TEST Engineering & Management

*July-August 2020*
*ISSN: 0193-4120 Page No. 4324 - 4345*

# Finding Main Paths of a Research Subject: a Thorough Demonstration，Key Problems Solving And a Case Study of Main-Path Study Itself

**Shao Zhiyi[1, *], Li Yongming[1, 2, 3], Wu Ke[3], Hui Fen[2], Niu Yanfen[2], Yuan Shen[2], Feng Fan[4]**

[1] The Library, Shaanxi Normal University, Xi'an,China

[2] School of Mathematics and Information Science, Shaanxi Normal University, Xi'an,China

3 School of Computer Science, Shaanxi Normal University, Xi'an,China

[4] Research Center for Clinical and Translational Medicine, The 302nd Hospital of Chinese PLA,Beijing, China

∗ Corresponding author: Shao Zhiyi. E-mail address: shaozy@snnu.edu.cn

***Article Info***
***Volume 83***
***Page Number: 4324 - 4345***
***Publication Issue:***
***July-August 2020***

***Article History***
***Article Received:*** *25 April 2020*
***Revised:*** *29 May 2020*
***Accepted:*** *20 June 2020*
***Publication:*** *10 August 2020*

***Abstract***

The main paths show the key events and key publications of a research subject. Finding main paths for a research subject could help scholars quickly grasp the main developing trajectory of the corresponding subject and thus plays an important role. Several previous studies have aimed to solve the existed technical problems, in order that a more perfect main paths finding method can be obtained. On the other hand, plenty of empirical main path analysis have also been launched to gain insights into various research subjects. However, some technical problems are still not addressed or clearly illustrated. For example, there still have no definite methods to obtain a comparatively more complete dataset based on which main paths will be found, and there also have no definite methods to transform a cyclic paper citation network into an acyclic one, after which main paths finding technique could be used. At least, none of these two problems has been illustrated exactly in detail. Furthermore, the main path analysis method has always been used as a black box, and no studies have illustrated in detail that how to obtain main paths for a research subject step by step. Thus, beginners, who just start to launch main path studies or who come from different research backgrounds and use main path analysis as a technical tool in order to contribute in their own research subject, could easily be confused. This could hinder the rapid development and widely use of main path studies. To address these problems, this paper demonstrates systematically that how to find main paths, in order that beginners or scholars of various research subject could grasp and use this technique easily; proposes the backward-forward data collection method, in order that a comparatively more complete dataset can be obtained; proposes a technical method to transform an acyclic network into a cyclic one, based on the directional property of knowledge flows. The empirical main path analysis in this study is launched for the research subject of the main-path study itself, and the empirical result could also be beneficial for scholars who are interested in main path studies. Some internal mechanisms of the SCI2 software which is used to construct the direct paper citation network in this study are also explained in detail.

***Keywords:*** *Main path · Data collection · Acyclic network · Direct paper citation network*

TEST Engineering & Management

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

## Introduction

Innovations or technological development is a discovery process, in which previous problems are solved while new problems and new knowledges are proposed. In this process, big break throughs are rare, instead, sequential and progressively minor refines and improvements are usually achieved. These sequential and progressively minor achievements thus construct technological trajectories of a research subject, and therefore, technological trajectories are suggested to interpret directions and determinants of technical changes (Dosi 1982). Finding the developing tajectories of a research subject from millions of scientific papers is not easy. Traditional methods require scholars to read plenty of related papers, and then scholars could form a rough understanding of the development of a subject. This is time consuming, as a huge number of scientific papers may exist for a concrete research subject, nowadays. Still, scholars may easily sum up with an inaccurate result, as they can hardly read all the related scientific papers. Thanks to the techniqual method of main path analysis, finding developing trajectories of a research subject becomes much easier, faster and more accurate.

Main paths analysis can disclose the main developing mechanism of a research subject and can provide scholars with the key papers and key events happening in the developing process. Usually, main paths may not be single, and they are made up of a successive backbone papers which are expected to contain main findings in a research subject. Main path analysis is a special kind of citation analysis which has become prevalent since Garfield's pioneering work (Garfield 1955). Citation analysis plays an important role in the distinguishing of backbone papers, as they could explicitly demonstrate the relationships among papers and describe the strengths of these relationships. Base on the assumption that the history of science is a chronological sequence of events in which new discoveries are based on earlier discoveries, Garfield suggests that citation analysis of scientific papers makes the writing of science history possible (Garfield 1964). If a scientific paper receives more citations, this paper will have a more possibility to be a milestone or a key event in the related research subject (Garfield 1970). Still, citation relationships have been proved to be invaluable during the study of technical changes (Jaffe 2002). Citation studies can be accordingly grouped into two categories, one is measuring the prominence of scientific papers and another is analyzing the structure of citation networks. Structure analysis of citation networks will often be launched if specific network relationships are to be considered. Relationships among papers include the direct citation relationship of "citing" or "cited by", bibliometric coupling (Kessler 1963), and co-citation (Small 1973). Traditional methods for network structure analysis is to cluster scientific papers based on the above-mentioned network relationships. These methods mainly concentrate on the network nodes such as scientific papers, not on the relationships among these nodes. When Garifield analyzes the paper citation network of DNA theories, there are only about 40 papers. Thus, the analysis was easy to be managed. Along with the rapid growth of scientific papers, directly structure analysis of these papers using traditional methods become rather inefficient. The main path method is thus proposed by Hummon to address this problem (Hummon 1989). Different from concentrating on the network nodes as traditional methods does, the main paths analysis method puts its attention mainly on the relationships among these network nodes. In other words, traditional methods for network structure analysis pay attention to the network nodes, while the main paths methods mainly pay attention to the network links among these nodes. Main paths are those network paths carrying the most knowledge flows, and they are the most important paths in a citation network. To

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

measure the importance of a link in a paper citation network, Hummon proposes the concept of traverse weight. The traverse weight of a link tells us the number of paths going through this link, and its value expresses the importance of this link. Hummon has developed three methods to compute the traverse weight, SPLC (Search Path Link Count), SPNP (Search Path Node Pair) and NPPC (Node Pair Projection Count). Having measured the importance of links based on these three methods, Hummon suggests that the depth first search method could be used to find the main path. Based on the paper citation network of DNA studies, Hummon has obtained the key events and the key papers leading to the discovery and modeling of DNA theories. The citation network that Hummon has used is the same as that Garifield had ever used (Garfield 1964), and the results obtained by main paths analysis are quite convincing compared with Garifield's analysis. However, that paper citation network is rather small, and there are only 40 network nodes constructed from 69 scientific papers. Later, Hummon uses the main path analysis method on a larger dataset which contains 119 network nodes and 632 citation links among these nodes. These nodes include scientific papers, reports and books, which are about centrality and productivity research. That constructed network is a "is cited by" network, in which the influence of earlier research on later research is mapped. In that study, traverse count computation methods of NPPC, SPLC and SPNP are all used, and a clear main path for centrality and productivity research is found (Hummon 1990-1). In the same year, Hummon illustrates in detail the depth first search method which comes from the computer science research fields (Hummon 1990). Based on the papers published in volumes 1 to 12 of the Social Networks journal, Hummon has also found main paths that are consistent with what the ancestor Kuhn had already found. This suggests that the main path method could provide scholars with accurate analysis result (Hummon 1993). Later, Carley analyzes the main path structure of the Journal of Conflict Resolution (Carley 1993). However, main path analysis has not been widely used for very large citation networks with several thousands of nodes until Batagelj developes efficient algorithms for searching main paths and embeds these computer programs into the Pajek software (Batagelj 2003). In that study, Batagelj develops highly efficient algorithms for Hummon's SPLC and SPNP. These algorithms are linear in number of arcs and thus can be used for analysis of very large networks. Batagelj also developes his own traverse weight computation method SPC, and compared it with SPLC and SPNP. That comparison shows that results of these three methods are nearly the same, but SPC is recommended due to some of its nice properties. Batagelj also provides two examples based on Pajek software, one is about the paper citation network and the other is about the patent citation network. If we see Garfield's proposal of citation network as the first step (Garfield 1964), see Hummon's proposal of main paths as the second important step (Hummon 1989), then Batagelj's contribution can be seen as the third important step for citation network analysis (Batagelj 2003). Based on previous pioneering works, Morre studies the genealogy of the concept of social capital in public health (Moore 2006). Mina uses this technique to investigate the development of a most important medical innovation, Percutaneous Transluminal Coronary Angioplasty, which is an important treatment developed by medical community against coronary artery disease, the most common cause of death in developed countries in that period. That investigation is based on two dataset, scientific papers and patent documents, and the analysis is launched using the Pajek software (Mina 2007). Verspagen examines the developing trajectories of fuel cell technology based on patent analysis (Verspagen 2007). Carlero-Medina investigates the developing trajectories of the absorptive capacity research (Carlero-Medina 2008).



