean Integrated Square Error
CART	Classification And Regression Tree	NAEP	National Assessment of Educational Progress
CUSUM	Cumulative Sum	ODE	Ordinary Differential Equation
FDR	False Discovery Rate	PCA	Principal Component Analysis
FWER	Family Wise Error Rate	RKHS	Reproducing Kernel Hilbert Space
FRK	Fixed Rank Kriging	REML	Restricted Maximum Likelihood
FMRI	Functional Magnetic Resonance Imaging	SVD	Singular Value Decomposition
GAM	Generalized Additive Model	SAVE	Sliced Average Variance Estimation
GEE	Generalized Estimating Equation	SIR	Sliced Inverse Regression
GLMM	Generalized Linear Mixed Effect Model	SCAD	Smoothly Clipped Absolute Deviation
GMM	Generalized Method Of Moment	SURE	Stein'S Unbiased Risk Estimate
GWAS	Genome Wide Association Study	TAR	Threshold Autoregressive
GIS	Geographic Information System	VAR	Vector Autoregressive

Three hundred and forty one keywords met our threshold. Then we calculated the co-occurrence matrix for all of the keywords. Beside the co-occurrence matrix for the keywords, we created an articles' keyword co-citation matrix. Then, we converted the co-occurrence matrix to a similarity matrix using the four well-known similarity measures, and for our new method we used the two-created matrices to get a similarity matrix. From there, we just needed to choose the best way to cluster our data.

The main purpose for data mining and bibliometrics is to find hidden patterns. Since we did not have information about our data items or classification of the data, we used the unsupervised clustering technique Tan et al. (2006) to reveal the intellectual structure of the ASA data-set. We believe that the keyword co-word analysis should be clustered through the unsupervised clustering method, since we do not have classified data to compare our data to. Consequently, the analysis must be completely unsupervised, so we chose to use a hierarchical clustering method to assign keywords to clusters.

For each of the five similarity measures that we explained above, clustering analyses were constructed. We performed hierarchical clustering on the similarity matrices, that we got from applying the above similarity measures on our data-set. The clustering evaluation

was used to prove the importance of our new method. We used the CoPhenetic Correlation Coefficient (CPCC), which is the most popular evaluation measure for hierarchical clustering (Rohlf & Fisher, 1968). We used it to compare the clustering efficiency and the goodness of fit of each method. We used the MultiDendrograms program Fernández & Gómez (2008) to get the CPCC for each similarity measure. The result is presented below in Table 3.5. The new method's CPCC is higher than all of the well-known direct similarity measures in bibliometrics science. The CPCC for our new method is 83%, while the CPCC for the association strength, the cosine, the inclusion index, and the Jaccard index are 56%, 53%, 48%, and 57%, respectively. This result shows that the new method has more of an ability to cluster the similar keywords into the same cluster than the other methods. Our new method has improved the ability of the hierarchical method to assign keywords into the right cluster by 48.21% than the association strength, and by 56.60% than the cosine, which is the most used similarity measure in bibliometrics, including 72.91% and 40.61% increases for the inclusion index and the Jaccard index, respectively.

Table 3.5: CPCC evaluation results.

Similarity Measure	Cophenetic Correlation Coefficient
New Method	83%
Association strength	56%
Cosine	53%
Inclusion index	48%
Jaccard index	57%

Keywords with high similarity tend to be in the same cluster. The distance between two keywords represent their relationship: the closer the two keywords, the more similar they are. The size of the nodes will not be presented here because we are going to discuss it in more detail in chapter 4. The top ten keywords with a high frequency of occurrences in the data are are Markov Chain Monte Carlo (440), R (349), Kernel (280), Bayesian

(267), Asymptotic (257), Bootstrap (225), Non-Parametric (201), Gibbs Sampler (199), Non-Parametric Regression (199), and EM Algorithm (189).

Our data divided into 18 clusters. Each cluster has a different number of keywords. The detailed information of clusters is shown in Table 3.6. 3.6 shows that cluster 2 has 68 keywords, making it the largest number of keywords among all the other clusters. We chose the best name that can represent the cluster keywords. Then we chose the top 5 keywords for each cluster as a sample. Remarkably, the total number of ASA journals is 17 journals, and we got 18 clusters, a notable sign that our new method has some merit.

Table 3.6: Clusters of our Keywords data-set.

	Name of the cluster	Terms per cluster	Main Keywords				
cluster 1	Data science	35	Missing Data	Clustering	Longitudinal Data	Conditional	GEE
cluster 2	Expectation Maximization	68	EM Algorithm	Generalized Linear	Measurement Error	Maximum Likelihood	Mixture
cluster 3	Dimension Reduction	18	Functional Data	PCA	Factor	Dimension Reduction	Visualization
cluster 4	Time Series	8	Time Series	State Space	Kalman Filter	Hidden Markov	Stochastic Volatility
cluster 5	Reliability	29	Selection	Variable Selection	Lasso	Shrinkage	Regularization
cluster 6	Bayesian and simulations	22	Markov Chain Monte Carlo	Bayesian	Gibbs Sampler	Hierarchical	Bayesian Inference
cluster 7	statistical analysis and Density Estimation	49	Kernel	Asymptotic	Non-Parametric Regression	Random Effect	Mixed Effect
cluster 8	Design of Experiment	14	Prediction	Gaussian Process	Kriging	Optimality	Spatial Statistic
cluster 9	Hypothesis Testing	26	Goodness of Fit	Spline	Confidence Interval	Calibration	Clinical Trial
cluster 10	Data Transformation	5	Nonlinear	Binary	Non-Parametric Density estimate	D Optimality	Box Cox Transformation
cluster 11	Sampling	11	Bootstrap	Non-Parametric	Monte Carlo	Hypothesis Testing	Panel Data
cluster 12	Statistical Process Control	9	Statistical Process Control	Change Point	Markov Chain	Autocorrelation	CUSUM
cluster 13	Categorical data and psychometrics	8	Multilevel	IRT	Hierarchical Linear	Structural Equation	Capture Recapture
cluster 14	Machine Learning	5	R	Classification	Competing Risk	Discriminant	Machine Learning
cluster 15	Robust Statistics	16	Robust	Regression	M Estimate	Outlier	Robust Estimate
cluster 16	high-dimensional dependence modeling	4	Imputation	Copula	MRF	Penalized Spline	
cluster 17	Unit-Root and Cointegration Tests	9	Selection Bias	Growth Curve	Brownian Motion	Identification	Nonstationarity
cluster 18	Ecology	5	Climate Change	SAS	Pseudo Likelihood	Data	Statistic

3.7 Conclusions And Future Work

3.7.1 An Overview of the Impacts and Contributions of this Paper

The main objectives of this paper were to study the effect of new similarity measure on the output of co-word analysis. As scientific literature continues its near exponential growth, methods for increasing the efficacy of data analytics are needed warranting this present discourse. Hence, the aim of this paper is to introduce a hybrid approach for calculating the similarity measures for keywords in a co-word analysis. The proposed approach integrated the co-citations of keywords' articles within the co-word analysis of keywords to get accurate similarity metrics between keywords. We have presented a new hybrid approach for calculating the similarity between two keywords. We argue that the new approach will show the inner relation between keywords within intended field. That based on that the more data we have of keywords relation the better information we can get. Usually, in bibliometrics science, the similarity measures are used for normalization. Hence, we have studied the most used direct similarity measures for co-occurrence data. The similarity measures are the association strength, the cosine, the inclusion index, and the Jaccard index. Then, we performed an empirical comparison to study the new similarity method and the other four direct similarity measures. We extracted 8,758 articles that were published from 1991-through 2016 in ASA journals. We subsequently studied and compared our new method to the well-known four direct similarity measures. We used keywords as the unit of analysis, and we also explained the importance of keywords to any scientific field. We used the co-word analysis to utilize our analyses, and we applied the hierarchical technique to cluster our data.

The cluster analysis for 485 keywords showed that the research fields of statistical analysis in ASA journals are varied. We have identified 18 clusters, and we believe each cluster belongs to one topic or more. We used the CPCC to evaluate our method and compare it to the old ones. The evaluation result showed a substantial difference between these methods.

Many believe that the CPCC is the best validation method for hierarchical clustering because it shows the ability of the methods to put items into the right group. Future work will focus on improving the visualization of the keywords on a map. The size of the keyword's node presented in the map should correlate to the strength of the keyword in the intended field, instead of only showing the quantity of the keyword. In future work, we will investigate the use of hybrid method on the strength or the size of keywords on a 2D map. Our main contributions and results can be summarized as follows:

- (A) In this paper, we have studied the similarity measures in co-occurrence data and the hybrid method that used in the bibliometrics area. As shown in sections 3.2 and 3.3 introduction and related work and background. In the literature of similarity measures for keyword analysis, only the co-occurrence of keywords used to measure the relatedness between keywords.
- (B) This work is the first to integrate articles' co-citation in co-word analysis to boost the calculation of similarity measures between keywords.
- (C) The result shows that our proposed method improved the clustering of keywords since it revealed the inner relation between keywords.

3.7.2 Practical Implications from our Work

Every new researcher or scholar is eager to have the tools that will help in understanding a field of interest within a reasonable time frame. These tools need to have the ability to analyze and visualize huge amount of scientific literature data. The development bibliometrics science gives researchers the hope of using these tools and techniques to expedite and facilitate the process of comprehensively perceiving a given field of study. As a result, this paper aims to help researchers capture the evolution and development of a given field by using the co-word analysis approach to cluster keywords better.

