Scientometrics (2019) 119:379–391
https://doi.org/10.1007/s11192-019-03034-x

A few notes on main path analysis

John S. Liu[1] · **Louis Y. Y. Lu**[2] · **Mei Hsiu-Ching Ho**[1]

Received: 4 September 2018 / Published online: 25 February 2019

Abstract
The last few years have seen a growing interest in main path analysis among scholars across a wide spectrum of disciplines. Hummon and Doreian first introduced this method, and it has since become an effective technique for mapping technological trajectories, exploring scientific knowledge flows, and conducting literature reviews. Nevertheless, there are issues not broadly discussed in applying the method, including the handling of citation data, choosing a proper traversal weight scheme, search options, and interpretation of the resulting paths. This note aims to deepen the discussions and concludes with several suggestions and strategies in applying main path analysis.

Keywords Main path analysis · Citation networks · Bibliometric analysis

Introduction

Amongst various measures and methods that exploit citation information, main path analysis (MPA) is one of the most interesting ones. It not only traces the development of a scientific or technological field, but also highlights critical documents in the field. The two most important contributions for MPA are from Hummon and Doreian (1989), who propose the method, and Batagelj (2003), who designs efficient algorithms that make the method feasible on large datasets.

Studies related to MPA have increased steadily in recent years and have aggregated into a tightly linked strand of the literature. Papers in this literature can be categorized into two types: those that enhance and extend the method as well as those that apply the method. The former focuses on methodology, and the latter dwells upon applications. Liu and Lu (2012), Yeo et al. (2014), Tu and Hsu (2016) and Kim and Shin (2018) are typical papers

✉ John S. Liu
johnliu@mail.ntust.edu.tw

Louis Y. Y. Lu
louislu@staurn.yzu.edu.tw

Mei Hsiu-Ching Ho
mei.ho@mail.ntust.edu.tw

[1] Graduate Institute of Technology Management, National Taiwan University of Science and Technology, Taipei, Taiwan

[2] College of Management, Yuan Ze University, 135 Yuan-Tung Road, Chung-Li, Taoyuan, Taiwan

of the 1st type. Traditional MPA provides only a single path, which is limited in real application as it ignores secondary development. The idea proposed in Liu and Lu (2012) allows for multiple paths and suggests ways to "zoom" in/out when using MPA as a lens to capture the major citation chains. Yeo et al. (2014) propose an aggregative approach as an alternative to a traversal weight and suggest 2nd-order Markov chains to conduct the path search. Tu and Hsu (2016) merge papers on the main path that have similar topics to form a "conceptual path". Kim and Shin (2018) combine MPA with technology juncture analysis to identify derivative paths emanating from some technology junctures on the main path.

The majority of MPA papers are application-oriented studies. These studies map technological trajectories (Verspagen 2007; Fontana et al. 2009; Lu et al. 2012; Consoli and Mina 2009), detect technological changes (Bekkers and Martinelli 2012; Lucio-Arias and Leydesdorff 2008), conduct literature reviews (Bhupatiraju et al. 2012; Calero-Medina and Noyons 2008; Colicchia and Strozzi 2012; Harris et al. 2011; Liu et al. 2016; Chuang et al. 2017; Lu and Liu 2013, 2016), etc. One common characteristic of these application-oriented studies is that they all attempt to trace historical development in order to illuminate the evolution of a scientific or technological field. In contrast to traditional citation-count analysis, progress in time is the essence and advantage of MPA. As suggested in Mina et al. (2007), the path identified by MPA "does not result from a horizontal count of the papers (patents) yields of the different years considered, which would result in a trail of the papers (patents) with the highest citation counts per year. This would not serve the purpose of showing the diachronic connectedness of the system".

MPA works on citation networks that are built from dataset consisting of documents such as patents, papers, or court decisions. Patent citation networks are usually used to explore technological issues (Epicoco 2013; Martinelli 2012) while paper citation networks are the basis for literature review and examining knowledge diffusion paths (Batagelj et al. 2017; Mina et al. 2007; Liu et al. 2013a, b). There is only one study in the literature so far using citation network constructed from court decisions (Liu et al. 2014).

It is not without reasons that MPA is widely used by academia in recent years. The drastic improvement in computation technology, including software and hardware, makes sophisticated network computation more feasible. In addition, the availability of packaged tools further makes the analysis accessible for scholars. In July 2001,[1] Pajek (Batagelj and Mrvar 1998; De Nooy et al. 2005), one of the most widely used social network analysis software, implemented MPA-related functions, which were further enhanced in July 2015 by adding the key-route approach and other variations (Liu and Lu 2012). It is around this time that one notices more fruitful MPA research. Although MPA-related research continues to grow, some issues regarding MPA applications, in particular the handling of citation data, the choice of traversal weight, branch searching, and interpretation of the results, are not well discussed. The purpose of this note is to discuss and clarify these issues.