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

Lucio-Arias applies HistCite to construct paper citation networks for fullerene research and nanotube research respectively, and adopts SPLC algorithm for main paths finding in Pajek (Lucio-Arias 2008). Harris studies the main paths of the secondhand smoke research (Harris 2009). Lu examines the research history of the ethics of nanotechnology (Lu 2012). In 2012, another milestone work is launched by Liu, in which several technical contributions are provided (Liu 2012). That study can be seen as another important step in main path study after Garfield's, Hummon's, and Batagelj's previous work. In Hummon's work, main path is selected by repeatedly choosing the link which emanates from the current node and has the largest traversal count. Using this method, the link selected each time has the largest traversal count, however, the resulted overall traversal count of all these selected links is not definitely the largest. Liu calls Hummon's priority first search algorithm as the local approach, and he thus designs a global approach. In the global search method, the link selected each time may not all have the current largest traversal count, however, the overall summed traversal count of the main path is the largest among all the paths in the network. This searching method is somewhat similar to a reverse version of the shortest paths searching method in graph theory (Liu 2012). Actually, these two methods of main paths finding have their own features, which method to adopt is decided by what information the researchers want to express. The local main path search method highlights the progressing significance; while the global main path search method emphasizes the overall importance in knowledge flows. Liu's later work (Liu 2013-a) shows that main paths found by the local search method and by the global search method are quite similar and differs only slightly in the early stage and late stage of the main paths. Still, Hummon's main path searching method is forward, which selects those papers attracting the most followers as nodes on the main paths. This is like finding the offspring of important contributions. From the opposite perspective, Liu proposes the backward searching method (Liu 2012). Backward searching selects those papers absorbing ideas from the widest sources as nodes on the main paths. Forward searching is based on the outdegree of network nodes, while backward searching is based on the indegree. However, backward searching is only applicable for local searching method. Actually, the backward searching idea has already been tried in Lucio-Arias' earlier work, where this method is called codification (Lucio-Arias 2008). Liu also mentions that, by relaxing the search constraint, more paths whose importance are next to the main path can be found, and more details can be revealed. This method is called multiple main paths searching. "Multiple main paths searching" is different from "the main paths is not single". The former is finding multiple main paths actively, and these different paths have different level of importance. The latter is induced because that several different main paths have the same traverse weights, and these main paths have the same level of importance. Another contribution of Liu's work is the design of key-route searching, which is proposed because of the realization of the problem that the link with the highest traversal count may not definitely be selected for the main path. Liu names such link as key-route. Before we move toward to the concrete searching method, we should introduce some basic knowledges to avoid confusion. A key-route is actual an arc, it has an arc tail and an arc head. If this arc points from the left to the right, then the left node is called the arc tail and the right node is called the arc head. As arch tail and arc head is often confused, we can understand them by imaging the bow and arrow. When the arrow is shot out, then the end which is in the front is called the arrowhead, and the end which is in the rear is called the arrow nock. The situation of arc tail and arc head is similar to this. Now, we come back to illustrate the concrete key-route searching method. In



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

this method, the key-route is selected first, then searching forward is launched from the head of this link until reaching the sink node which is not cited, at the same time, searching backward is launched from the tail of this link until reaching the source node which cites no other nodes. Similarly, by relaxing the constraint, multiple key-route main paths can be obtained. Based on the published papers of H-index studies, Liu collects several main paths for the developing trajectories of H-index studies, including the single forward local main paths, the single forward global main path, the single backward local main path, the multiple forward local main paths, and the key-route main paths with both local and global searching (Liu 2012). In Liu's later work, the local key-route main paths method is applied to the five major DEA (data envelopment analysis) applications, and thus a survey of DEA applications is obtained (Liu 2013-b). Chuang applied the local multiple key-route main path searching method to medical tourism research publications, and two distinctive developing paths are found for this research field (Chuang 2014). Still, in Liu's another later work, the edge-betweenness based network clustering method is applied to DEA literatures and four research fronts are identified. Then, the global key-route main path analysis is applied to each front to see details, where SPLC is adopted to measure the significance of links (Liu 2016). Kaffash investigates the DEA method and its applications in financial services based on the critical path method proposed by Batagelj (Batagelj 2003), and main paths for three thematic group are obtained (Kaffash 2017). Based on eTourism research papers, Chuang adopts the Girvan-Newman edge-betweenness clustering algorithm to divide the citation network into six groups, and then applies the key-route main path analysis to trace the historical development of each of these 6 groups (Chuang 2017). Based on the research papers of e-tourism, Ho investigates the influence of review papers on citation-based analysis, and finds that review papers can introduce bias during the main path analysis. However, on whether including the review papers or not, Ho suggests that this should be dependent on the distinctive research purposes (Ho 2017). Using SPC to measure the importance of links, Zhou finds the local main paths of the research field of Data Envelopment Analysis (Zhou 2018). In Pajek software, using SPC to compute the traverse count of links, Lee investigates the key-route main paths of the human space exploration research, and thus helps scholars recognize the underlying knowledge diffusion flow to establish future research directions (Lee 2018).

**Contributions of this paper**

(1) systematically demonstrates how to find main paths of a research topic.

(2) proposes the backward-forward data collection method, in order that a comparatively complete dataset can be obtained.

(3) proposes a method to translate a cyclic paper citation network into an acyclic one, in order that various main paths finding methods can be used.

(4) investigates the SPC-Local-Forward main paths of the research subject of main-path study itself.

(5) explains in detail the internal mechanism of the SCI2 software which is used to construct the direct paper citation network, in order that beginners can use it properly.

**Data collection**

Data collection is the first step in the whole operations. The data that we collect are not actual papers, but their bibliographic data. In our study, these bibliographic data are collected from Web of Science. A complete dataset is important, as all subsequent analysis will be launched based on this dataset. Obtaining a thoroughly complete dataset for a certain research subject seems impossible, however,



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

various measures can be taken to gain a comparatively complete dataset. This study proposes the ***Backward-Forward Data Collection Method*** as follows.

***Firstly,*** collect data of Core Papers. Core Papers are those most relevant to a research subject. They are found by retrieving in a concrete database, such as Web of Science, through keywords search. Thus, all the possible keywords that can be used to represent this research area should be listed out. Then, topic searches will be launched using these keywords. Compared with title search and other search methods, topic search provides us with more relevant papers. Taking Web of Science as an example, when topic search is launched, this database finds those papers containing a concrete keyword in their Titles, Abstracts, Author Keywords, or Keywords Plus. Papers retrieved in this step are called Core Papers. Then, Full Record and Cited References of these Core Papers are saved as a Plain Text format file.