3.7.3 Limitations and Future Work

Despite the outstanding results of our proposed method, there are a number of limitations that need to be highlighted. First, we have only applied our method on one scientific field, statistics. However, we believe that if our method has a great success with a multidimensional scientific field such as statistics, then we anticipate an excellent output on a simpler scientific field. Second, applying our proposed method on other unit of analysis may not have the same output since the characteristics of keywords are different than other units of analysis. Third, we have faced difficulties in collecting and preprocessing our data due to the lack of standardization between publishers and resource platforms. For future research, there are two potential opportunities: first, researchers can study the effect of similarity measures on each units of analysis and the impact of them on the output. Second, researchers can implement more data into similarity measure, such as citation time, prestige, and popularity of articles.

Chapter 4

Qualitative Weighted Keywords for Clustering Analysis in Bibliometrics Data.

4.1 Abstract

With the dilemma of measuring the scientific production and overwhelming information available online these days, new researchers and scholars find it difficult to grasp the vast storehouses of information within a given scientific field. For that reason, the visualization of bibliometrics data is a helpful way to have an insightful look at a field through the visualization of its data. However, when it comes to using keywords as unite of analysis, one of the main flaws of visualization is that does not represent the importance, or the influence, of the keywords on the 2D map; instead it visualizes the number of keyword appearances in its data-set, which is represented in the map by the node size. Hence, the aim of this paper is to introduce a hybrid approach for calculating the importance of keywords. The proposed approach integrated four aspects of the keyword: number of appearance, citation time interval, popularity and prestige. This leads to make the node size of a keyword on the 2D map more informative, where the size of the keyword represents its strength or influence on the data-set.

4.2 Introduction

The main motivation for this task is to construct a way to visualize the quality of a unit of analysis for scientific papers. Measuring scientific production has raised a lot of criticism. The justified way to evaluate the individual papers or research is by reading and understanding the work. The disadvantage of this way is that it is time-consuming and dependent upon the reader's expertise. The first acknowledged attempt to measure

science was made in 1873 by Alphonse de Cansolle in his book “*Histoire Des Science et Des Savants Depuis Deux Siecles.*” he studied how the environmental factors of scientific society memberships affect the scientific strength of nations (De Candolle, 1885). In 1926, Lotka published his work “*The Frequency Distribution of Scientific Productivity*” in the Journal of Washington Academy Sciences. He introduced Lotka law for author productivity (Lotka, 1926).

In 1927, Gross & Gross (1927) introduced citation counting, the aim of which was to show the importance of citation in evaluating a scientific paper. Since its inception, citation counting has had a tremendous influence on evaluating scientific articles, journals, institutions, domains, and countries (Yan & Ding, 2010). After that, in 1955, Garfield proposed the journal impact factor, which is an indicator for evaluating the average number of citations per published article within a given journal (Garfield et al., 1964). To satisfy such a need, bibliometrics science was developed. Bibliometrics measures the degree to which scientific performance of knowledge adds to science every day (Walter et al., 2003).

Applying bibliometrics analysis on any scientific field is for equipping scholars or researchers with a coherent knowledge of the intended field, within a reasonable time frame. At the same time, an ancillary goal is to provide depth perception for the connections between units of analysis (e.g. authors, keywords, documents, and so forth). The main focus of bibliometrics science involves the collection of data to present information in an insightful way. Bibliometrics tools and methodologies transform bibliographic data into formats that are more intellectually digestible.

By way of overview, Börner et al. (2003) used the following sequential steps to online the procedure behind Knowledge Domain Visualization, which is also interchangeable with “bibliometrics” or “scientometrics.” (1) data extraction,(2) definition of unit of analysis, (3) selection of measures, (4) calculation of similarity between units. and (5) data visualization and analytics. Data extraction is the first step of constructing a visual representation of the intended subject and comprises the collection of bibliographic. Bibliographic literature

presents the background needed for identifying the measures and for representing the relationships between them using graphs and networks. Next, the step of defining the unit of analysis consists of two parts. The first part is to choose the unit of analysis that the one wants to study and present in a visualized map. The second part of defining the unit of analysis is preparing the units for analysis, a process that consists of cleaning and making data ready for analysis; it is this preparation phase that raises the efficiency and effectiveness of any applied analysis. This step relies heavily on methods from text mining (Rodrigues et al., 2014). Text mining employs natural language processing (NLP) methods to extract bibliographic data from the publications and preprocesses the unit of analysis. In this paper, we used NLP to preprocess keywords, our unit of analysis. We used the stemming technique to convert keywords into their basic (Francis & Flynn, 2010).

The third step of bibliometrics is the selection of measures. The main goal of this step is to measure relatedness between the unit of analysis. The fourth step of bibliometrics visualization is the calculation of unit similarity. A similarity measure is one of the essential steps of visualizing the bibliometrics outcomes, especially since it reflects the relation between two items. For that reason, choosing an appropriate similarity measure usually results in an accurate visualization for the data-set (Huang, 2008). The relation between two items is represented graphically in a 2D map where the distance between the items reflect their similarity. The smaller the distance indicated, the stronger the relation between the items. In this paper, the items correspond with our original keywords. The fifth step of bibliometrics is data visualization and analytics. The main goal of this step is to represent the output of data visually to the user. Data visualization, be it interactive or non-interactive, is a representation of information on a graphical platform that helps the user to examine, explore, discover, comprehend, and analyze a large amount of data in a short period of time (Börner et al., 2003; Khan & Khan, 2011). This step includes grouping similar items into clusters, laying out the items on a 2D space to represent the relationship between them. Present vast

amount of data visually makes the information of inquiry easier to understand aids in the process of analyzing it effectively.

In bibliometrics science, there has been many attempts to integrate quantitative and qualitative methods to create innovative approaches to outperform old approaches. there are a few hybrid works that have been done in the bibliometrics research. For example, one of the first attempts to integrate the quantitative and qualitative method in bibliometrics was introduced by Braam et al. (1991) when they introduced their approach to co-cluster documents by combining co-citation and word analysis. Then Hammouda & Kamel (2002) proposed a new approach to improve web document clustering by combining semi-structures inherent in web documents, document index graphs, and phrase-based similarity measures. Chim & Deng (2008) introduced the concept of using a suffix tree to cluster new documents based on phrase-based document similarity. Another researcher proposed a new approach to automatically construct a term taxonomy and the relation of terms based on weighted keywords co-occurrence (Li et al., 2015).

In light of other attempts at hybridization, we have pushed forward in the spirit of innovation, the need for new methods and techniques to expedite and facilitate the process of perceiving an overview of a scientific topic, specifically with respect to volume magnitude and visual data representation. At the present time, the bibliometrics mapping of science is a quantitative method for studying bibliographic data (e.g. titles, keywords, authors, and so on) and visually representing the information. Furthermore, when it comes to dealing with a large body of literature, bibliometrics maps are exceptionally useful and have been used in different contexts like research literature surveys, government decision-making, and scientific publications. The aforementioned considerations are precisely why we saw an opportunity to improve visualization of keywords' node on a 2D map.

Mapping a scientific field is useful method for gaining significant insight into a scientific field. This is especially the case when mapping from a macro-level to a micro-level. Keywords is one of the most used units of analysis, so knowing its importance can help researcher

gain the insight s/he seek. With respect to visualizing bibliometrics data, the clustering analysis is an efficient approach for visualizing the data and compared to other approaches, it is easiest to understand. Usually the weight of keywords corresponds to the frequency of each keyword in the data-set, which implies that keywords that appear more frequently are more substantial to the field. Nonetheless, we argue that counting is not a suitable measure for several reasons. First, scholars put keywords into their articles to catch the eyes of researchers, but in reality they are not always maximally relevant to their research. Second, some keywords could fall into different subfields of the field being investigated, which can cause confusion since some keywords have different content based on their subfield. In addition, the quantity of a keyword does not represent its quality or influence in the data-set since some keywords are used to increase the searching of an article. These flaws raised the need to have better visualization of keywords.

Given this, we will introduce in this study, a new qualitative method for visualizing the weight of a keyword in clustering mapping. We adopted the concept of journal and article popularity and prestige and applied it to keywords, where keyword status can be defined by four factors: keyword appearance, which is the number of times it appears in the data-set; keyword popularity, that is, the number of citations its articles received; keyword prestige, namely, the article influence scores of the articles that cite the keyword's articles; and the citation interval time, the time between the publishing and citing of the keyword's articles.

Bollen et al. (2006) explained the difference between prestige and popularity. For example, let's compare two authors. One author has received a Nobel Prize in literature and is highly respected by colleagues, but he is not known for any bestselling works. The other has a spectacular sales rate but lacks the accolades of the first. Concerning prestige and popularity, the first author is referred to as prestigious author, while the second author is referred to as popular author (Bollen et al., 2006). In bibliometrics science, academicians consider articles or journals to be popular if they garner a lot of citations, but the truth is that the number of citations does not necessarily signify a work's level of importance.

Kruskal & Wish (1978) were the first to note the difference between popularity and prestige within bibliometrics.

Basically, the popularity of an article or journal is measured by the number of citations it receives, even if the citations are unweighted. For example, even though they make no real contribution to a given field, review articles usually receive a high number of citations because they are referenced when an author conducts a literature review. On the other hand, the prestige of a journal or article is measured by the number of times it is cited by a prestigious citing journal or article. For example, getting cited by journals like *Nature* or *Science* is valued much more than being referenced by a journal with a low impact factor (Franceschet, 2010). The gauging both the popularity and prestige of an article is an ideal way of finding its real impact in the respective scientific field. However, newer articles simply cannot compete with old articles. For that Sayyadi & Getoor (2009) introduced FutureRank for articles to give it a fair chance to pop up. This indicator gives weight for citing time. Also, considering the time of the citation is a reliable indicator of its importance (Sayyadi & Getoor, 2009).