This article continues in the following section on issues associated with citation data, which determine the structure of a citation network. "Citation data" section looks to clarify the usage of traversal weight. "Traversal weight" section elaborates on the branch search. "Searching for branches" section discusses on how to properly interpret the main path. The last section concludes.

[1] Revision history of Pajek can be found on the website http://mrvar.fdv.uni-lj.si/pajek/history.htm.

Revisiting main path analysis

MPA begins with a citation network constructed from documents along with their mutual citation relationships. In a citation network, the relationship between any two nodes is sensitive to directionality, and in theory one can never visit the same node twice following the directed link from node to node. This type of network is commonly called acyclic directed network (ADN) in graph theory. ADN can be constructed from any dataset, wherein any two objects exhibit referencing or inheritance relationships. Examples of datasets bearing such properties are those consisting of academic papers, patents, or court decisions.

Given an ADN, Hummon and Doreian (1989) propose a two-step procedure. The first step makes a distinction between citation links that are originally all equally important by assigning each link a value called "traversal weight". The second step searches along the links that have the highest traversal weight to obtain the most significant citation chain. The resulting chain is the "main path", because it is regarded as a representative path in the citation network.

Traversal weight

Traversal weight is one of the core concepts in MPA. In their sentinel paper, Hummon and Doreian (1989) suggest three types of traversal weight: search path link count (SPLC), search path node pair (SPNP), and node pair projection count (NPPC). Batagelj (2003) later suggests the search path count (SPC). Among these four traversal weights, NPPC is not suitable for large networks since its computation time is of the order N^2 (Batagelj 2003). We therefore do not discuss NPPC here. In this article, we use SPX to represent the remaining three variations of traversal weights.

Several terms are defined before defining the three types of traversal weights. In a citation network, sources are the nodes that are cited while referring to no other nodes. Sinks are the reverse; they refer to other nodes, but are not cited. Intermediates are the nodes that refer to and are cited by others. The ancestors of a target node are those that can reach the target through citation chains. The descendants of a target node are those that can be found through citation chains emanating out from the target. Each citation link can be thought of as an arrow with its sharp end as the head and the other end as the tail. The head node for a link is the node attached to the arrow head, while the tail node is node attached to the arrow tail.

A citation link's SPC is the number of times the link is traversed if one runs through all the possible citation chains from all the sources to all the sinks in a citation network. To find SPC for a specific link, one needs to enumerate all the possible citation chains that emanate from all the sources and terminate at all the sinks. A citation link's SPLC is the number of times the link is traversed if one runs through all the possible citation chains from all the ancestors of the tail node (including itself) to all the sinks. To obtain SPLC for a specific link, one needs to enumerate all the possible citation chains that emanate from all the ancestors of the tail node (including itself) and terminate at all the sinks. There are apparently much more citation chains to enumerate. SPNP adds further complications, which is the number of times the link is traversed if one runs through all the possible citation chains from all the ancestors of the tail node (including itself) to all the descendants of the head node (including itself). Thus, to obtain SPNP, one needs to enumerate all the possible citation chains that emanate from all the ancestors of the tail node (including itself)

and terminate at all the descendants of the head node (including itself). "Which one to use?" section further clarifies the differences among the three SPX.

Searching for main path

The second step of MPA searches for the most significant citation chain(s) based on the traversal weights obtained in the previous step. Hummon and Doreian (1989) adopt the priority search algorithm, which begins from a source and along the way makes the locally best choice when selecting the next stop until a sink is encountered. This is the reason why Liu and Lu (2012) call it the local search. Batagelj (2003) suggests the critical path algorithm, which is a concept from operations research. The approach identifies the citation chains with the highest overall traversal weights. The resulting main path bears the names such as top path (Martinelli 2012; Fontana et al. 2009) and global main path (Liu and Lu 2012).

The local and global main paths usually include only few documents in the dataset, because only the most significant citation chain is observed. One can nevertheless gather multiple local or global main paths to obtain a more holistic picture through the "network of main paths" concept originated from Hummon and Doreian (1989) and brought to attention by Verspagen (2007). A network of main paths is a union of all citation chains in interest and is so-called, because a main path is a network so does a collection of main paths that are merged together. An example of the network of main paths approach appears in Fontana et al. (2009), which merges the first few highest top paths so that paths at the next significance levels can be observed at the same time with the original main path.