***Secondly,*** collect data of Core Papers' Cited Papers. Treat Core Papers as the original point, this step is the **Backward Collection**. Although the data collected in the first step contain cited references, they are only auxiliary data which will be used to judge relationships among Core Papers. When the main path analysis is launched, Core Papers are treated as nodes of the citation network, while their Cited Papers are not. Cited Papers may have similar topics with Core Papers, and they may be key nodes on the main paths. However, if they are not nodes of the citation network, they will never appear on any main paths. To solve this problem, data of these Cited Papers should be collected independently. Full Record and Cited References of these Cited Papers are also saved as a Plain Text format file. Then, in data analysis software such as Pajek, Cited Papers of Core Papers will also be seen as network nodes.

***Thirdly,*** collect data of Core Papers' Citing Papers. Treat Core Papers as the original point, this step is the **Forward Collection**. Just as Cited Papers, Citing Papers may also have similar topics with Core Paper. Data of Citing Papers should also be collected independently, in order that they have the possibility to appear on the main paths. Full Record and Cited References of these Citing Papers are also saved as a Plain Text format file.

***Fourthly,*** collect data of **un-WOS-indexed milestone papers and un-WOS-indexed important papers** which are not indexed in Web of Science. These papers will be seen as parts of the Core Papers, and will be manually put at the beginning of the dataset. Our dataset is collected from Web of Science, however, not all the papers about main-path study are indexed in this database. Of course, there is hardly a database in which all the related papers corresponding with a certain research subject could be index. However, we still should include those milestone papers and important papers, which could not be indexed in Web of Science, in our dataset as they may significantly affect the analysis results. For example, those papers written by the leaders of the related research subject, or those papers receiving high citation counts. This task could be finished by carefully reading the review papers which are related with the studied subject.

Through reading the related reviews, two milestone papers and four important papers are found not indexed in Web of Science. The two milestone papers are "HUMMON NP, 1989, Connectivity in a citation network: The development of DNA theory, SOC NETWORKS, V11, P39, DOI 10.1016/0378-8733(89)90017-8" and "Batagelj V, 2003, Efficient Algorithms for Citation Network Analysis, arXiv:cs/0309023". The first paper is the start point of main-path study, and the second paper is an important technical improvement. The four important papers are "HUMMON NP, 1990, KNOWLEDGE, V11, P459, DOI 10.1177/107554709001100405",



Published by: The Mattingley Publishing Co., Inc.

TEST Engineering & Management

*July-August 2020*
*ISSN: 0193-4120 Page No. 4324 - 4345*

"HUMMON NP, 1990, SOC NETWORKS, V12, P273, DOI 10.1016/0378-8733(90)90011-W", "HUMMON NP, 1993, SOC NETWORKS, V15, P71, DOI 10.1016/0378-8733(93)90022-D", and "CARLEY KM, 1993, KNOWLEDGE, V14, P417, DOI 10.1177/107554709301400406". The first three papers are written by the beginner HUMMON NP, and HUMMON NP is also a participate author in the fourth paper. Thus, we collect manually the information that needed for these six papers.

**PT** J

**AU** Batagelj, V

**TI** Efficient Algorithms for Citation Network Analysis

**AB** In the paper very efficient, linear in number of arcs, algorithms for determining Hummon and Doreian's arc weights SPLC and SPNP in citation network are proposed…

**DE** large network; acyclic; citation network; main path; CPM path; arc weight; algorithm; self organizing maps; patent

**CR** Hummon NP, 1989, SOC NETWORKS, V11, P39, DOI 10.1016/0378-8733(89)90017-8…

**J9** ARXIV

**PY** 2003

**VL** arXiv:cs/0309023

**BP** 1

**UT** ARXIV:BATAGELJ2003

**ER**

However, there is an important problem to be addressed for these manually collected data, i.e., data format. All the data collected from Web of Science automatically have the same data format. Each paper indexed in Web of Science has a unique number, represented by the data field of UT in the bibliometric data. UT will be used to distinguish different papers when constructing paper citation network using SCI2. For example, when the value of UT is WOS:000452245000013, then the paper is confirmed to be "How academic librarians involve and contribute in research activities of universities? A systematic demonstration in practice through comparative studies of research productivities and research impact" (Shao 2018). However, these manually collected papers have no UT. We thus provide each paper with a virtual UT value by experiences. For example, the UT value for these six papers are set to be SOCIAL NETWORKS: HUMMON1989, KNOWLEDGE: HUMMON1990, ARXIV: BATAGELJ2003, SOCIAL NETWORKS: HUMMON1990, SOCIAL NETWORKS: HUMMON1993, and KNOWLEDGE: CARLEY1993. Beside the UT field, we have to construct some other bibliometric data fields. Take Batagelj's paper as an example, we show how to construct an intact data block.

The above data block begins with the data field of

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

PT and ends with the data field of ER. There is no strict requirement for the sequence of these other data fields. In these data fields, PT means the kind of the paper; AU means the authors; TI means the title of the paper; AB means the abstracts; DE means the keywords; PY means the publication year; VL means the number of the volume; BP means the beginning page of the paper; UT is the unique identifier of the paper; J9 means the abbreviated title of the journal. The data filed of TI, AB, and DE could be managed flexibly. The reason of the setting of these three data fields is that they will be used in the followed data preprocess. In order to filter out the unrelated papers, these data fields will be checked. In fact, it is enough that one or some of these data fields include the needed keywords. When constructing other data fields, the format of these data fields must be the same as those collected automatically in the database of Web of Science, or some unexpected problems may arise.

After setting the virtual UT value and the other bibliographic data fields, these manually collected bibliographic data can be processed as they are collected from Web of Science. We must mention again that a that the format of these manually collected bibliometric data must be the same as those downloaded from Web of Science, and their information should be the same as they appear in the data filed of cited references CR. Otherwise, these manually added papers will not be correctly recognized, for example, SCI2 may wrongly recognized them as different papers, and the result of finding main paths will be affected. A little trick is to manually construct the paper records according to the data field of "CiteMeAs". For example, the data filed of CiteMeAs for Hummon's milestone paper is "HUMMON NP, 1989, Connectivity in a citation network: The development of DNA theory, SOC NETWORKS, V11, P39, DOI 10.1016/0378-8733(89)90017-8". Thus, we can set the AU data field as "Hummon, NP", PY as "1989", TI as "Connectivity in a citation network: The development of DNA theory", J9 as "SOC NETWORKS", VL as "11", BP as "39", and DI as "10.1016/0378-8733(89)90017-8". Then, other information can be added. Among these other information, DE and CR should be put special attention. The DE data field includes the keywords of a paper, and the CR data field includes its references. In the preprocess procedure, unrelated papers will be filtered by the keywords such as "main path", thus DE must be set appropriately. The links among papers are determined by their cited references, thus CR also must be set appropriately according to the references of papers.

***Finally,*** we put the data of un-WOS-indexed milestone papers and un-WOS-indexed important papers at the very beginning of the file, and append the data of Core Papers, the data of Cited Papers and the data of Citing Papers, to them. By doing this, we obtain the dataset on which all the subsequent operations will be built. In the rests of this paper, this dataset is called ***the original dataset***. In the section of "Data Preprocess", we will illustrate why the proposed ***Backward-Forward Data Collection Method*** works well.

In our empirical study, the bibliographic data of ***the original dataset*** is collected from Web of Science Core Collection, during October 15 to October 18, 2018. The keyword which we use to retrieve these data was "main path". After setting the time span to be All years (1986-2018), through topic search, data of 258 Core Papers, 4105 Cited Papers, and 2619 Citing Papers, 6 un-WOS-indexed papers (including 2 milestone papers and 4 important papers) are collected. Thus, our ***original dataset*** contains Full Record and Cited References of 6988 papers.