Table 4.1 below shows the most used visualization tools in bibliometrics mapping (Belter, 2012), where all of them use the counting of the unit of analysis as the size or weight of the item's node. Despite this, our method is designed more specifically for keywords to be the unit of analysis. The rest of this paper is organized as follows: section 4.3 where we study the related work and background. Then, section 4.4 , where we describe the method. Next, in section 4.5, we introduce a empirical comparison. Then, in section 4.6, we present a case study for our method. Finally, in section 4.7 is the conclusion.

Table 4.1: Visualization tools used bibliometrics mapping.

software	Types of Bibliometrics Maps	Method	References	website
VOSviewer	distance-based maps	counting of items	(van Eck & Waltman, 2009)	http://www.vosviewer.com/
Science of Science (Sci2) Tool	graph-based maps	counting of items	(Team, 2009)	https://sci2.cns.iu.edu/user/index.php
Citespace	graph-based maps	counting of items	(C. Chen, 2006)	http://cluster.cis.drexel.edu/~cchen/citespace/
NodeXL	graph-based maps	counting of items	(Smith et al., 2010)	http://nodexl.codeplex.com/
NetDraw	graph-based maps	counting of items	(Borgatti, 2002)	https://sites.google.com/site/netdrawsoftware/home
Cytoscape	graph-based maps	counting of items	(Smoot et al., 2011)	http://www.cytoscape.org/
New Method	distance-based maps	Frequency, citation time interval, Popularity and Prestige		
	graph-based maps			

4.3 Related Work and Background

Generally speaking, we can divide indicators in bibliometrics, (also known as scientometrics, information systems or information science) into three levels (Yan & Ding, 2010): journal-level, article-level, and author-level. Also, there are two types of indicators: altmetrics and metrics. Altmetrics, Short for “alternative metrics,” measure the impact of bibliometrics meat-analysis on online sources such as websites, social media platforms, and blogs. Typically, they use downloads or click-through as an indicator for impact of the meat-analysis (Galligan & Dyas-Correia, 2013). There are several altmetrics available online. For example, at the journal-level of altmetrics there are two well-known altmetrics that can be found on this website <https://www.altmetric.com/>. The Altmetric explorer and PLoS impact explorer collect feedback on journals from social media platforms, blog websites and other websites to measure the impact of a journal online (Brigham, 2014). At the article-level there are Plum analytics <http://plumanalytics.com> which gather data about articles from different online sources, and ReadMeter <http://readermeter.org>, a tool for measuring the impact of articles and authors within a scientific field by tabulating the approximate readership. At the author-level, in addition to ReadMeter, there is CitedIn, which measures author citation feedback from numerous online sources.

The second indicator type is metrics, also known as traditional metrics, which measure the impact of bibliometrics meat-analysis by studying citation weights. It also has three levels, journal-level: article-level, and author-level. Metrics at the journal-level measure the impact and quality of a scientific journal on its field. The higher the impact of a journal the higher its prestige and endorsement in its own field or on the science in general. There are several indicators that measure the impact of a journal. The impact factor is used to measure journal prestige (Garfield, 1999). The most used and noticed one is the Impact Factor (Garfield, 1999), which was introduced by Garfield in 1999, and has become the most respected measure for journals and widely used in the world (Bordons et al., 2002; Nederhof et al., 2001; Bornmann et al., 2011). Bornmann et al. (2011) introduced the PageRank

method in bibliometrics, which is an adaptation of Google's PageRank method and is used to evaluate the scientific impact of a journal by measuring its use online (Bollen et al., 2006, 2009). Similar to the PageRank indicator is the eigenfactor, a way of assessing the impact of a journal by the amount of citations it receives from highly ranked journals. The eigenfactor is available for free on <http://www.eigenfactor.org>. Additionally, there is SCImago, this measures the productivity and prestige of journals (García-Peñalvo et al., 2010).

The purpose of article-level metrics is to study the impact of a single article on its scientific field or on science in general. Yan & Ding (2010) proposed a new method for measuring article prestige (Yan & Ding, 2010); they used the weighted citation of an article's journal and citation time to measure an article's prestige. Also, Google's PageRank has been adapted to measure the impact of an article (P. Chen et al., 2007). In addition, CiteRank is a derivative of Google's PageRank, which accounts for newer articles having higher chance of being cited (Walker et al., 2007).

At the author-level, since Hirsch (2005) has proposed the h-index as a single number to evaluate author production, the h-index has become an important indicator for author productivity. Since then, a debate about the h-index's efficacy ensued, and many have proposed either similar indexes or modified ones (Jin, 2007; Sidiropoulos et al., 2007). Bornmann et al. (2011) have studied the h-index and the other indices. He reported 38 indices that have been presented as indicators of author output. Besides that, there have been different methods for determining author output, as well as its quality. For example, Schubert et al. (2006) studied the impact of author self-citation on the author's overall production. Also, Pan & Fortunato (2013) introduced the Author Impact Factor (AIF), which measures the dynamism of author production in a period of time. In 1990, Egghe & Rousseau (1990) suggested a different way to count citation for the first author, and co-authors. The eigenfactor score, which is mostly used at the journal-level, was adapted by West et al. (2013) to propose author-level eigenfactor metrics.

Bibliometrics literature states that the most common units of analysis employed within bibliometrics are journal, article, author, and concept “term” or “keywords” (Börner et al., 2003). We found there are no term-level metrics to measure a term's impact on its field, data-set, or even science in general. For that reason, we propose a new method to measure the impact of keywords and present it at a 2D map. We going to use co-word analysis to show the relation between keywords and to visualize them. The first proposition for a co-word analysis was made by Callon et al. (1986) in their book “*Mapping the Dynamics of Science and Technology*” in 1986. Since the appearance of this book, co-word analysis has become an important method in bibliometrics science. Co-word analysis is a content analysis technique that uses patterns of concept pairs co-occurrence (i.e., words, items, noun phrases) in a corpus of texts or keywords lists to identify the relationships between the concepts within the subject areas presented in these texts (He, 1999).

4.4 Method

In this paper, we will introduce the concept of measuring the importance of a keyword by evaluating its appearance, citation time interval, popularity and prestige. This method aims to weigh the strength of keywords within the clustering visualization based on four elements: frequency of keywords, citation time interval, popularity, and prestige. The frequency is the number of a keyword's appearance within a data-set. The citation time interval is the timespan from the publication of the keyword's article to the publication of the citing article. The popularity of a keyword can be measured by the number of times its articles have been cited. The prestige of a keyword can be quantified by the number of citations from highly cited publications (Yan & Ding, 2010).

That being said, the keywords' size “strength” on the 2D map of clustering, will not be presented by its frequency in the data-set only, but it weight will also be determined by the four elements to have a better presentation for keywords on the map. For example, when a keyword's article is cited by a highly prestigious journal, it has greater importance that

when cited by a journal of lower prestige. Also, the time factor is crucial for measuring the importance of the keyword's impact on its field, as well as its dynamics. This means that the keywords' strength will change from year to year based on the previously mentioned elements. This would help scholars to see the evaluation of the importance of keywords through an interactive visualization. To add to this, by differentiating among self-citations, grad student citations, and prestigious journal citations, the new method will strike a balance between popularity and prestige. We should mention that the proposed method does not intend to rank keywords, but it does intend to give keywords a better representation in a visualized map.

Now, we are going to explain the four elements that influence keywords strength in our new method. The keyword is represented by kw_i ; let n be the total number of i . The keyword's strength or influence is represented by ks . The frequency of keywords is the number of keyword appearances in the data-set, which is represented by f ; we extracted the number of appearances from our own data-set. Also, not all articles are cited; we represent the total number of cited article by f' . We assume that f equals the number of articles that holds keywords in its own keywords' list, and it is represented by a .

$$f_i \geq f'_i \text{ for all } i = 1, \dots, n \quad (4.1)$$

The popularity of a keyword is the summation of citations each of the keyword's articles received, represented by p .

$$p_i \geq f'_i \text{ for all } i = 1, \dots, n \quad (4.2)$$

Also, p_i equals zero only when none of the articles received any citations.

$$p_i = 0 \Rightarrow f'_i = 0 \quad (4.3)$$

The keyword prestige is driven by the eigenfactor score. The eigenfactor score is a measurement for scientific journal importance over a period of 5 years, by rating journals based on the incoming citations from highly ranked journals. Then we got the article influence (AI) score for each citing journal from which the eigenfactor is derived. The article influence score is calculated by dividing the eigenfactor over the number of articles published by its journals over a 5-year span. We got the article influence scores from (www.Eigenfactor.org).