As Liu and Lu (2012) point out, either the local search or the global search may not include the links with the highest traversal weight, and thus they introduce the key-route search to address this issue. The key-route search algorithm begins with a seed link, usually the link with the highest traversal weight, and then searches forward until a sink is hit and searches backward from the tail node until a source is hit. The resulting main paths are merged together as a network of main paths. Either the local or global search can be adopted when searching forward or backward. The algorithm guarantees that the top link is included in the resulting main path, because it is the seed link. One can certainly conduct the key-route search from as many as seed links possible—for example, selecting the top 5, 10 or 20 links as the seed. A larger number of seed links will reveal greater details of the major citation chains.

Citation data

A citation connects documents and determines the structure in a citation network. MPA results are thus heavily determined on the network structure and are sensitive to citation data. This section discusses issues related to citation data.

Loop and the solution

In theory, a document can only refer to prior documents such that no document will loop back to itself when tracing along citation chains in a citation network. Nevertheless, in reality, data extracted from a publication database, e.g., Web of Science (WoS), can include two-way or even three-way citation loops. A two-way citation loop has document A citing B, and B citing A, while a three-way citation loop has document A citing B, B citing C,

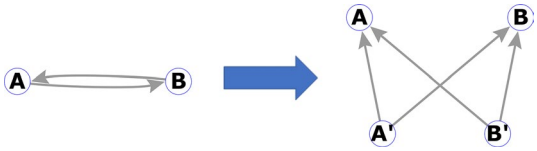

Fig. 1 Pre-print transformation proposed in Batagelj (2003)

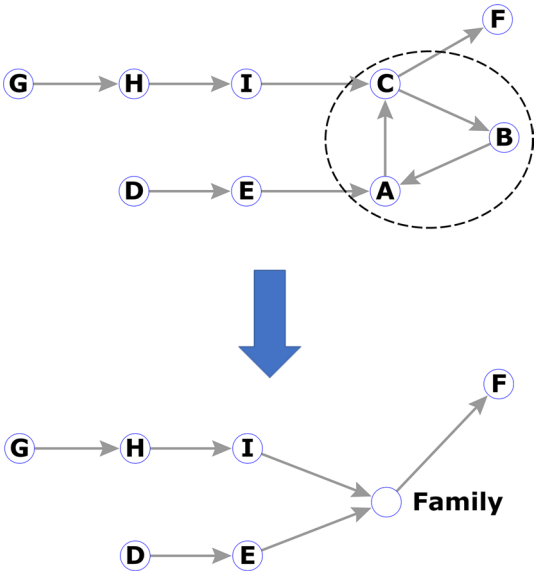

Fig. 2 Transforming loops into a family

and C citing A. Loops cause problems as certain documents will relentlessly come back to themselves in the path search process. There are three major causes for the existence of citation loops. First, the two documents are aware of each other, which often occur in papers published in special issues of the same journal. Second, a cited document for some reason is published at a later date than the citing document. Third, the database has errors in citation data. The solution for the third cause is straightforward—just correct the error.

For the two other causes, Batagelj (2003) mentions several approaches to resolve the problem. The more promising ones are those that reorganize citation data and transform/modify the almost acyclic network into an acyclic network. For that, there are three alternatives. First, delete one of the citation links within the loop. This approach effectively stops looping in the path search. As for which link to delete, it is an open question, but usually does not affect the result much. Second, add pre-prints to transform the loop into an acyclic structure. As shown in Fig. 1, a pre-print transformation replaces documents A and B with the pair (A, A′) and (B, B′), respectively, where A′ and B′ are the pre-prints of documents A and B. The pre-prints A′ and B′ replace documents A and B as the cited target, thus avoiding the loop. In Fig. 1, arrows indicate knowledge flow direction, whereby the citation is from the cited document (A′) to the citing document (A). Batagelj et al. (2017) apply this approach in a quantitative analysis of the "peer review" field.

The third alternative, also mentioned by (Batagelj 2003), is grouping the documents involved in a loop into a single "family", thus containing the loop within the family and making the loop disappear in the global structure. Figure 2 shows how the approach works.

In the figure, documents A, B, and C form a three-way loop. One treats them as a family and replaces them with a new node, say A′. All the inward and outward citations to A, B, and C are then aggregated to become the inward and outward citations to A′. After the transformation, the cyclic group is now integrated into one single entity A′ (labeled as Family in Fig. 2), and thus infinite looping is gone forever in the path search.