**Data Preprocess**

The main task of data preprocess is to filter out the unrelated data of papers from the original dataset. This is launched automatically by a computer



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

program that we write. This program checks the following data filed, TI (title)、AB (abstract)、DE (author keywords)、ID (keywords plus), and see whether these data fields contain the keyword of "main path", "evolution", and "trajectory". If none of these four data fields contain anyone of these three keywords, then the record of this paper will be removed from the original dataset. This program also tells us which papers are left and which papers are removed, how many papers are left and how many papers are removed. After the data filtering, 919 papers are left. From this we can see that our collection method works. If we retrieve the papers by using keyword "main path", we can only obtain 258 papers. Through the ***Backward-Forward Data Collection Method***, 661(991-258) more papers are obtained, and these papers have close relationships with main-path studies. 6069 records of unrelated papers are removed. Although the number of the unrelated papers is somewhat large, however, there are still 661 closely related papers are obtained. In our opinion, it deserves because a complete dataset is rather important for data analysis. We could never obtain a thoroughly complete dataset for a research subject, however, the obtaining of these 661 additional papers proves that ***Backward-Forward Data Collection Method*** could provide us with a comparatively more complete dataset. By filtering out the unrelated paper data, the data analysis could be launched more efficiently.

Another important task of data preprocess is unifying the reference format of a paper. To avoid unnecessary mistakes, this is done manually. Take Batagelj's milestone paper, "Batagelj V, 2003, Efficient Algorithms for Citation Network Analysis, arXiv:cs/0309023", as an example. This paper is published in arXiv, and its cited formats are various. For example, "BATAGELJ V, 2003, CSDL0309023", "Batagelj V., 2003, ARXIVCS0309023", and "Batagelj V., 2003, EFFICIENT ALGORITHMS" are all found in the original dataset. This leads to a problem that paper citation network construction software such as SCI2 treats these various formats as different papers, and thus different nodes will be created, and the cited times of the corresponding paper will be wrongly counted. What' more, the relationships of this paper with other papers will be wrongly recognized. Then, the result of main-path finding thus will be affected. As Batagelj' s paper is a milestone, thus the result of main-path finding will be affected significantly. To address this problem, we preprocess the dataset by unifying the cited format of this paper as "Batagelj V, 2003, ARXIV: CS/0309023". By unifying the reference format of a paper, the data can be used more properly.

After these two tasks having been finished, we can obtain ***the preprocessed dataset*** which ensures that the original dataset can be efficiently and properly used.

**Construct Acyclic Direct Paper Citation Network**

Data Collection and Data Preprocess in above sections are very important for data analysis, as a comparatively complete dataset is the primary guarantee of the accurate analysis results, and data preprocess ensures that the collected data can be efficiently and properly used. After these two important tasks having been finished, we can then extract the paper citation network based on the preprocessed dataset containing 919 paper records.

Batagelj's milestone work has ever mentioned three ideas to transform a cyclic network to be acyclic. The first idea is to shrink the cyclic networks; the second idea is to delete some arcs; the third idea is the preprint transformation (Batagelj 2003). However, this milestone work does not systematically tell scholars in an operational way that how to transform a cyclic paper citation network to be acyclic. This paper proposes our own theory, and demonstrate in detail that how to transform a cyclic paper citation network into acyclic.



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

Our method is launched based on the preprocessed dataset, and the whole procedure is divided into two steps. The first step is to construct a Direct Paper Citation Network, and the second step is to make this network Acyclic. These two steps are demonstrated in detail individually in the following two subsections.

**First Step: Construct the Direct Paper Citation Network**

This section demonstrates how to construct a Direct Paper Citation Network using the Sci2 software. Studies about Sci2 are limited, thus some details are shown in this study. This not only helps scholars master how to construct a Direct Paper Citation Network using Sci2 software, but also helps scholars understand the internal working mechanism of this software.

A Direct Paper Citation Network shows the direct citation relationships among papers. Nodes of the network represent papers and arcs represent direct citation relationships among these papers. Arc tails represent cited papers and arc heads represent citing papers. A paper citation network should be a directed network in which knowledges flow from arc tails to arc heads.

Virous software can be used to construct the Direct Paper Citation Network. For example, HistCite, CiteSpace, VosViewer, Sci2 and so on. Sci2 is chosen in this study, because it is more suitable for constructing very large networks and because the resulted network file can be easily analyzed in the Pajek software. Sci2 is the abbreviation of Science of Science, and is a modular tool set specially designed for the study of science. The project investigators of the design of Sci2 are Katy Börner, who works at Indiana University, and Kevin W. Boyack, who works in SciTech Strategies Inc. Its development is supported in part by the Cyberinfrastructure for Network Science center and the Department of Information and Library Science at Indiana University (Sci2 Team 2009). The version of Sci2 that we use in this study is 1.2 beta.

The preprocessed dataset is recognized and loaded as ISI Flat Format by Sci2. There are 919 paper records in the preprocessed dataset, and each record has 41 data fields. For example, the data field TI shows the title of a paper. Among these 41 data fields, 10 cannot be recognized by Sci2, as shown in Table 1. Sci2 gives a new name called “is” for these unrecognized data fields, and fortunately these unrecognized data fields have no impact on the construction of the network.

After the preprocessed dataset is loaded, Sci2 automatically saves the dataset as a CSV format file called Unique ISI Records. All the subsequent operations will be based on this file. Compared with the original dataset and the preprocessed dataset, all the data fields in this file is renamed, and some new data fields, such as Cite Me As, are created. “Cite Me As” is a data field internally created by Sci2 for each record, and is used to match paper records. It is constructed by combining some data fields of the ISI records in the original dataset. These combined data fields include the first author (AU), publication year (PY), journal abbreviation (J9), volume (VL), beginning page (BP), and DOI. However, there definitely exist duplicate records in the preprocessed dataset. This dataset consists of four parts, data of Core Papers, data of Cited Papers, data of Citing Papers, and data of un-WOS-indexed papers. Obviously, a Citing Paper of paper A may also be a Cited Paper of paper B, when A and B both appear in the Core Papers, then duplications appear because their Citing Papers and Cited Papers are also be collected. Thus, duplication records definitely appear in the preprocessed dataset. Because paper A and paper B have similar topics, there has a high possibility that they appear at the same time in the Core Papers. And this makes it a high possibility that duplications exist. Thus, Removing ISI Duplicate Records should be launched for Data Preparation.



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

Actually, Sci2 removes these duplications automatically when the dataset is loaded. This is based on the attribute of ISI Unique ID in each paper record, in the preprocessed dataset. In the preprocessed dataset, the ISI Unique ID is represented by the data filed UT; in the Unique ISI Records file, the ISI Unique ID is represented by the data filed Unique ID. The value of these two data filed is the same. For example, in the preprocessed dataset, the value of the UT data filed of a paper record is "WOS:000452245000013; in the Unique ISI Records file, the value of the Unique ID data filed of the same paper record is also "WOS:000452245000013". They both represent the paper "How academic librarians involve and contribute in research activities of universities? A systematic demonstration in practice through comparative studies of research productivities and research impact (Shao et al. 2018)". Actually, the duplicate records are removed by Sci2 automatically when the preprocessed dataset is loaded, but Sci2 also provide a function called Remove ISI Duplicate Record to remove the duplications manually. Finally, 812 records are left, and the obtained dataset is thus called Unique ISI Record.