Next, citation time interval plays a significant role in measuring the impact of meat-analysis in bibliometrics. For example, the sooner an article gets cited, the more importance it gains relative to other articles in the same field. In addition, it shows that the article has higher influence than other articles, and that could happen for multiple reasons such as a breakthrough in its field or the author is a very trusted scientist (Yan & Ding, 2010). At the journal-level the immediacy index is used to determine citation time for articles in the journal. At article-level, it used to determine the citation time for each article to be cited. We should consider giving different weight according to the length of time between the citation and publishing date of a given article. Citations for articles decrease exponentially. This has been proven by Sayyadi & Getoor (2009), who introduced the FutureRanking method. The data fit the trend curve: $f(x) \sim e^{-0.117x}$. The citation time interval gives a great indication for new scholars and researchers to notice the emerging keywords or concepts in their field or subfield. The year of publication of the keyword's article is represented by t_p , and the year it got cited by t_i . The x in the trend curve is equal to the difference between the cited year and published year. Figure 4.1 below, shows the time-line of cited articles, citing articles, and citing journals.

$$x = t_i - t_p \tag{4.4}$$

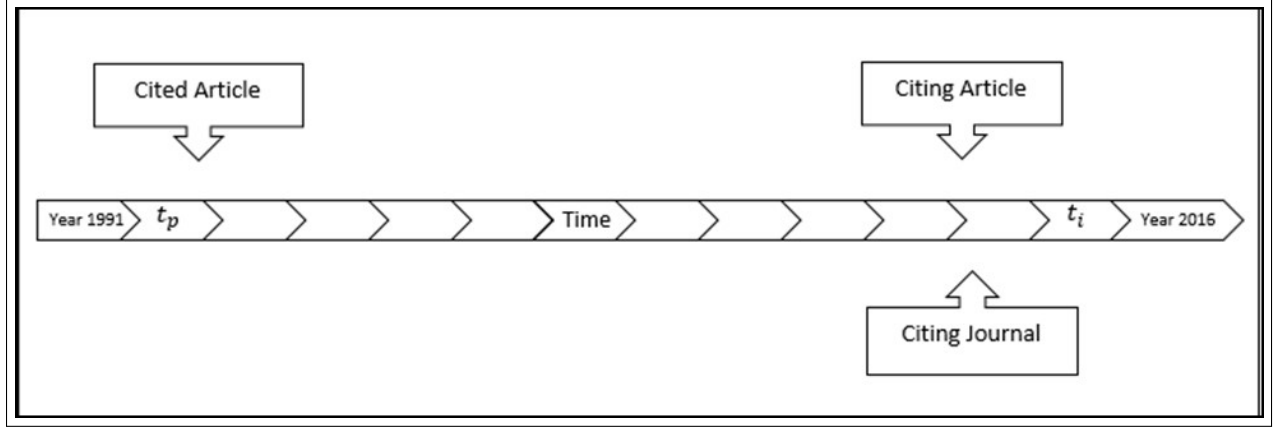


Figure 4.1: Timeline of cited articles and citing articles.

The new method in equation 4.5 . The keywords strength is weighted among four elements.

$$KS = \sum_{i=1}^p \left(\frac{1}{e^{f_i}} \right) * AI_{j_i} * e^{-0.117*(t_i-t_p)} \quad (4.5)$$

Below is a mathematical notation for our new method.

- KS = Keyword Strength.
- p_i = Total number of citations that all a_i received.
- f_i = Total number of appearance of keyword i .
- f'_i = Total number of a_i got cited.
- a_i = Article that contains keyword i in its keyword list.
- a'_i is the citing article of a_i .
- AI_{j_i} = The article influence score for citing journal j for keyword i .
- t_i = Year when keyword i article got cited.
- t_p = Year when keyword i got published.

Figure 4.2 shows the model to calculate the keyword's strength for one keyword. The figure shows the flow of data for a keyword. It shows the steps that needed to gather data for one keyword, from keyword's list in its article to the citing article.

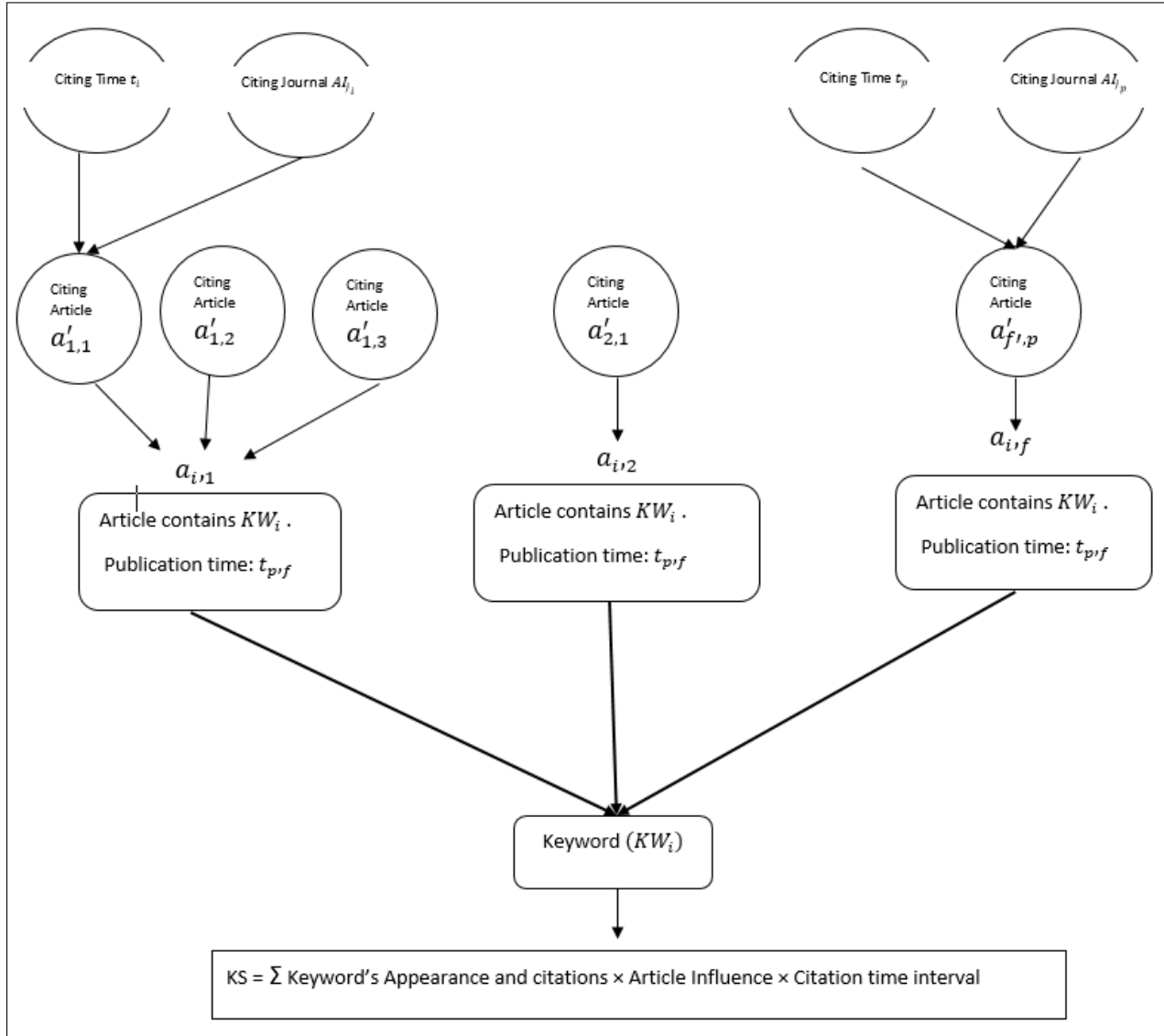


Figure 4.2: Flow diagram of our method to calculate keywords' strength.

4.5 Empirical Comparison

After describing the proposed method, we present an example to illustrate how it works. Let us consider two keywords, for example, "Control Chart" and "P value". Both of the two keywords appeared 10 times in our data-set. The keyword "Control Chart" (KW_1)

appeared in 10 articles ($a_1 = 10$ and $f_1 = 10$) and all of the articles got cited ($a'_1 = 10$ and $f'_1 = 10$) each article receiving two citations ($p = 20$). We are going to show the calculation for the first step, and the table 4.2 below has all of the information. The publishing time for the article is 2013, ($t_p = 2013$.) The citing article cited the first article in year 2015 ($t_i = 2015$), and its journal J_A , and the Article Influence score for the citing journal in year 2015 is $AI_{ji} = 0.5$. The keyword's strength for keyword "Control Chart" is a sum up of all the scores for each of the citation its articles received, which in our example just 20 times. For the scores of the Article Influence score of the citing article's journal, and the Citation time interval, the $e^{\frac{1}{10}}$, and the AI = 0.5, and the citation time interval: $e^{-0.117 \times (2015-2013)}$. After calculating the strength or influence of the two keywords we got KS for "Control Chart" = 56.52, and "P value" = 11.53. We believe that these two numbers are more representative and informative of the old method where these two keywords node sizes, 10, will be equal on the 2D map. However, in our method, we perceive the strength and the influence of the first keyword over the second one by looking at the strength.

We considered more factors to have more informative visualization. Thus, we looked more into the keyword to get more information out of it instead of just frequency. For that reason, to get a better understanding of the influence or the strength of the keywords on the data-set, we looked at each one of the articles that the keywords appeared in and noted number of times each article got cited, who cited it and when.

Table 4.2: Two keywords' example data

Keyword "Control Chart"	publication time	Citing time	AI	Keyword "P value"	publication time	Citing time	AI
kw_1	t_p	t_i		kw_2	t_p	t_i	
$a_{1,1}$	2013	2015	0.5	$a_{2,1}$	2010	2013	0.5
		2013	1			2013	1
$a_{1,2}$	2013	2013	2	$a_{2,2}$	2010	2013	2
		2014	3			2014	3
$a_{1,3}$	2013	2014	4	$a_{2,3}$	2010	2014	4
		2014	0.5			2014	0.5
$a_{1,4}$	2013	2014	0.5	$a_{2,4}$	2010	2014	0.5
		2015	0.5			2015	0.5
$a_{1,5}$	2013	2015	3	$a_{2,5}$	2015	2015	3
		2015	20			2015	2
$a_{1,6}$	2015	2015	16	$a_{2,6}$	2015		
		2015	11				
$a_{1,7}$	2015	2015	3	$a_{2,7}$	2015		
		2015	0.5				
$a_{1,8}$	2015	2016	0.6	$a_{2,8}$	2015		
		2016	0.1				
$a_{1,9}$	2015	2016	0.4	$a_{2,9}$	2015		
		2016	0.5				
$a_{1,10}$	2015	2016	0.6	$a_{2,10}$	2015		
		2016	0.9				

Below, we going to illustrated the steps to calculated the keyword's strength. The workflow of our proposed method is shown in figure 4.3, We will take the keyword “chart control” as an example. First, we extract all the records of the ASA that are in the WOS website. Next, we choose our unit of analysis which is the keywords list. Then, we perform text mining and preprocess on the data. Set up a threshold is very important test point to get only the items that meet your criteria, and to have informative visualization. Then, conduct the co-word analysis for the keywords. after that, choose the keywords that you want to get its strength or important or the top 100 keywords for example. Next, get each article that contains the keyword, for example “chart control” We calculate the total appearance of the keywords and the total number of citations for all of the articles, if one of the article has no citation then move to the next one. From the cited article, we get the publication year, and we get the citing article (the article that have cited the article that contains the keyword) data, we need to get the publication year of the citing article and the Article Influence of its journal. For each time of the citation we apply our new method and at the end we sum them up to get the keyword's strength, then we move to the next keyword.

