Analysis and Prediction of Patent Trajectory

Yun-Chia Liang
Department of Industrial Engineering & Management
Innovation Center for Big Data and Digital Convergence
Yuan Ze University, Taoyuan, Taiwan
Tel: (+886) 3-463-8800 ext. 2521, Email: ycliang@saturn.yzu.edu.tw

Angela Hsiang-Ling Chen
Department of Marketing and Distribution Management
Nanya Institute of Technology, Taoyuan, Taiwan
Tel: (+886) 3-436-1070 ext. 5507, Email: achen@tiit.edu.tw

René Josué Aparicio Monjarás
Department of Industrial Engineering & Management
Yuan Ze University, Taoyuan, Taiwan
Email: s1035460@mail.yzu.edu.tw

Abstract. The best way to identify the flow of knowledge on a patent citation network is to apply a technique that can process the relationship between each patent and obtain as a result the most important discoveries that allow further evolution of the technology. Patents are a unique source of information containing not only legal but also technical, business and potentially policy-related information. Therefore, how to determine the life cycle of a technology and forecast its development for future years becomes the focus of this study. Fuel cell, a technology been long recognized as a good energy solution and an environmental friendly alternative to the fossil fuels. Fuel cell topics satisfy a series of criteria that makes it ideal as a case study. The purpose of this research is then to develop a main path analysis implementation for patent analysis case specifically on the fuel cell technology. By exploring this information, it's possible to determine the impact that has over different areas of an organization such as marketing, research and development, finance, etc. The outcomes of the case study shows that the proposed method is a useful tool to provide inputs and information on strategic decision and competitive intelligence done in a company or organization, such as investments, management decisions or venture in research for new technologies.

Keywords: Patent Analysis, Fuel Cell Technology, Main Path Analysis

1. INTRODUCTION

Back in 1997 every business day more than 200 new patents documents were issued worldwide (Karki, 1997), and by 2014 there were 2.68 million applications done worldwide with 44.03% granted during that same year[1]. So its noticeable the activity that patents provide yearly in this globalized world in which intellectual property is protected from any infringement or misuse of technologies (or knowledge) that doesn't recognize the rightful owner in eyes of the law. As it is well known patents represent a long-term intangible asset and has its own value for the company (or institution who owns it).

This brings relevance on the topic of patents and its sequential analysis, by the year 2014, having enforced a quantity of 10.2 million of patents worldwide. It is necessary to process this data with the main objective of obtaining useful information that can create possibilities to make decisions and benefit an organization (lucrative or non lucrative ones). For example, acquire a new technology or not, strengthen the research and development department, manage patents portfolios, investments, inputs for strategic business planning, etc.

The research focuses on fuel cell technology which in its general idea is devices that combine hydrogen fuel with oxygen to produce electricity, heat and water. This

---

[1] Data from World Intellectual Property Organization (WIPO)

technology has been recognized as a good energy solution in a world where the cost of fossil fuels and the climatic impact that produces are really high. It is necessary to look for some more environmental friendly alternatives. The investment in developing clean and renewable energy sources has increased over the last 5 years (Ho et al., 2014). Fuel cell topics satisfy a series of criteria that makes it ideal to a case study, i.e., large datasets of patents available, relatively new (the origins in published papers are around 1988 which determined the base for development of technology), high relevancy in the present world. Therefore, the purpose of this research is to develop a main path analysis implementation for a patent analysis case specifically the fuel cell technology, to help identify the most important discoveries since the first idea and establish the evolution of the novelty technology through the years.

2. LITERATURE REVIEW

Patents in their most basic concept are "the right to appropriate returns from research, and this in effect excludes other firms from practicing or producing the same process and products." When you put this information in writing (all the claims, descriptions and specifications of the knowledge) within a patent document, it makes it a legal resource in which the knowledge is protected by law against infringement (Reitzig, 2004). It is important to mention for a patent to be granted, the knowledge must be novel, inventive, industrially applicable and useful. Patents are a form of intellectual property which also contains copyright, industrial design rights, plant varieties, trade dress, trade secrets and trademarks. In a more technical term patent is defined by the United States Patent and Trademark Office (USPTO) as: "The right conferred by the patent grant is, in the language of the statute and of the grant itself, the right to exclude others from making, using, offering for sale, or selling the invention in the United States or importing the invention into the United States. What is granted is not the right to make, use, offer for sale, sell or import, but *the right to exclude others* from making, using, offering for sale, selling or importing the invention. Once a patent is issued, the patentee must enforce the patent without aid of the USPTO."