Based on these 812 records, the direct paper citation network will be built. Sci2 provides a function called Extract Paper Citation Network to fulfill this task. All the paper records of the network nodes are put together into a single file automatically. These records include bibliometric data of the Core Papers, Cited Papers, Citing Papers, un-WOS-indexed papers, cited references of the Core Papers, cited references of the Cited Papers, cited references of the Citing Papers, and cited references of the un-WOS-indexed papers. Data of the Core Papers, Cited Papers, Citing Papers, and un-WOS-indexed papers come from our original dataset and is collected directly from the Unique ISI Record file, while their respectively cited references are collected by Sci2 automatically from the data field of Cited References in the Unique ISI Record file. Still, there certainly exist duplicate records, and these duplications are removed by Sci2 automatically. However, if we do not include all these paper records in the dataset in order to avoide duplications, many important nodes will not be created because of the absence of the bibliometric data of these papers. But these nodes may be very important, because they may appear on the main paths. Sci2 puts all the paper records related with the network nodes into a single file together. In this file, each record of the cited references of a paper is put at the front by Sci2, and the record of the corresponding paper is put at the rear. Similar operations are repeated until all these 812 records and the data of their cited references are arranged, as Fig 1 shows. Then, Sci2 removes the duplicated records automatically. Among these duplicated records, those appear for the first time in the file is left and those appear later as duplications are deleted. After removing the duplications, Sci2 creates one node for each of the left paper records. These nodes can be divided into two kinds. The first kind is constructed based on the paper records of Core Papers, Cited Papers, Citing Papers and un-WOS-indexed important papers, and these records come from the original dataset; the second kind is constructed based on the paper records of cited references of these papers, and they can be found in the data field of CR of each paper records appearing in the original dataset. These two kinds of node can be easily distinguished by their data filed of "inoriginaldataset". If a node belongs to the first kind, then a new data filed of "inoriginaldataset" is created, while the second kind of node has not this data field. Take the nodes shown below as an example. Node 1 belongs to the first kind of node as it has the data field of "inoriginaldataset", and its data directly comes from the related paper record in the original dataset; while Node 2 belongs to the second kind of node as it has not such a data field, and its data comes from the CR data field of some paper record, which



Published by: The Mattingley Publishing Co., Inc.

*July-August 2020*
*ISSN: 0193-4120 Page No. 4324 - 4345*

reveals that this paper is a cited reference of some paper in the original dataset.

***Node 1:*** "Yuan J, 2018, Ieee T Ind Electron, V65, P9625, Doi 10.1109/tie.2018.2823691" localcitationcount -1 inoriginaldataset "Yuan, J|Yang, SK|Cai, JX" globalcitationcount 0

***Node 2:*** "Lv C, 2018, Ieee T Ind Inform, V14, P3436, Doi 10.1109/tii.2017.2777460" localcitationcount 1

In the constructed network, nodes represent papers and arcs represent the citation relationships among papers. Finally, 43302 nodes and 51182 arcs are created for this direct paper citation network.

This network is then automatically saved as a .net format file called Extracted Paper-Citation Network. The first half of this file describes the nodes of this network, and the second half of this file describes the arcs of this network. Thanks to this file, we can analyze the constructed network in detail in the Pajek software later.

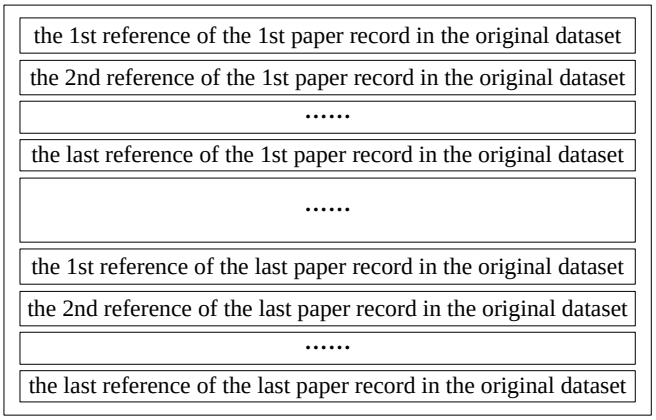



**Fig 1.** File structure

**Table 1** 10 fields unrecognized by Sci2

| **Field name** | **Meaning of the field** |
|---|---|
| U1 | Number of uses in the last 180 days |
| U2 | Number of uses since 2013 |
| EI | Electronic International Standard Serial Number, eISSN |
| DA | Generation data of this report |
| PM | PubMed PMID |
| OA | Indicator of open access |



*Published by: The Mattingley Publishing Co., Inc.*

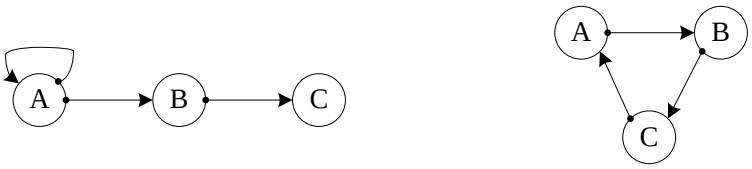
July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

| | |
|---|---|
| HC | Indictor of ESI high citation paper |
| HP | Indicator of ESI hot paper |
| EA | Month of early access |
| EY | Year of early access |

**Second Step: Make the Network Acyclic**

The Direct Paper Citation Network is not Acyclic with a high possibility, and this subsection aims to make it acyclic in order that current techniques for main path analysis can be used. Several classical algorithms have been embedded in Pajek, and this makes Pajek to be the main software for main path studies and very large network analysis. However, all the operations of main path studies in Pajek are based on acyclic network. It is very common that a large direct paper citation network is not acyclic, and the original network that we have obtained in this study is indeed not acyclic. However, few studies are found to explain in detail that how to make a direct paper citation network acyclic. This section aims to solve this problem. The following studies are based on the Extracted Paper-Citation Network that we have obtained using Sci2, and this network contains 43302 vertices and 51182 arcs. The main software we use in this section is 64-bit Pajek of version 5.06a.

A network is not acyclic because self-loops or cycles exist. Fig 2 shows what a self-loop is and what a cycle is. To make the network acyclic, both self-loops and cycles should be removed. The reasons that why self-loops and cycles exist are various. For example, self-citations; the delay between the time that a paper is formally published and the time that it can be cited as a reference; database error, and so on.




(1) Directed Network with a self-loop (2) Directed network with a cycle

**Fig 2.** Self-Loops and Cycles in directed paper-citation network

***In order to remove the self-loops as shown in Fig 2. (1),*** existed functions in Pajek could be used. The Extracted Paper-Citation Network is first provided to Pajek as input, then the function of Get Loops is run to find self-loops and a vector file called Get Loops is created automatically. The Get Loops file provides a description for all the nodes. Originally, all the vector values for the network nodes are 0. If self-loop appears for a node, then the value for this node is set to be 1. By checking the vector values, the nodes with self-loops can be fixed. In our previous attempts, we do find self-loops in a larger network, and the reason for these self-loops is database error. But fortunately, no self-loops are found in this network.

***After checking the self-loops, the following contents aims to remove the cycles.*** There is no



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

existed tool on hand to remove the cycles as shown in Fig 2. (2), and we propose our own theory and method in this section. The theoretical basis of our method is the **Directional Property of Knowledge Flows**. Concretely, knowledges should always flow from earlier papers to later papers. By removing the arcs which violate this rule, we can thus remove the cycles in the direct paper citation network. Assume that paper C is published later than paper B, and paper B is published later than paper A. Fig 3 shows that for paper A and paper B, knowledge should flow only from paper A to paper B. Fig 4 shows that the arc C→A violates the Directional Property of Knowledge Flows, and remove this arc could make this network acyclic. Arcs such as C→A should not exist, but they indeed exist in practice, and we thus name them as False Arcs. Technical details about removing False Arcs in order to remove cycles are demonstrated in the following procedures, and illustrations about these procedures are demonstrated in the subsequent contents.

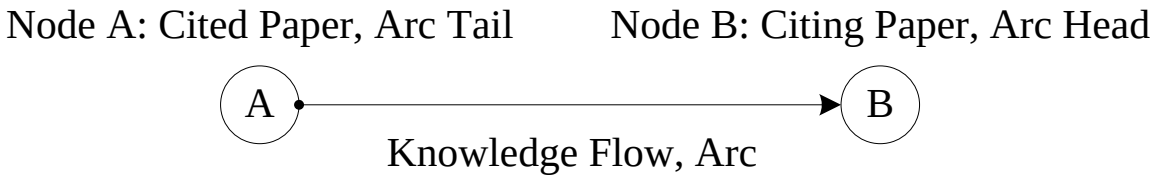



**Fig 4.** Remove the Cycle

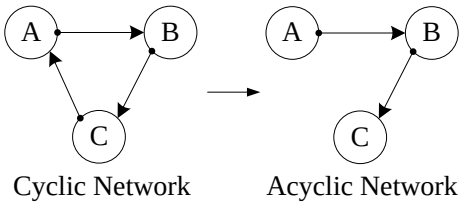



**Fig 3.** Directional Property of Knowledge Flows

**Procedures:**

Step 1. Find the subnetworks in which cycles exist.