We suggest that the "family" approach is preferred over the other alternatives for two reasons. First, the documents in the loops can be closely related in topics thus are quite proper to examine them together. Second, the approach is in line the patent family concept. A patent family is a group of closely related patents. It is usually derived from the same core technology and consists of patents issued from patent office in various countries (OuYang et al. 2011). The approach allows us to investigate the importance of a certain patent family whether its members form a loop or not. In this regard, it is particularly useful for analyzing the position of a patent family. In summary, the approach is useful in examining the role of a document group, no matter whether they are papers of similar topics, patents of the same technology, or different court decisions on the same case.

The effects of self-citations

Brysbaert and Smyth (2011) and Hyland (2003) show that scholars tend to cite their own works in order to increase the scientific importance of their studies. Whether self-citation is "a neutral form of reporting" or "an unsavory kind of academic egotism" (Hyland 2003) has long been discussed in the academic community. Many past studies examine if self-citation creates a biased result on bibliometric analysis (Glänzel and Thijs 2004; MacRoberts and MacRoberts 1989, 1996; Hyland 2003; Glänzel et al. 2006). There even exist suggestions to remove self-citations before making comparisons on individual impact (Aksnes 2003).

Most prior discussions on self-citation center on its effect on citation count. From the MPA point of view, the effect of self-citation is more complicated than just reducing citation count. Removing self-citation changes the citation network structure and hence the main path results. Removing self-citation is good that unwanted linkages resulting from superfluous citations. Nevertheless, desired linkages are also removed at the same time. For a self-citation that reflects genuine knowledge diffusion, e.g., a thoughtful extension of a scholar's prior work, removing the citation wrongly ignores a true knowledge inheritance. As such, keeping or removing self-citations both have problems. Facing this to-be-or-not-to-be question, it seems that removing self-citation is riskier, while keeping self-citations is a better choice, because one would rather tolerate insignificant documents than risk losing significant knowledge diffusion information on the main path. The quandary can only be resolved by introducing techniques to differentiate the truthfulness of a self-citation. Before that happens, it is suggested that one should not eliminate self-citations when conducting MPA.

Traversal weight

This section discusses the differences among various traversal weights by applying a messenger and tollway analogy. The dissimilarities in measuring the impact of a paper via using traversal weight or citation count are also clarified.

Which one to use?

Most studies in the application-oriented literature adopt SPLC, SPNP, or SPC without detailing the rationale for their choices. Several studies have reported that different types of traversal weights produce similar results (Martinelli 2012; Batagelj 2003), i.e., the resulting main paths are almost the same even though individual links are assigned different type of traversal weights. It should be noted that traversal weights are very sensitive to the network structure. The behavior in one network structure is not guaranteed in another one. This is why (Batagelj 2003) suggests "… additional experiences should be gained from the analyses of real-life large citation networks" in order to understand the patterns resulting from different types of traversal weight. Here, we attempt to clarify the differences of these three measures by applying a messenger and tollway analogy.

We first assume that knowledge flow in a citation network is carried out by an imaginary messenger who takes knowledge from an origin document and sends it to a destination document through citation chains that connect the documents. For each pair of the specified origin and destination documents, many alternative paths exist running from the origin to the destination. While traversing the chains, the messenger is obliged to pay a toll when passing each citation link. For a citation link situated at a structural position where the messengers are more likely to pass through to complete the mission, it eventually collects more toll than those otherwise. As a reminder, the traversal weight for a citation link is the number of times the link is traversed upon a mission assigned to the messenger. In this analogy, the link's traversal weight is like the total toll the link can collect. At the end of a mission, the more toll a link collects, the higher the traversal weight is; and it is reasonable to then infer that the more significant the citation link is.

Under the messenger and tollway analogy, the three types of traversal weights are the results of three different messenger missions. The simplest mission is taking knowledge from all the sources and sending them to all the sinks running exhaustively through all possible paths. To be more precise, the messenger picks a pair of source-sink; takes knowledge from the source and sends it to the sink through one path; repeats the job on all possible paths linking the source and sink of the pair; picks another source-sink pair; and then repeats the job on all possible paths. The mission is completed when all source-sink pairs are exhausted. In this mission, origins are all the sources and destinations are all the sinks. Intermediates play the role of a middleman. After the mission is completed, the total toll collected for each citation link is its SPC.