1. Extract the bibliographic data from the WOS.
2. Data mining and preprocess of that data.
3. Choose your unit of analysis (keywords).
4. Set up the thresholds for your keywords.
5. Construct a co-word analysis for the keywords.
6. Get all the articles that contain the keyword “chart control”.
7. Get the article meta-data; such as
 - publication year
 - total appearance

- sum of citation.
8. Count the total number of appearance.
 9. Count the number of articles that have got cited.
 10. Sum up citations.
 11. Get the citing article information, such as
 - Publication year
 - Journal AI
 - Calculate the keyword strength for each citation.
 - Sum up the keyword strengths of each citations.

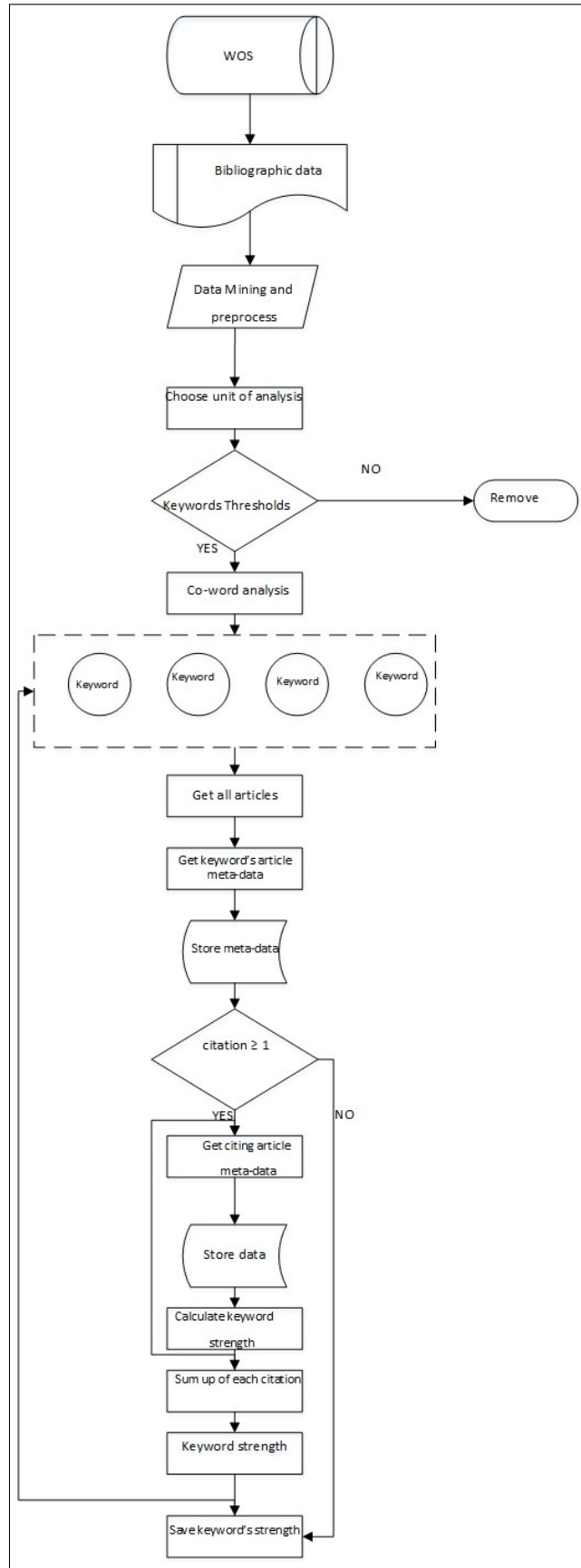


Figure 4.3: Flow chart of our method.

4.6 Case Study

In this study, we used the data-set of the ASA from 1991 to 2016. All articles and their full records were collected from Web of Sciences (WOS). In total, we retrieved 10,618 records with 136,762 Citations. We omit all other document types such as proceedings paper, biographical-item, book review, correction, editorial material, letter, meeting abstract, note, reprint, review, and software review, and we kept only articles with keywords. We were sure to note keywords frequency, leading to the identification of 17,766 unique keywords. Also, we collected the citing articles for each keyword and notated their journals' as well as publishing times.

Usually, in co-word analysis, the frequency of the keyword reflects the size of the node (keyword) for visualizing the outcome of the clustering analysis. This measurement does not reflect the impact of a given keyword on the intended field. This old method does not consider the importance of other factors that affect keyword importance. For that reason, we are going to refer to the size of the keyword “node” on the mapped clustering as the strength of the keyword because it shows the impact and influence of the keyword instead of its size.

We applied our method on four keywords that have the same number of appearances, 15. The aim is to study the effect of the appearances, citation time interval, popularity and prestige on each one of them. Overall results of the methods are synthesized in Table 4.3. Appearance, citation time interval, popularity and prestige measure four aspects of a keywords. Prestige, represented by the AI, increases the importance of a keyword by 25%. For example, if keyword 1 has an AI of 1, keyword 2 has an AI of 2, and all of the other elements are equal, we can state that the keyword 2 has 25% more influence than keyword 1. In Table 4.3 , we can see that the keyword “central limit theorem” has the highest average of AI, and keyword “variance reduction” has the lowest AI. On the other hand, keyword “multinomial distribution” has a better

average AI than keyword “variance reduction,” and this affects the final result by ranking “multinomial distribution” higher than “variance reduction”. We can state that the prestige of a keyword affects the weight of the citing article, which causes an increase in the keyword strength. Popularity significantly affects the keyword's strength as well. In fact, the effect of the popularity is around 1%. Also, the citation time interval has an effect on the strength of a keyword from 10% to 15%. The less the citation time interval for a keyword, the more influence it has.

Table 4.3: Data of the four keywords data

Keyword	No. of appearance	No. of cited articles	No. of citations	Average citations	Average citations/ cited articles	Average AI	Average time interval	Keyword's strength
central limit theorem	15	12	116	7.73	9.67	1.57	7.91	90.88
constrained optimization	15	11	64	4.27	5.82	1.42	8.72	31.71
multinomial distribution	15	13	205	13.67	15.77	1.45	6.69	156.15
variance reduction	15	12	377	25.13	31.42	1.15	11.32	129.61

In the Table 4.4. below, we ranked the four keywords based on the data we got. The keyword “multinomial distribution” has the highest strength between the four keywords, and this is due to its low average citation time interval. Moreover, it has the second highest of amount of citations and average of citations. As we mentioned above this method does not intend to rank keywords but focuses on giving the user better visualization on the 2D map. However, we are showing the ranking to give and explain our idea more perfectly.

Table 4.4: Ranking of four keywords data

Keyword	Rank of appearance	Rank of cited articles	Rank of citations	Rank of Average citations	Rank of Average citations/ cited articles	Rank of Average AI	Rank of Average time interval	Rank of Keyword's strength
central limit theorem	1	1	3	3	3	1	2	3
constrained optimization	1	4	4	4	4	3	3	4
multinomial distribution	1	1	2	2	2	2	1	1
variance reduction	1	2	1	1	1	4	4	2

The figure 4.4 below shows the difference between the old method of keyword node size and our proposed method. The red color represents the node size based on the old method and the green color represents the node size of our method. We can see that the size of the keyword node is more informative than in the old ones. From

just looking at the visualization we can see the importance of keyword “constrained optimization” over the other keywords, especially on keywords “central limit theorem” and “constrained optimization”. The figure 4.4 below is created for this example only, to show the difference between the two methods in a real visualization tool.

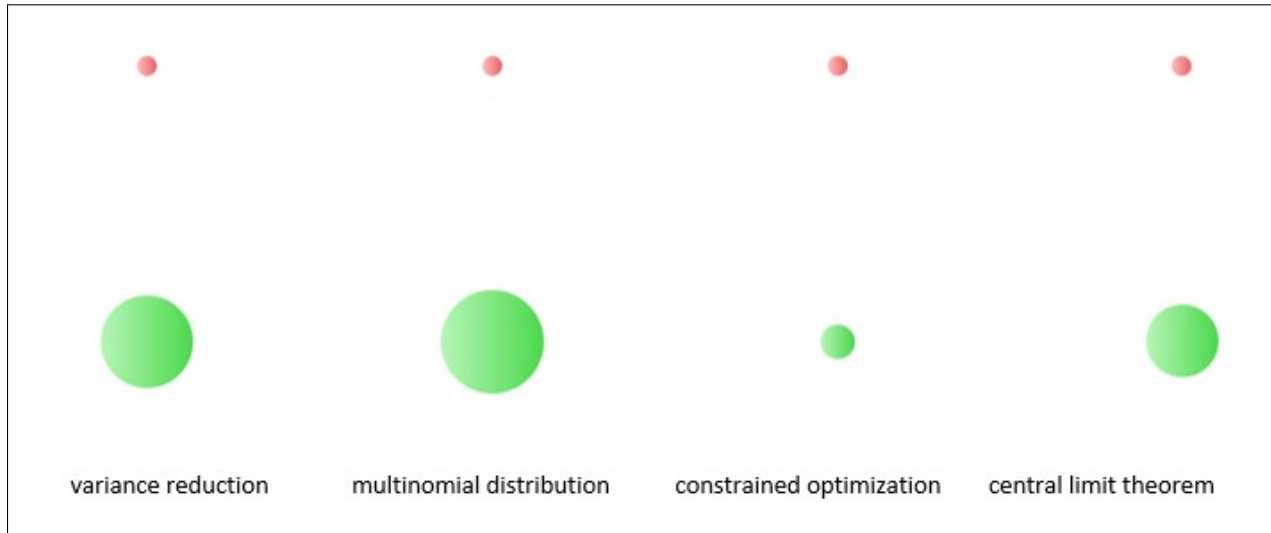


Figure 4.4: New method keyword's strength versus old method node size.