(1) Reduce the original network by confining the input degree of each node to be at least 1.

(2) Continue to reduce the above network by confining the output degree of each node to be at least 1.

(3) Obtain the subnetwork in which cycles exist.

Step 2. Find, in this subnetwork, the false arcs that should be removed.

(1) Confirm the publication time of all papers (vertices) in the subnetwork.

(2) Record all the false arcs whose arc tail is published later than the arc head.

Step 3. Remove these false arcs in the corresponding original network.

(1) Find the false arcs in the corresponding original network.

(2) Remove these false arcs from the original network to obtain an acyclic network.

The above procedure demonstrates how to transform



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

a cyclic paper-citation network into an acyclic one. Now, we will give some explanations in detail about this procedure. The original paper citation network we have obtained is a bit large, containing 43302 nodes and 51182 arcs. To find false arcs directly in this original network is not easy. Thus, launching data process in the reduced subnetwork in which the cycles exist will reduce the workloads enormously. Step 1 aims to achieve this goal. The subnetwork in which cycles exist has a common feature: the input degree of each node is at least 1 and the output degree of each node is also at least 1. By confining the input degree and output degree, the subnetwork that contains cycles can be obtained. Pajek provides a function to reduce a network by the input and output degree of its nodes, and this function is called Reduction whose parent function is Create New Network. By running this function, a subnetwork containing only 7 vertices and 17 arcs is obtained. This subnetwork is constructed directly based on the original Extracted Paper-Citation Network, as no self-loops are found and thus no new network is constructed after it. We name the Extracted Paper-Citation Network as the original network. Compared with the original network, this subnetwork makes subsequent work much easier. The false arcs will be found in this subnetwork. In a paper citation network, the arc tail is the cited reference of the arc head. According to the Directional Property of Knowledge Flows, arc tail should be published earlier than arc head. Step 2 tries to find the false arcs by comparing the publication time of the tail and the head of each arc in the subnetwork. These comparisons are launched in three ways, through comparing the data field of Publication Year, through comparing the data field of DOI, and through checking the publication time in actual papers. These three ways are launched progressively, and not necessarily all these three ways of comparisons should be done. For example, if we can make sure the publication time sequence of the arc tail and the arc head by comparing their data field of Publication Year, then we will no longer compare the DOI or check the actual papers. What we actually want to do is to fix the publication time sequence of the arc tail and the arc head. To achieve this goal, the cyclic subnetwork is first loaded into Excel. Then, the data fields of Publication Year and DOI of each node are extracted. The Publication Year of the tail and the head of each arc are compared. If the Publication Year of an arc tail is later than that of an arc head, then this is a false arc and should be recorded. If the Publication Year of an arc tail and an arc head is the same, then DOI of these nodes will be compared. DOI of papers published in different journals cannot be compared, because they have different format. However, DOI of papers published in the same journal could be compared to confirm the publication time sequence of these papers. In this study, the subnetwork containing cycles is very small, only 7 vertices and 17 arcs in it, and we can operate manually without difficulty. However, in previous studies, we meet much larger subnetwork and it is formidable for scholars to operate manually. Thus, we have written some VBA programs to automatically launch the extractions of the above-mentioned data fields and comparisons of these data fields. If the publication sequence still cannot be confirmed, the corresponding papers should be downloaded, and concrete publication time could be confirmed. Generally, the publication sequence of papers could certainly be confirmed in these three ways, and all the false arcs in the subnetwork are recorded when Step 2 is finished. Having find out all false arcs, they should be removed from the parent network, not from the subnetwork directly. The subnetwork is only used to easily find out the false arcs. However, because it is a new network compared with the parent network, the sequence number of vertices are changed. Thus, Step 3 first finds out the false arcs in the parent network. This can be achieved by comparing the DOI of the vertices in the subnetwork and in the parent network, as the DOI of a concrete paper will never be changed in different networks. Finally, 6 false arcs are found



Published by: The Mattingley Publishing Co., Inc.

*July-August 2020*
*ISSN: 0193-4120 Page No. 4324 - 4345*

as shown in Table 3, and these 6 false arcs correspond with 7 vertices whose bibliometric information are shown in Table 4. Changes of the sequence number of these vertices are also shown in Table 4. After these false arcs are confirmed, they will be directly removed from the parent network. Till now, all these procedures are finished, and the parent network is made acyclic.

**Table 3** 6 false arcs

| **Arcs in the subnetwork** | | **Arcs in the original network** | |
|---|---|---|---|
| **Arc Tail** | **Arc Head** | **Arc Tail** | **Arc Head** |
| 5 | 6 | 8500 | 8561 |
| 1 | 2 | 870 | 7393 |
| 1 | 4 | 870 | 7409 |
| 2 | 3 | 7393 | 7400 |
| 4 | 3 | 7409 | 7400 |
| 7 | 2 | 10822 | 7393 |

**Table 4** Bibliometric information of the 7 vertices

| **Vertices** | **Doi** | **Bibliometric information** |
|---|---|---|
| 1 (870) | 10.1016/j.respol.2012.03.011 | Research Policy, 2012, 41(7): 1205-1218 |
| 2 (7393) | 10.1016/j.respol.2012.03.010 | Research Policy, 2012, 41(7): 1182-1204 |
| 3 (7400) | 10.1016/j.respol.2012.03.008 | Research Policy, 2012, 41(7): 1132-1153 |
| 4 (7409) | 10.1016/j.respol.2012.03.009 | Research Policy, 2012, 41(7): 1154-1181 |
| 5 (8500) | 10.1016/j.dsr2.2010.02.015 | Deep-Sea Research Part II, 2010, 57(16): 1460-1477 |
| 6 (8561) | 10.1016/j.dsr2.2010.02.014 | Deep-Sea Research Part II, 2010, 57(16): 1446-1459 |
| 7 (10822) | 10.1016/j.respol.2012.03.012 | Research Policy, 2012, 41(7): 1219-1239 |

**Note:** the sequence numbers of vertices in the subnetwork and in the original network are changed. The sequence numbers outside the brackets are from the subnetwork, and the sequence numbers inside the brackets are from the original network.



*Published by: The Mattingley Publishing Co., Inc.*

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

We further analyze these 6 false arcs, and find that they mainly appear in two journals, Research Policy and Deep-Sea Research Part II. Still, the arc tail and the arc head of each false arcs are published in the same issue of the same journal. We download the 7 papers corresponding with these false arcs in order to see the details. We find that, (1) in each of these false arcs, the arc tail and arc head are papers published in the same issue (This is a special issue named "Exploring the Emerging Knowledge Base of 'The Knowledge Society'"), in the same volume, and in the same journal; (2) in all the related 7 papers, the citation relationships indeed exist for these 6 false arcs. For example, the arc (6, 5) is recognized as a false arc as paper 5 is seen published before than paper 6 according to their DOI, however, paper 5 has indeed cited paper 6 as a reference; (3) the time difference, between the time that a paper can be cited and the time that a paper is formally published, may be the reason that why arc tails and arc heads could cite each other at the same time. However, the reason that why the cycles are caused should be further investigated. This is our future work and out of the scope of this study.

The operations about removing these false arcs in order to remove the cycles are based on the original network. Actually, the technique in this section can be used to remove both self-loops and cycles, because self-loops also violate the Directional Property of Knowledge Flows. Thus, the two steps of removing self-loops and removing cycles can also be finished in one step.