In comparison with the 1st mission, in the 2nd mission the messenger takes knowledge not only from all the sources, but also from all the intermediates and sends them to all the sinks. It adds a complication into the 1st mission whereby all the intermediates also emanate out knowledge. At the end of the mission, the total toll collected for each citation link is its SPLC.

The corresponding mission that produces SPNP adds yet another complication to the 2nd mission. This mission takes knowledge from all the sources as well as intermediates and sends them to all the connectable intermediates as well as sinks, running exhaustively through all possible paths. All the intermediates are not only seen as origins of knowledge but are also seen as the final destinations.

From this messenger and tollway analogy, one can easily infer that $\mathrm{SPNP} \geq \mathrm{SPLC} \geq \mathrm{SPC}$, because the traversal weight increases with the complication of the missions. This has been shown mathematically in Batagelj (2003). Furthermore, in comparison with SPC, earlier citations "receive lower weights in SPLC because they

cannot be part of paths emanating from later articles" (De Nooy et al. 2005), and citations in the middle receive higher traversal weights for SPNP (De Nooy et al. 2005).

We can now discuss which traversal weight is the most suitable for MPA. Batagelj (2003) shows that SPC follows Kirchhoff's node law, which means that the sum of the inflow traversal weights is equal to that of the outflow traversal weights. Based on this "nice" property, a preference on SPC is suggested. Nevertheless, knowledge diffusion in a scientific and technological world works in a different way than the SPC mission analogy, in which the intermediates just pass on the knowledge. The intermediates, nevertheless, also generate knowledge. In the mission analogy for SPLC, intermediates not only pass knowledge through, but also play the role as the origins of knowledge. On the other hand, SPNP goes a little far, as it considers intermediates to also be knowledge depositories, which is certainly not the case. Therefore, among the three, the SPLC mission is the closest to the knowledge diffusion scenario in science and technology development.

In summary, for the purpose of tracing knowledge diffusion trajectory, we recommend using SPLC as it emulates most closely the knowledge diffusion scenario in the scientific and technological development where individual papers or patents not only pass knowledge, but also are knowledge sources themselves.

Traversal weight versus citation count

A commonly used index for measuring a paper's impact is citation count, which is the number of times a paper is cited by all literature. MPA provides another index for measurement, the document traversal weight. As discussed previously, SPX is defined on citation links. One can nevertheless extend the concept to a document (De Nooy et al. 2018). A document's SPX is the average SPX of all citations' links. It is obtained by dividing the SPX sum of all inward and outward links by the number of inward and outward links. A high-SPX document is more likely to be on the main paths and it can be regarded as having a great impact as it is situated in a massive knowledge flow juncture.

Whether document SPX or citation count is a better measure for a document's impact is an open question. Here, we clarify the major differences between the two. The first difference has something to do with research scope. Citation count is the number of times a paper is cited, no matter whether the citing paper belongs to the target research domain or not, while document SPX is usually defined within the target research domain. When calculating document SPX, one considers only knowledge flows within the documents in a dataset, whereas knowledge flows into and out of the research domain are not considered. Second, document SPX is an indirect index, while citation count is a direct one. Document SPX takes all indirect citations into consideration, but citation count considers only a document's immediate citation. The overall citation network structure determines document SPX, yet only the local network property decides citation count.

A paper with a high citation count does not necessarily have a high document SPX. As mentioned earlier, document SPX and citation count are defined on different bases. If a paper is cited by many papers that do not belong to the dataset of the research scope, then these citations do not contribute to the traversal weight. For example, paper A in Fig. 3 is cited by 6 documents, but citing documents D, E, and F are out of the research domain. They cannot contribute to the traversal weight of the link (S, A). We note here that arrows in the figure point to the citing documents. Furthermore, if the citing papers do not attract further citations, then the traversal weight of a highly cited document cannot be high.