For more explanation and simplicity, we applied our method on the top 100 keywords of the ASA from 1991 to 2016. After calculating the keyword's strength for each keyword. Figure 4.5 shows how the clustering of keywords when we have the number of appearance as the node size representer, and figure 4.6 shows how the clustering of keywords when we have the strength of keywords as the node size representer. For example, let's compare the keyword “Variable selection” nodes size in figures 4.5 and 4.6, the red arrow points its place on the maps. In the first figure 4.5, the node size is barely recognizable. However in the second figure 4.6, the node size is very recognizable for the user. Also, if we look at the table 4.5 below, we see that the keyword “Variable selection” is ranked in sixth place between keywords based on the keyword's strength and ranked the 21st based on the keywords appearances. For deep analysis, we looked at the keyness value of the keyword “Variable selection” in our data-set from 1990 to

2016, (see more explanation about keyness in section 2.5.1), Figure 4.6 shows that the keyness value of the keyword is increasing exponentially.

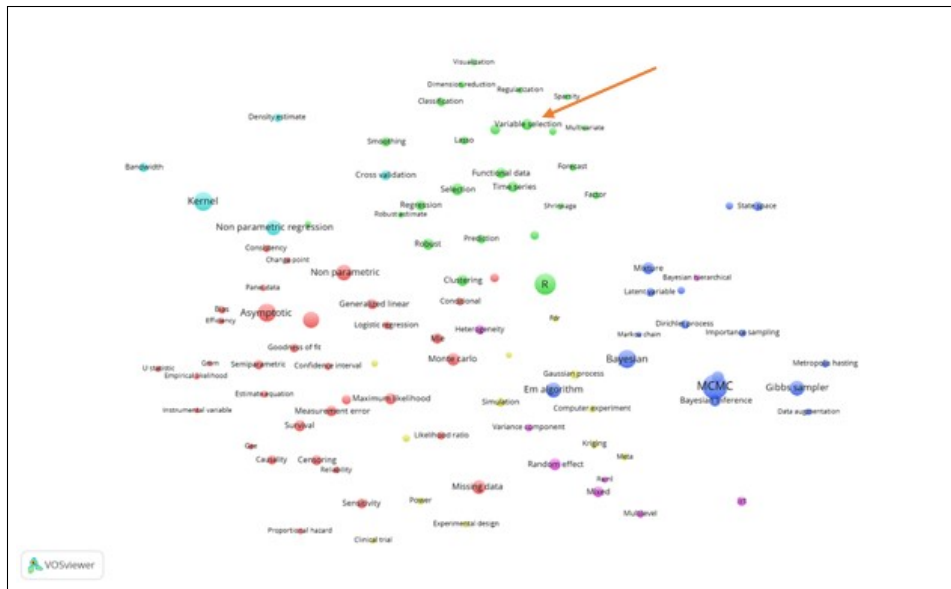


Figure 4.5: The node size represented by the frequency of appearance in the data-set.

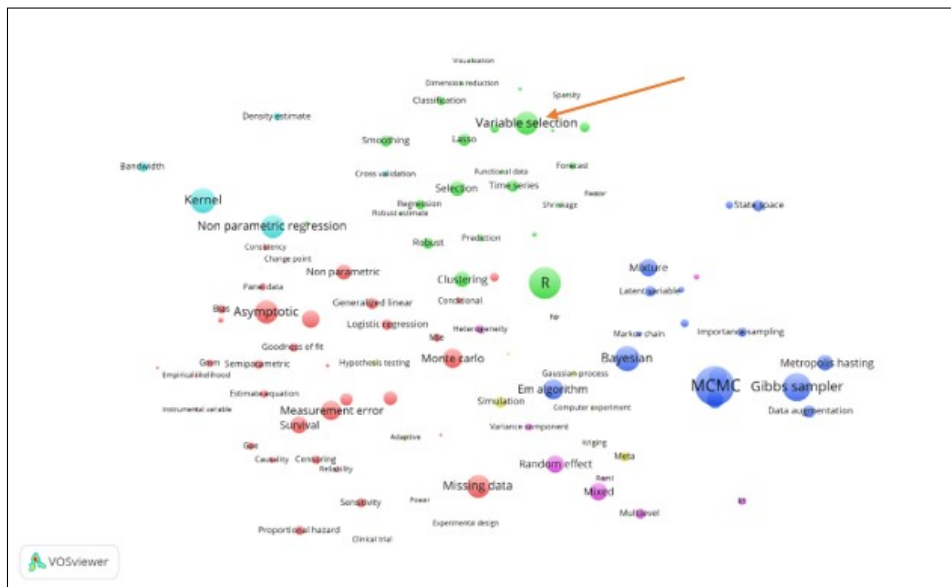


Figure 4.6: The node size represented by keyword's strength.

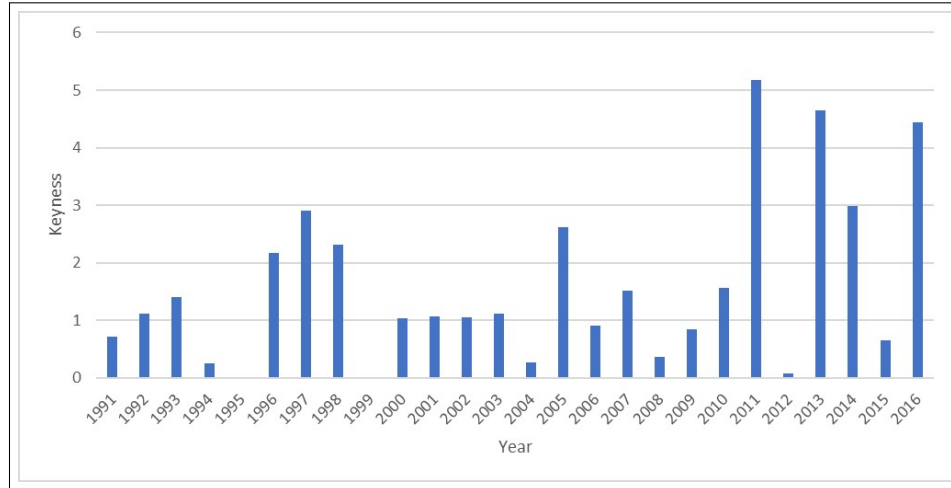


Figure 4.7: Keyness values of the keyword Variable selection.

In table 4.5, we ranked the top 20 keywords based on our new method. The table shows the keywords' name, strength value, number of appearances, rank based on the strength of the keywords, and rank based on the keywords' appearances in the data-set. The old method has many flaws. For one, node size is a representation of the keyword number of appearances in the data-set, which is not value-added information within the visualization. Also, the quantity of a keyword does not represent its quality or influence in the data-set since some keywords are used to increase the searching of an article. There are several advantages to using the new method in visualizing keywords on a bibliometrics map. The size of the node that represents a specific keyword correlates to its strength within the data-set. For example; citations time intervals helps new researchers and scholars to see the emergence of important keywords, because the keywords that have a better citations time intervals are ones that are trending and are getting cited faster than the other keywords, even if they appear the same amount of times or less in the data-set. Also, a keyword with a high AI employed in articles that are cited within prestigious journals, indicating a topic that has a higher tendency to be trending in the near future, even if it has low appearances in the data-set.

Table 4.5: The top 20 keywords based on our method.

Keyword	Keyword's Strength	Number of Appearance	Keyword's Strength Rank	Number of Appearance Rank
MCMC	4880.76	440	1	1
R	3394.65	349	2	2
Gibbs sampler	2739.92	199	3	8
Kernel	2318.49	280	4	3
Bayesian	2248.02	267	5	4
Variable selection	2049.47	118	6	21
Asymptotic	2035.69	257	7	5
Non parametric regression	1971.80	199	8	8
Missing data	1969.01	166	9	11
Em algorithm	1572.75	189	10	10
Monte carlo	1527.94	145	11	12
Measurement error	1478.01	111	12	25
Hierarchical	1330.42	136	13	13
Bootstrap	1291.43	225	14	6
Mixture	1287.55	131	15	15
Random effect	1282.60	130	16	16
Mixed	1243.08	112	17	24
Survival	1209.44	121	18	19
Bayesian inference	1136.10	122	19	18
Metropolis hastig	1090.04	67	20	55

4.7 Conclusions And Future Work

The main objective of this paper was to integrate four new factors into measuring the strength of keywords on a 2D map. The biggest challenge of data visualization is to show a huge amount of data on a small canvas and presenting it in informative manner. In the old method, the number of keyword appearances in a data-set is used to represent the size of the keyword's node on a 2D map visualization, which is not very informative to the researchers. Hence, we proposed a novel method to visualize keywords of co-word analysis on 2D map. We introduced four factors to measure the strength of keywords. The first factor is the citation time interval, which is represented by the difference between the publication year of the keyword's article and the time of the citing article publication year. The second factor is the popularity of a keyword, which is represented by the number of citations its articles received. The third factor is the prestige of a keyword, which is represented by the Article Influence score that the citing article has. Finally, the fourth factor is the total number of keyword's appearances. We argue that the new approach gives an important rule for the keyword's size on a 2D map.