**Find main paths**

This section finds main paths of the research filed of main-path studies, in the newly generated subnetwork in which the Cycles have been removed. This process can be divided into 2 steps. The first step is to weight the importance of the links. The importance of links is measured by Traversal Weights. If a link has bigger traversal weights, then it plays a more important role in the process of knowledge dissemination. Pajek provides 3 ways, SPC, SPLC and SPNP, to compute traversal weights. Usually, SPC is recommended as it is considered to have some special benefits. Thus, in this study, we adopt SPC to compute the traversal weights. Details of SPC will be introduced in the following contents. The second step is to search the main paths which contains the most knowledge flows. Still, several ways are available to search the main paths, for example, Local Forward, Local Backward, Global, Local Key Route, and Global Key Route. Different methods have different advantages, and we adopt the Local Forward method in this study. All in all, in this empirical study, we adopt SPC Local Forward searching method to find the concrete main paths.

SPC Local Forward searching consists of three parts, SPC, Local, and Forward. SPC, Search Path Count, is the total number of times that a link is traversed which is proposed by Batagelj (Batagelj 2003). SPC is recommended over SPLC and SPNP by Batagelj as it owns some nice properties. Many followed studies adopt SPC to measure the importance of links, and we also use it in this empirical study. In order to explain in detail that how SPC works, in Fig 5, we borrow the figure that Liu has ever used (Liu 2012). In the paper citation network shown in Fig 5, nodes A and B are called Source, as there are only knowledge outflows for these two nodes. Correspondingly, nodes C, D, E and F are called Sink, as there are only knowledge inflows for these four nodes. Various paths can be found from a source to a sink, and this represents knowledge flows in different ways. However, different paths could carry different amounts of knowledges, and the one that carries the most knowledges is called main path. Hummon suggests that, begin with the source, a link owning the largest SPC is fixed each time, and then this process is repeated when the sink node is touched (Hummon 1989). For example, if we begin with



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

Source B, then the link B->I is first fixed, followed by the link I->G, G->D or G->E. As link G->D and G->E have the same SPC, thus two main paths are found. The first main path is B->I->G->D, and the second main path is B->I->G->E. These two main paths carry the same amounts of knowledges, but they may lead to different research directions.

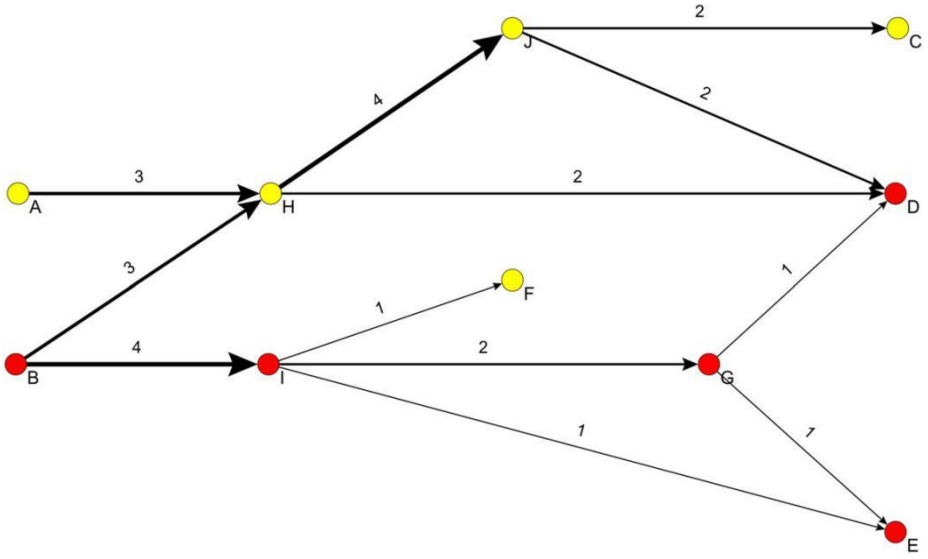



**Fig 5** A simple paper citation network

Based on the acyclic network obtained, we adopt the above SPC Local Forward searching method and get the main path as shown in Fig 6. The bibliometric information of each node on the main paths is shown in Table 5, in order that they can be easily obtained.

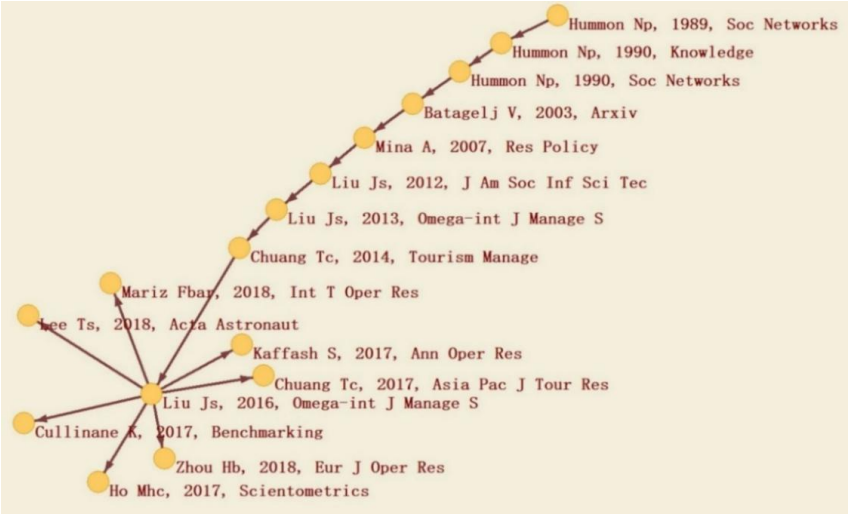



**Fig 6** SPC local forward search main path

**Table 5** Details of the nodes on the main paths

| First | Yea | Journal | Volumn | Begin | DOI |
|---|---|---|---|---|---|



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345


| author | r | | | page | |
|---|---|---|---|---|---|
| Hummon Np | 1989 | Soc Networks | V11 | P39 | Doi 10.1016/0378-8733(89)90017-8 |
| Hummon Np | 1990 | Knowledge | V11 | P459 | Doi 10.1177/107554709001100405 |
| Hummon Np | 1990 | Soc Networks | V12 | P273 | Doi 10.1016/0378-8733(90)90011-w |
| Batagelj V | 2003 | Arxiv | V0309023 | P1 | Doi 10.1111/1111 |
| Mina A | 2007 | Res Policy | V36 | P789 | Doi 10.1016/j.respol.2006.12.007 |
| Liu Js | 2012 | J Am Soc Inf Sci Tec | V63 | P528 | Doi 10.1002/asi.21692 |
| Liu Js | 2013 | Omega-int J Manage S | V41 | P893 | Doi 10.1016/j.omega.2012.11.004 |
| Chuang Tc | 2014 | Tourism Manage | V45 | P49 | Doi 10.1016/j.tourman.2014.03.016 |
| Liu Js | 2016 | Omega-int J Manage S | V58 | P33 | Doi 10.1016/j.omega.2015.04.004 |
| Kaffash S | 2017 | Ann Oper Res | V253 | P307 | Doi 10.1007/s10479-016-2294-1 |
| Chuang Tc | 2017 | Asia Pac J Tour Res | V22 | P213 | Doi 10.1080/10941665.2016.1220963 |
| Ho Mhc | 2017 | Scientometrics | V110 | P65 | Doi 10.1007/s11192-016-2158-0 |
| Cullinane K | 2017 | Benchmarking | V24 | P1552 | Doi 10.1108/bij-01-2016-0015 |
| Chuang Tc | 2017 | Asia Pac J Tour Res | V22 | P213 | Doi 10.1080/10941665.2016.1220963 |
| Ho Mhc | 2017 | Scientometrics | V110 | P65 | Doi 10.1007/s11192-016-2158-0 |
| Cullinane K | 2017 | Benchmarking | V24 | P1552 | Doi 10.1108/bij-01-2016-0015 |
| Lee Ts | 2018 | Acta Astronaut | V143 | P169 | Doi 10.1016/j.actaastro.2017.11.032 |
| Zhou Hb | 2018 | Eur J Oper Res | V264 | P1 | Doi 10.1016/j.ejor2017.06.023 |
| Mariz Fbar | 2018 | Int T Oper Res | V25 | P469 | Doi 10.1111/itor.12468 |