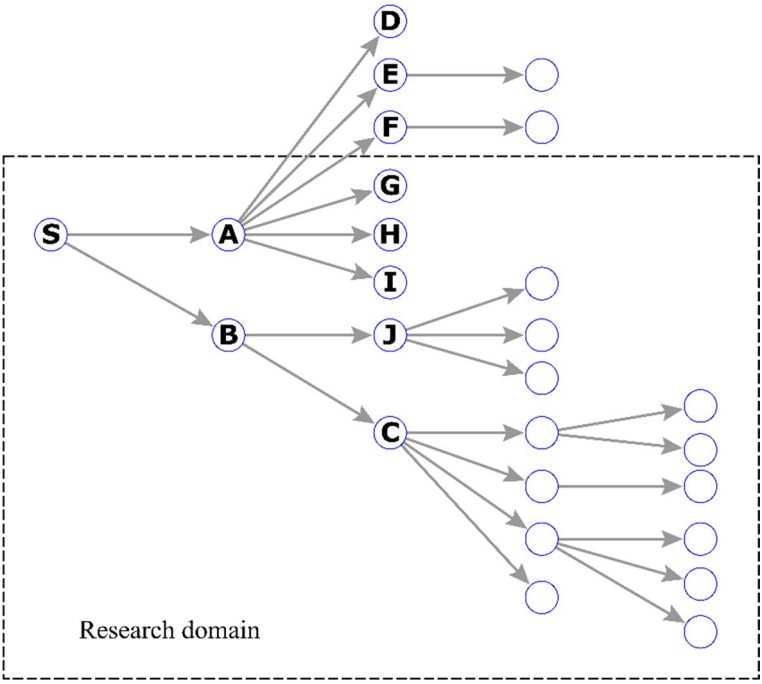

Fig. 3 A sample citation network

Document A's descendants G, H and I do not attract further citations, and so their contribution to (S, A)'s traversal weight is very limited.

A paper receiving a relatively low citation count may conversely have a high document SPX and is thus included in the main path. Fontana et al. (2009) and Barberá-Tomás et al. (2011) mention such a phenomenon. It is likely to be caused by the reverse-inheritance and the integrator effects, which are discussed later in "Interpretation of the results" section. In summary, document SPX is quite different from citation count and can be an interesting index to measure a document's impact, because it considers the effects of indirect citations.

Searching for branches

Branches are paths extending out from certain nodes on the main paths. Researchers are likely to be interested in examining branches for various reasons in practice. First, there is a need to look at the specifics in different development stages to better comprehend a scientific field or technology. These specifics are often outshined by a few significant paths and hence are not seen on the main paths. Second, one may want to examine why some seemly important documents are not on the main path—for example, highly cited academic papers or a market leader's patents. The branches leading to these documents provide us with information on their roles. Third, for the situation wherein earlier developments do not continue onto the later part of the main path, a previous major research topic may be overwhelmed by new research topics, or an earlier technology can be overridden by later technologies. Observing branches in such a situation is a good way to identify the later

developments of an earlier research topic or technology. All branches and the traditional main paths usually merge together to illustrate a holistic picture through the "network of main paths" concept. This section describes two major types of approaches adopted in the MPA literature for obtaining branches.

Time-based approach

The time-based approach obtains branches by identifying the main paths at different time periods and then merging the results into a network of main paths. A common treatment is fixing the starting year, while changing the ending year for each time period (Fontana et al. 2009; Martinelli 2012; Barberá-Tomás et al. 2011; Verspagen 2007). For example, in investigating the technological changes of telecommunication switches, Martinelli (2012) presents the network of main paths that combines the top paths for time periods 1924–1979, 1924–1984, 1924–1989, 1924–1994, 1924–1999 and 1924–2003, respectively, thus allowing us to observe significant developments in each time period.

Designated-document approach

The designated-document approach traces both forward and backward the knowledge flow of a designated document until encountering a node on the main path or a sink, hence revealing this document's relationship with the main path. This type of branch search illuminates the recent development paths that are overlooked by the traditional MPA. An example is shown in Ho et al. (2014), who conduct MPA on fuel cell papers. Proton-exchange membrane fuel cell (PEMFC) was the major technology of a fuel cell before 2000, but direct methanol fuel cell (DMFC) dominated the development after 2000. PEMFC development is overwhelmed by DFMFC and is thus only seen on the main path before 2004. After selecting some recent highly cited PEMFC papers and applying the branch search, the resulting network of main paths clearly depicts the more recent development of PEMFC and its relationship with the traditional main paths.

Interpretation of the results

When interpreting the MPA results, it is suggested to consider the reverse-inheritance effect and the integrator effect. Both are the phenomena resulting from the way the traversal weights are assigned. The reverse-inheritance effect increases the significance of a document's ancestor if the document itself is significant. Document B in Fig. 3 exhibits this effect. It is cited only by two documents, but its descendants, documents J and C bring honor to it. Many later documents cite documents J and C. Hence, the traversal weight of the link (S, B) is relatively big. The significance of the document is mainly contributed from J and C rather than by B itself. When interpreting the result, one cannot overemphasize the contribution of document B and should highlight that its honor is brought by the successful descendants.