Chapter 5

Conclusion and Summary of Dissertation Contributions

The main objective of my dissertation is to study bibliometrics science and to develop ideal methods to facilitate the process of examination and understanding of scientific literature in an intended field, especially with the enormous availability of literature nowadays. A full study of the literature of ASA journals was explored from macro-level to micro-level, and the findings provided a helpful insight into the field of statistics. In the next chapter of the dissertation, a new method was introduced to determine an appropriate similarity measure between keywords in the co-word analysis. After a deep study of the literature of the similarity measures, I can say that there are two main types of similarity measures used into the co-word analysis; the direct and the indirect similarity measures. It has been concluded that direct similarity measure is more appropriate for the co-occurrence data in the co-word analysis than the indirect similarity measure. An empirical comparison study between the new proposed method and the most well-known direct similarity measures in co-word analysis was conducted to illustrate the procedure and to highlight the differences between the two methods. WE used the CoPhenetic Correlation Coefficient (CPCC) to evaluate our new method and its findings. The results show that the new method improves the ability of the clustering to assign keywords into the right cluster by an average of 50% over the current methods. After that, a new method of visualizing the quality of a unit of analysis “keywords” was introduced. The main objective of this method is to make the nodes on the map more informative than the current method. The current method restricts the node size to the count of appearances of the keywords on the data-set,

which have been argued in this dissertation. However, the new method gives more accurate size to the keyword's node based on its importance into the data-set.

References

- Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, *3*(1), 49–63.
- Ahlgren, P., & Jarneving, B. (2008). Bibliographic coupling, common abstract stems and clustering: A comparison of two document-document similarity approaches in the context of science mapping. *Scientometrics*, *76*(2), 273–290.
- Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to pearson’s correlation coefficient. *Journal of the American Society for Information Science and Technology*, *54*(6), 550–560.
- Altman, D. G., & Goodman, S. N. (1994). Transfer of technology from statistical journals to the biomedical literature: past trends and future predictions. *Jama*, *272*(2), 129–132.
- An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics*, *88*(1), 133–144.
- Anderberg, M. R. (1973). *Cluster analysis for applications. monographs and textbooks on probability and mathematical statistics*. Academic Press, Inc., New York.
- Baccini, A., Barabesi, L., & Marcheselli, M. (2009). How are statistical journals linked? a network analysis. *Chance*, *22*(3), 35–45.
- Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval* (Vol. 463). ACM press New York.
- Ballou, D. P., & Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output information systems. *Management science*, *31*(2), 153.
- Bassecoulard, E., & Zitt, M. (1999). Indicators in a research institute: A multi-level classification of scientific journals. *Scientometrics*, *44*(3), 323–345.
- Baulieu, F. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, *6*(1), 233–246.
- Baulieu, F. (1997). Two variant axiom systems for presence/absence based dissimilarity coefficients. *Journal of Classification*, *14*(1), 159–170.
- Belter, C. (2012). Visualizing networks of scientific research. *Online-Medford*, *36*(3), 14.

- Bhattacharya, S., & Basu, P. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics*, *43*(3), 359–372.
- Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure* (Vol. 28). John Benjamins Publishing.
- Bjork, B.-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: yearly volume and open access availability. *Information Research: An International Electronic Journal*, *14*(1).
- Bollen, J., Rodriquez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, *69*(3), 669–687.
- Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS one*, *4*(6), e6022.
- Bordons, M., Fernández, M., & Gómez, I. (2002). Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics*, *53*(2), 195–206.
- Borgatti, S. P. (2002). Netdraw software for network visualization. *Lexington, KY: Analytic Technologies*, 95.
- Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual review of information science and technology*, *37*(1), 179–255.
- Börner, K., Penumarthy, S., Meiss, M., & Ke, W. (2006). Mapping the diffusion of scholarly knowledge among major us research institutions. *Scientometrics*, *68*(3), 415–426.
- Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H.-D. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics*, *5*(3), 346–359.
- Bornmann, L., Stefaner, M., de Moya Anegón, F., & Mutz, R. (2014). Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualisation of results from multi-level models. *Online Information Review*, *38*(1), 43–58.
- Box, G. E., & Woodall, W. H. (2012). Innovation, quality engineering, and statistics. *Quality Engineering*, *24*(1), 20–29.
- Braam, R. R., Moed, H. F., & Van Raan, A. F. (1991). Mapping of science by combined co-citation and word analysis i. structural aspects. *Journal of the American Society for information science*, *42*(4), 233.
- Brigham, T. J. (2014). An introduction to altmetrics. *Medical reference services quarterly*, *33*(4), 438–447.

- Cahlik, T. (2000). Comparison of the maps of science. *Scientometrics*, *49*(3), 373–387.
- Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics*, *22*(1), 155–205.
- Callon, M., Rip, A., & Law, J. (1986). *Mapping the dynamics of science and technology: Sociology of science in the real world*. Springer.
- Cao, M., & Gao, X. (2005). Combining contents and citations for scientific document classification. *AI 2005: Advances in artificial intelligence*, 143–152.
- Chen, C. (2006). Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377.
- Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with googles pagerank algorithm. *Journal of Informetrics*, *1*(1), 8–15.
- Chim, H., & Deng, X. (2008). Efficient phrase-based document similarity for clustering. *IEEE Transactions on Knowledge and Data Engineering*, *20*(9), 1217–1229.
- Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22–29.
- Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, *62*(7), 1382–1402.
- Coulter, N., Monarch, I., & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the Association for Information Science and Technology*, *49*(13), 1206–1223.
- Cox, T. F., & Cox, M. A. (2000). *Multidimensional scaling*. CRC press.
- De Candolle, A. (1885). *Origin of cultivated plants* (Vol. 48). D. Appleton.
- Ding, Y., Chowdhury, G. G., & Foo, S. (2000). Journal as markers of intellectual space: Journal co-citation analysis of information retrieval area, 1987–1997. *Scientometrics*, *47*(1), 55–73.
- Ding, Y., Chowdhury, G. G., & Foo, S. (2001). Bibliometric cartography of information retrieval research by using co-word analysis. *Information processing & management*, *37*(6), 817–842.
- Eck, N. J. v., & Waltman, L. (2009). How to normalize cooccurrence data? an analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology*, *60*(8), 1635–1651.

- Egghe, L. (2009). New relations between similarity measures for vectors based on vector norms. *Journal of the American Society for Information Science and Technology*, *60*(2), 232–239.
- Egghe, L., & Michel, C. (2002). Strong similarity measures for ordered sets of documents in information retrieval. *Information processing & management*, *38*(6), 823–848.
- Egghe, L., & Michel, C. (2003). Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. *Information processing & management*, *39*(5), 771–807.
- Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Elsevier Science Publishers.
- Egghe, L., & Rousseau, R. (2006). Classical retrieval and overlap measures satisfy the requirements for rankings based on a lorenz curve. *Information processing & management*, *42*(1), 106–120.
- Feinerer, I. (2017). *Introduction to the tm package text mining in r*.
- Fellows, I., Fellows, M. I., & Rcpp, L. (2012). Package wordcloud. Retrieved, 4, 2013.
- Fernández, A., & Gómez, S. (2008). Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification*, *25*(1), 43–65.
- Franceschet, M. (2010). The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. *Journal of Informetrics*, *4*(1), 55–63.
- Francis, L., & Flynn, M. (2010). Text mining handbook. In *Casualty actuarial society e-forum* (p. 1).
- Galligan, F., & Dyas-Correia, S. (2013). Altmetrics: Rethinking the way we measure. *Serials review*, *39*(1), 56–61.
- García-Peñalvo, F., de Figuerola, C., Merlo, J., & Jacsó, P. (2010). Comparison of journal impact rankings in the scimago journal & country rank and the journal citation reports databases. *Online information review*, *34*(4), 642–657.
- Garfield, E. (1999). Journal impact factor: a brief review. *Canadian Medical Association Journal*, *161*(8), 979–980.
- Garfield, E., et al. (1964). Science citation index-a new dimension in indexing. *Science*, *144*(3619), 649–654.
- Gbur Jr, E. E., & Trumbo, B. E. (1995). Key words and phrasesthe key to scholarly visibility and efficiency in an information explosion. *The American Statistician*, *49*(1), 29–33.

- Gil-Leiva, I., & Alonso-Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. *Journal of the American Society for Information Science and Technology*, *58*(8), 1175–1187.
- Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), 69–115.
- Gross, P., & Gross, E. (1927). College libraries and chemical education.
- Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity measures in scientometric research: the jaccard index versus salton's cosine formula. *Information Processing & Management*, *25*(3), 315–318.
- Hammouda, K. M., & Kamel, M. S. (2002). Phrase-based document similarity based on an index graph model. In *Data mining, 2002. icdm 2003. proceedings. 2002 ieee international conference on* (pp. 203–210).
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques*. Elsevier.
- Hartley, J., & Kostoff, R. N. (2003). How useful are key words' in scientific journals? *Journal of Information Science*, *29*(5), 433–438.
- Hazewinkel, M. (1999). Key words and key phrases in scientific databases. aspects of guaranteeing output quality.
- He, Q. (1999). Knowledge discovery through co-word analysis. *Library trends*, *48*(1), 133–133.
- Heidarian, A., & Dinneen, M. J. (2016). A hybrid geometric approach for measuring similarity level among documents and document clustering. In *Big data computing service and applications (bigdataservice), 2016 ieee second international conference on* (pp. 142–151).
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 16569–16572.
- Hoonlor, A., Szymanski, B. K., & Zaki, M. J. (2013). Trends in computer science research. *Communications of the ACM*, *56*(10), 74–83.
- Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (nzcsrsc2008), christchurch, new zealand* (pp. 49–56).
- Hubalek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. *Biological Reviews*, *57*(4), 669–689.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264–323.