## Conclusions



Published by: The Mattingley Publishing Co., Inc.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

This paper has demonstrated systematically for beginners that how to find main paths in a concrete research field. During this process, we have also tried to address some technical problems. Concretely, we have proposed the backward-forward data collection method by which a comparatively more complete dataset can be obtained; have proposed a technical method to transform an acyclic network into a cyclic one, and the theoretical base of this technical method is the directional property of knowledge flows; have launched an empirical main path analysis for the research subject of the main-path study itself; have explained internally in detail that how SCI2 works. Through this paper, beginners could grasp the technique of main path analysis easily, and experts could find optional methods for main path finding. Our future studies aim to distinguish which database can meet our needs for main paths finding more accurately. Is it Web of Science, Scopus, or others? We will also launch comparatively studies to make it clear that how to obtain more convincing main paths.

**Acknowledgements**

We thank all the funding providers. This work is supported by China Postdoctoral Science Foundation under grant 2016M600763; the Major bidding projects of National Social Science Foundation of China 19ZDA309; the National Natural Science Foundation of China under grants 11671244, 61602291. We thank Huang Licheng for his help during our developing of the computer programs.

**References**

1. Batagelj, V., (2003). Efficient Algorithms for Citation Network Analysis. arXiv:cs/0309023v1 [cs.DL]
2. Carlero-Medina, C., & Noyons, E.C.M. (2008). Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. Journal of Informetrics, 2(4), 272–279.
3. Carley, K.M., Hummon, N.P., & Harty, M. (1993). Scientific influence: An analysis of the main path structure in the Journal of Conflict Resolution. Knowledge: Creation, Diffusion, Utilization, 14(4), 417–447.
4. Chuang, T. C. , Liu, J. S. , Lu, L. Y. Y. , & Lee, Y. . (2014). The main paths of medical tourism: from transplantation to beautification. Tourism Management, 45, 49-58.
5. Chuang, T. C., Liu, J. S., Lu, L. Y., Tseng, F. M., Lee, Y., & Chang, C. T. (2017). The main paths of eTourism: trends of managing tourism through Internet. Asia Pacific Journal of Tourism Research, 22(2), 213-231.
6. Dosi, G. (1982). Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. Research Policy 11, 147–162.
7. Garfield E, Sher IH, and Torpie RJ. (1964). The Use of Citation Data in Writing the History of
8. Garfield, E. (1955). Citation indexes for science. Science, 122(3159), 108-111.
9. Garfield, E.G., (1970). Location of milestone papers through citation networks. Essays of an Information Scientist, 1(34), 139-141.
10. Harris, J.K., Luke, D.A., Zuckerman, R.B., & Shelton, S.C. (2009). Forty years of secondhand smoke research: The gap between discovery and delivery. American Journal of Preventive Medicine, 36(6), 538–548.

July-August 2020
ISSN: 0193-4120 Page No. 4324 - 4345

11. Ho, M. H. C., Liu, J. S., & Chang, K. C. T. (2017). To include or not: the role of review papers in citation-based analysis. Scientometrics, 110(1), 65-76.

12. Ho, M. H. C., Liu, J. S., & Chang, K. C. T. (2017). To include or not: the role of review papers in citation-based analysis. Scientometrics, 110(1), 65-76.

13. Hummon, N. P. & Doreian P. (1989). Connectivity in a citation network: the development of DNA theory. Social Networks, 11(1), 39-63.

14. Hummon, N. P., Doreian, P., & Freeman, L. C. (1990). Analyzing the structure of the centrality-productivity literature created between 1948 and 1979. Knowledge, 11(4), 459-480.

15. Hummon, N.P., & Carley, K. (1993). Social networks as normal science. Social Networks, 15(1), 71–106.

16. Hummon, N.P., Doreian, P.,&Freeman, L.C. (1990).Analyzing the structure of the centrality–roductivity literature created between 1948 and 1979. Science Communication, 11(4), 459–480.

17. Jaffe, A.B., & Trajtenberg, M., (2002). Patents, Citations, and Innovations: A Window on the Knowledge Economy, MIT Press, Cambridge and London.

18. Kaffash S., & Marra, M. (2017). Data envelopment analysis in financial services: a citations network analysis of banks, insurance companies and money market funds. Annals of Operations Research, 253: 307-344.

19. Kessler, M.M. (1963). Bibliographic coupling between scientific papers. American Documentation, 14: 10-25.

20. Lee, T. S., Lee, Y. S., Lee, J., & Chang, B. C. (2018). Analysis of the intellectual structure of human space exploration research using a bibliometric approach: Focus on human related factors. Acta Astronautica, 143, 169-182.

21. Liu, J. S. , & Lu, L. Y. Y. . (2012). An integrated approach for main path analysis: development of the hirsch index as an example. Journal of the American Society for Information Science and Technology, 63(3), 528-542.

22. Liu, J. S. , Lu, L. Y. Y. , Lu, W. M. (2016). Research fronts in data envelopment analysis. Omega, 58, 33-45.

23. Liu, J. S. , Lu, L. Y. Y. , Lu, W. M. , & Lin, B. J. Y. . (2013-a). Data envelopment analysis 1978–2010: a citation-based literature survey. Omega, 41(1), 3-15.

24. Liu, J. S. , Lu, L. Y. Y. , Lu, W. M. , & Lin, B. J. Y. . (2013-b). A survey of DEA applications. Omega, 41(5), 893-902.

25. Liu, J. S., Lu, L. Y., & Lu, W. M. (2016). Research fronts in data envelopment analysis. Omega, 58, 33-45.

26. Lu, L. Y. Y. , Lin, B. J. Y. , Liu, J. S. , & Yu, C. Y. . (2012). Ethics in nanotechnology: what's being done? what's missing?. Journal of Business Ethics, 109(4), 583-598.

27. Lucio-Arias, D., & Leydesdorff, L. (2008). Main-path analysis and path dependent transitions in HistCite-based historiograms. Journal of the American Society for Information Science and Technology, 59(12), 1948–1962.

28. Mina, A. , Ramlogan, R. , Tampubolon, G. , & Metcalfe, J. S. . (2007). Mapping evolutionary trajectories: applications to the growth and



Published by: The Mattingley Publishing Co., Inc.

*July-August 2020*
*ISSN: 0193-4120 Page No. 4324 - 4345*

transformation of medical knowledge. Research Policy, 36(5), 789-806.

29. Moore, S., Haines, V., Hawe, P., & Shiell, A. (2006). Lost in translation: A genealogy of the "social capital" concept in public health. Journal of Epidemiology & Community Health, 60, 729–734.

30. Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, https://sci2.cns.iu.edu.

31. Science. Philadelphia: The Institute for Scientific Information, December 1964.

32. Shao, Z. , Li, Y. , Wu, K. , Guo, Y. , Feng, F. , & Hui, F. , et al. (2018). How academic librarians involve and contribute in research activities of universities? a systematic demonstration in practice through comparative studies of research productivities and research impacts. The Journal of Academic Librarianship.

33. 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: 265-269.

34. Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. Advances in Complex Systems, 10(1), 93–115.

35. Zhou, H. , Yang, Y. , Chen, Y. , & Zhu, J. . (2017). Data envelopment analysis application in sustainability: the origins, development and future directions. European Journal of Operational Research, 264(1), 1-16.



*Published by: The Mattingley Publishing Co., Inc.*