The integrator effect increases the significance of a document that heavily references others. The effect is particularly noticeable in review papers, which usually cite many prior documents in order to integrate diverse perspectives that have been discussed in the past. In a dataset containing review papers, numerous streams of knowledge flow into the review papers, thus increasing the traversal weight of the links that cite the review paper. This is

one of the reasons why review papers are seen quite often on the main paths; the other being that they are usually highly cited themselves. Nevertheless, it is fair to say that the integrator effect boosts the significance of review papers. The main path results are usually different with and without review papers (Ho et al. 2017). Whether review papers should be included in the analysis is an interesting question. Including review papers helps to show the integrated development, but review papers can be noises that make the other developments unobservable. On the other hand, excluding review papers allows the observation of true theoretical development, but may result in missing papers that summarize the research field. Two alternatives are available to address the issue. One is recognizing the research purpose and making a contingent decision. The other is conducting the main path analysis with and without the review papers and then discussing and comparing both results.

Prior studies have reported that the documents identified by MPA are not necessarily the ones with highest citation counts (Fontana et al. 2009; Barberá-Tomás et al. 2011). Fontana et al. (2009) suggest that this is because they are important "at that point in time" and are thus positioned at a strategic 'junction' along the trajectory." Further to the interpretation, we propose from the network structure viewpoint that the integrator and reverse-inheritance effects can be the root cause for such a phenomenon. For example, patent 4751701 discussed in Fontana et al. (2009) references relatively more prior patents than the others.[2] The integrator effect is thus likely to boost its importance.

Conclusions

MPA maps complex citation networks into a single chain or combinations of significant citation chains. The results not only exhibit the main knowledge flow but sometimes also reveal interesting phenomenon such as the divergence-convergence development in technology (Hung et al. 2014). One may wonder whether the method artificially fabricates the phenomena, or the phenomena are uncovered by the method. Barberá-Tomás et al. (2011) conduct an external validation to the method using patents related to an artificial disc for treating spinal pain. They show that the main paths of artificial disc technology largely describe its evolutionary path and corroborate that MPA is a valid tool for identifying reliable knowledge flow. This suggests that the latter is closer to what truly happened. MPA in one sense is like the visualization techniques in fluid mechanics. One does not observe the turbulence in the fluid until dye is injected into it. MPA is the dye that helps visualize more clearly the knowledge flow.

This short note shares with the academic community our thoughts in applying MPA. We suggest that grouping documents together as a "family" is a good way to eliminate legitimate loops in a citation network, and that SPLC is a preferred choice for traversal weight as it fits the knowledge diffusion model better than the other traversal weights. Our suggestions extend to the technique of seeing branches within a large-scale development. We also discuss the reverse-inheritance effect and integrator effect, which need to be considered in interpreting the MPA results.

MPA is seen as a general tool that highlights historical events in order to illuminate the evolution of scientific and technological fields. Considering the enthusiasm given towards

[2] The number of references for 4751701 is 22, or much higher than that of its neighbors on the main paths, 5012467, 4560984 and 4598285, which are 10, 7 and 6, respectively.

improving the method and the need to quantitatively study the history of science and technology, future applications of MPA can go beyond our imagination.

Acknowledgements We thank two anonymous reviewers for their constructive comments which have greatly improved the accuracy and readability of this article. This work is partially supported by Taiwan's Ministry of Science and Technology grants MOST 105-2410-H-011-021-MY3, 107-2410-H-155-046, and 106-2410-H-011-028-MY2.

References

Aksnes, D. W. (2003). A macro study of self-citation. *Scientometrics, 56*(2), 235–246.

Barberá-Tomás, D., Jiménez-Sáez, F., & Castelló-Molina, I. (2011). Mapping the importance of the real world: The validity of connectivity analysis of patent citations networks. *Research Policy, 40*(3), 473–486.

Batagelj, V. (2003). Efficient algorithms for citation network analysis. In *Preprint series, University of Ljubljana, Institute of Mathematics, Physics and Mechanics, Department of Theoretical Computer Science*.

Batagelj, V., Ferligoj, A., & Squazzoni, F. (2017). The emergence of a field: a network analysis of research on peer review. *Scientometrics, 113*(1), 503–532.

Batagelj, V., & Mrvar, A. (1998). Pajek-program for large network analysis. *Connections, 21*(2), 47–57.

Bekkers, R., & Martinelli, A. (2012). Knowledge positions in high-tech markets: Trajectories, standards, strategies and true innovators. *Technological Forecasting and Social Change, 79*(7), 1192–1216.

Bhupatiraju, S., Nomaler, Ö., Triulzi, G., & Verspagen, B. (2012). Knowledge flows: Analyzing the core literature of innovation, entrepreneurship and science and technology studies. *Research Policy, 41*(7), 1205–1218.