- Janson, S., & Vegelius, J. (1981). Measures of ecological association. *Oecologia*, *49*(3), 371–376.
- Ji, P., Jin, J., et al. (2016). Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, *10*(4), 1779–1812.
- Jin, B. (2007). The ar-index: complementing the h-index. *ISSI newsletter*, *3*(1), 6.
- Jinha, A. E. (2010). Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, *23*(3), 258–263.
- Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American society for information science*, *38*(6), 420.
- Jones-Farmer, L. A., Ezell, J. D., & Hazen, B. T. (2014). Applying control chart methods to enhance data quality. *Technometrics*, *56*(1), 29–41.
- Katz, J. (1994). Geographical proximity and scientific collaboration. *Scientometrics*, *31*(1), 31–43.
- Kessler, M. M. (1962). *An experimental study of bibliographic coupling between technical papers* (Tech. Rep.). DTIC Document.
- Khan, M., & Khan, S. S. (2011). Data and information visualization methods, and interactive mechanisms: A survey. *International Journal of Computer Applications*, *34*(1), 1–14.
- Kostoff, R. N., del Rio, J. A., Humenik, J. A., Garcia, E. O., & Ramirez, A. M. (2001). Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*, *52*(13), 1148–1156.
- Kostoff, R. N., Eberhart, H. J., & Toothman, D. R. (1999). Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography. *Journal of the Association for Information Science and Technology*, *50*(5), 427.
- Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling* (Vol. 11). Sage.
- Larsen, B., & Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *Proceedings of the fifth acm sigkdd international conference on knowledge discovery and data mining* (pp. 16–22).
- Law, J., Bauin, S., Courtial, J., & Whittaker, J. (1988). Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification. *Scientometrics*, *14*(3-4), 251–264.
- Leydesdorff, L. (1987). Words and co-words as indicators of the intellectual organization of the sciences'. In *Easst workshop in amsterdam (december 1987)*.

- Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1), 77–85.
- Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: extending aca to the web environment. *Journal of the American Society for Information Science and technology*, 57(12), 1616–1628.
- Li, S., Sun, Y., & Soergel, D. (2015). A new method for automatically constructing domain-oriented term taxonomy based on weighted word co-occurrence analysis. *Scientometrics*, 103(3), 1023–1042.
- Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1), 145–151.
- Liu, Y., Goncalves, J., Ferreira, D., Xiao, B., Hosio, S., & Kostakos, V. (2014). Chi 1994-2013: mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the 32nd annual acm conference on human factors in computing systems* (pp. 3553–3562).
- Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington academy of sciences*, 16(12), 317–323.
- Luukkonen, T., Tijssen, R., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics*, 28(1), 15–36.
- McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American society for information science*, 41(6), 433.
- Megahed, F. M., & Jones-Farmer, L. A. (2015). Statistical perspectives on big data. In *Frontiers in statistical quality control 11* (pp. 29–47). Springer.
- Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36(3), 363–377.
- M.L., S. (2003). Statistics: The next generation. *Journal of the American Statistical Association*, 98, 1-6. Retrieved from <http://EconPapers.repec.org/RePEc:bes:jnlasa:v:98:y:2003:p:1-6>
- Nair, V., Hansen, M., & Shi, J. (2000). Statistics in advanced manufacturing. *Journal of the American Statistical Association*, 95(451), 1002–1005.
- Nederhof, A., Luwel, M., & Moed, H. (2001). Assessing the quality of scholarly journals in linguistics: An alternative to citation-based journal impact factors. *Scientometrics*, 51(1), 241–265.
- Nelson, M. L., Bollen, J., Calhoun, J. R., & Mackey, C. E. (2004). User evaluation of the nasa technical report server recommendation service. In *Proceedings of the 6th annual acm international workshop on web information and data management* (pp. 144–151).

- Newman, M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl 1), 5200–5205.
- Pan, R. K., & Fortunato, S. (2013). Author impact factor: tracking the dynamics of individual scientific impact. *arXiv preprint arXiv:1312.2650*.
- Peat, H. J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the american society for information science*, 42(5), 378.
- Peters, H., & van Raan, A. F. (1993). Co-word-based science maps of chemical engineering. part i: Representations by direct multidimensional scaling. *Research Policy*, 22(1), 23–45.
- Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5), 297–312.
- Rip, A., & Courtial, J. (1984). Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6(6), 381–400.
- Rodrigues, S., Van Eck, N., Waltman, L., & Jansen, F. (2014). Mapping patient safety: a large-scale literature review using bibliometric visualisation techniques. *BMJ open*, 4(3), e004468.
- Rohlf, F. J., & Fisher, D. R. (1968). Tests for hierarchical structure in random data sets. *Systematic Biology*, 17(4), 407–412.
- Ryan, T. P., & Woodall, W. H. (2005). The most-cited statistical papers. *Journal of Applied Statistics*, 32(5), 461–474.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513–523.
- Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval.
- Sayyadi, H., & Getoor, L. (2009). Futurerank: Ranking scientific articles by predicting their future pagerank. In *Proceedings of the 2009 siam international conference on data mining* (pp. 533–544).
- Schneider, J. W., & Borlund, P. (2007). Matrix comparison, part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology*, 58(11), 1586–1595.
- Schubert, A., Glänzel, W., & Thijs, B. (2006). The weight of author self-citations. a fractional approach to self-citation counting. *Scientometrics*, 67(3), 503–514.

- Shannon, C. (1948). A mathematical theory of communication, bell system technical journal 27: 379-423 and 623-656. *Mathematical Reviews (MathSciNet): MR10, 133e*.
- Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2), 253-280.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4), 265-269.
- Smith, M., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J., & Dunne, C. (2010). *Nodexl: a free and open network overview, discovery and exploration add-in for excel 2007/2010*.
- Smoot, M. E., Ono, K., Ruschinski, J., Wang, P.-L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3), 431-432.
- Sokal, R. R., Sneath, P. H., et al. (1963). Principles of numerical taxonomy. *Principles of numerical taxonomy*.
- Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science*, 94-108.
- Tan, P.-N., et al. (2006). *Introduction to data mining*. Pearson Education India.
- Team, S. (2009). *Science of science (sci2) tool. indiana university and scitech strategies*.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
- van Eck, N., & Waltman, L. (2009). Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.
- Van Eck, N. J., & Waltman, L. (2007). Bibliometric mapping of the computational intelligence field. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 15(05), 625-645.
- van Eck, N. J., & Waltman, L. (2009). Vosviewer: A computer program for bibliometric mapping.
- van Eck, N. J., Waltman, L., Dekker, R., & van den Berg, J. (2010). A comparison of two techniques for bibliometric mapping: Multidimensional scaling and vos. *Journal of the American Society for Information Science and Technology*, 61(12), 2405-2416.

- van Eck, N. J., Waltman, L., & van den Berg, J. (2005). A novel algorithm for visualizing concept associations. In *16th international workshop on database and expert systems applications (dexa'05)* (pp. 405–409).
- Vaughan, L., & You, J. (2010). Word co-occurrences on webpages as a measure of the relatedness of organizations: A new webometrics concept. *Journal of Informetrics*, *4*(4), 483–491.
- Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*(06), P06010.
- Walter, G., Bloch, S., Hunt, G., & Fisher, K. (2003). Counting on citations: a flawed way to measure quality. *Medical Journal of Australia*, *178*(6), 280–281.
- Waltman, L., & Eck, N. J. v. (2007). Some comments on the question whether co-occurrence data should be normalized. *Journal of the American Society for Information Science and Technology*, *58*(11), 1701–1703.
- Waltman, L., van Eck, N. J., & Noyons, E. C. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, *12*(4), 5–33.
- Ware, M., & Mabe, M. (2009). *The stm report-an overview of scientific and scholarly journals publishing international association of scientific, technical and medical publishers*. 68 p. Oxford.
- Ware, M., & Mabe, M. (2015). The stm report: An overview of scientific and scholarly journal publishing.
- Web of science @ONLINE*. (2016, June). Retrieved from <https://apps. webofknowledge.com>
- Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a. *MIS quarterly*, *26*(2), 13–23.
- West, J. D., Jensen, M. C., Dandrea, R. J., Gordon, G. J., & Bergstrom, C. T. (2013). Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community. *Journal of the American Society for Information Science and Technology*, *64*(4), 787–801.
- White, H. D., & Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for information Science*, *32*(3), 163–171.
- White, H. D., & McCain, K. W. (1997). Visualization of literatures. *Annual review of information science and technology (ARIST)*, *32*, 103.

- White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American society for information science*, 49(4), 327–355.
- Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science*, 19(3), 473–496.
- Woodall, W. H., & Montgomery, D. C. (1999). Research issues and ideas in statistical process control. *Journal of Quality Technology*, 31(4), 376.
- Woodall, W. H., & Montgomery, D. C. (2014). Some current directions in the theory and application of statistical process monitoring. *Journal of Quality Technology*, 46(1), 78.
- Yan, E., & Ding, Y. (2010). Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*, 61(8), 1635–1643.
- Yang, C., Zhu, D., & Wang, X. (2017). Sao semantic information identification for text mining.
- Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). term clumping for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, 85, 26–39.
- Zitt, M., Bassecouard, E., & Okubo, Y. (2000). Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics*, 47(3), 627–657.
- Zitt, M., Lelu, A., & Bassecouard, E. (2011). Hybrid citation-word representations in science mapping: Portolan charts of research fields? *Journal of the American Society for Information Science and Technology*, 62(1), 19–39.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418–1429.