Brysbaert, M., & Smyth, S. (2011). Self-enhancement in scientific research: The self-citation bias. *Psychologica Belgica, 51*(2), 129–137.

Calero-Medina, C., & Noyons, E. C. (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.

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.

Colicchia, C., & Strozzi, F. (2012). Supply chain risk management: A new methodology for a systematic literature review. *Supply Chain Management: An International Journal, 17*(4), 403–418.

Consoli, D., & Mina, A. (2009). An evolutionary perspective on health innovation systems. *Journal of Evolutionary Economics, 19*(2), 297.

De Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University Press.

De Nooy, W., Mrvar, A., & Batagelj, V. (2018). *Exploratory social network analysis with Pajek* (3rd ed.). Cambridge: Cambridge University Press.

Epicoco, M. (2013). Knowledge patterns and sources of leadership: Mapping the semiconductor miniaturization trajectory. *Research Policy, 42*(1), 180–195.

Fontana, R., Nuvolari, A., & Verspagen, B. (2009). Mapping technological trajectories as patent citation networks. An application to data communication standards. *Economics of Innovation and New Technology, 18*(4), 311–336.

Glänzel, W., Debackere, K., Thijs, B., & Schubert, A. (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics, 67*(2), 263–277.

Glänzel, W., & Thijs, B. (2004). The influence of author self-citations on bibliometric macro indicators. *Scientometrics, 59*(3), 281–310.

Harris, J. K., Beatty, K. E., Lecy, J. D., Cyr, J. M., & Shapiro, R. M. (2011). Mapping the multidisciplinary field of public health services and systems research. *American Journal of Preventive Medicine, 41*(1), 105–111.

Ho, J. C., Saw, E.-C., Lu, L. Y., & Liu, J. S. (2014). Technological barriers and research trends in fuel cell technologies: A citation network analysis. *Technological Forecasting and Social Change, 82,* 66–79.

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.

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

Hung, S.-C., Liu, J. S., Lu, L. Y., & Tseng, Y.-C. (2014). Technological change in lithium iron phosphate battery: The key-route main path analysis. *Scientometrics, 100*(1), 97–120.

Hyland, K. (2003). Self-citation and self-reference: Credibility and promotion in academic publication. *Journal of the American Society for Information Science and Technology, 54*(3), 251–259.

Kim, J., & Shin, J. (2018). Mapping extended technological trajectories: integration of main path, derivative paths, and technology junctures. *Scientometrics, 116*(3), 1439–1459.

Liu, J. S., Chen, H. H., Ho, M. H. C., & Li, Y. C. (2014). Citations with different levels of relevancy: Tracing the main paths of legal opinions. *Journal of the Association for Information Science and Technology, 65*(12), 2479–2488.

Liu, J. S., & Lu, L. Y. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. *Journal of the Association for Information Science and Technology, 63*(3), 528–542.

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

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

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

Lu, L. Y., Lan, Y., & Liu, J. S. (2012). A novel approach for exploring technological development trajectories. In: *2012 IEEE International Conference on Management of Innovation and Technology (ICMIT)* (pp. 504–509). IEEE.

Lu, L. Y., & Liu, J. S. (2013). An innovative approach to identify the knowledge diffusion path: The case of resource-based theory. *Scientometrics, 94*(1), 225–246.

Lu, L. Y., & Liu, J. S. (2016). A novel approach to identify the major research themes and development trajectory: The case of patenting research. *Technological Forecasting and Social Change, 103,* 71–82.

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

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for information Science, 40*(5), 342–349.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics, 36*(3), 435–444.

Martinelli, A. (2012). An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry. *Research Policy, 41*(2), 414–429.

Mina, A., Ramlogan, R., Tampubolon, G., & Metcalfe, J. S. (2007). Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge. *Research Policy, 36*(5), 789–806.

OuYang, K., Weng, C. S. J. T. F., & Change, S. (2011). A new comprehensive patent analysis approach for new product design in mechanical engineering. *Technological Forecasting and Social Change, 78*(7), 1183–1199.

Tu, Y. N., & Hsu, S. L. (2016). Constructing conceptual trajectory maps to trace the development of research fields. *Journal of the Association for Information Science and Technology, 67*(8), 2016–2031.

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems, 10*(01), 93–115.

Yeo, W., Kim, S., Lee, J.-M., & Kang, J. (2014). Aggregative and stochastic model of main path identification: A case study on graphene. *Scientometrics, 98*(1), 633–655.