Patent analysis is a management tools for addressing the strategic management of the firm's technology and product or service development process. Converting patent data into competitive intelligence allows to measure its current technical competitiveness, to forecast trends and to plan for potential competition based on new technologies (Breitzman & Mogee, 2002). It is important to note that an analysis of patents does not involve a simple procedure of patent count for any organization of firm (von Wartburg et al., 2005). To determine the value of a patent it should also be addressed the technical value and the citations involved in the patent to describe it as a valuable asset. It is necessary to shift the type of focus to the patent citation analysis that concentrates on the study of the relationship among different patents.

Patent citation analysis approaches are studied to structure a large number of patents. The following are examples of patents citation analysis: Verspagen (2007) explored technological trajectories in fuel cell research by using citation analysis of patents. Cho & Shih (2011) identified core and emerging technologies for Taiwan technological innovation system. Lee et al. (2012) addressed the issue to assess future technological impacts based on patent citation information by employing the future citation count as a proxy. Breitzman & Thomas (2015) studied the phenomenon of forward citation which is valuable for impact evaluation. Altuntas & Dereli (2015) used patent citation analysis for prioritizing the portfolio of investment projects. These past examples are just a few of the applications of citation patents information.

Academic citations are explicit linkages between papers that have some important content in common. The idea of papers being linked by citations is the foundation on which the construction of the Science Citation Index (SCI) is based upon (Hummon & Dereian, 1989). Once these linkages are put together by mapping them, the result is a network conformed by components known as: nodes, links and flow. The sequences of links and nodes are called search paths, and to put it simple it's a path which starts in a specified node and ends in another one. Following one single direction, to travel between two nodes the number of routes can start from one single to hundreds all depending on the number of nodes (distance) between them. The citation network analysis started with the study of Garfield et al. (1964) when they demonstrated a relationship between discovery events on DNA fields and the citation relationship between these events.

The definition of main path analysis is provided by Liu & Lu (2012) which described as a bibliometric method capable of tracing the most significant paths in a citation network, and commonly used to trace the development trajectory of a research field. This is a technique conceived by Hummon & Dereian (1989), in which they proposed a different approach in regard of the importance of connectivity threads in a network and not only in the nodes of it, they first applied this method to the development of the DNA theory. The main path analysis has been applied to several different journal's citation networks in different fields such as the DNA theory (Hummon & Dereian, 1989), Hirsch Index (Liu & Lu, 2012), Data Envelopment Analysis (DEA) applications (Liu et al., 2013), and fuel cell technologies (Ho et al., 2014).

3. METHODOLOGY

In this section, the methodology for a patent analysis using main path analysis will be explained, with the objective to create a stepwise procedure to use on a case study which is fuel cell technology. By applying this methodology there is an opportunity to explore new approaches for patent analysis, taking into consideration the development of inventions over the years, and obtaining useful information to apply for decision making in areas such as marketing, research and development, investments, purchases, etc. Also this study represents an occasion to compare the results obtained via “patents” with the ones obtain through “research published papers” done by Ho et al. (2014). This section consists of three main phases – data extraction, network construction, and main path analysis.

Data extraction step consists of extracting all the data necessary to do the following analysis; this is done by scrapping the patents related to fuel cell technology, using the designated keywords and USPTO database in which the study will be based on. That means it is necessary to do a pre-processing on the data collected to eliminate any unfit data.

Afterwards, it will be necessary to construct the network based on the citations of the patents collected. This is done by first analyzing the citing relationships found in the database, identifying key elements such as the chronology, cyclic citations and transitive reduction, and starting to build the citation network for this research.

Once the network is obtained, it will be the time to apply the main path analysis technique. This is a two-step method. First, it will be necessary to compute the traversal count of the whole network which is a way to determine the weight or significance index of each link; the second step is to apply a search algorithm to determine which of the links has the most significance to build the main path of the complete network.

The details of each phase will be described in the following subsections.

3.1 Data Extraction

To extract the necessary data, the software called WebHarvy[2] is used, enabling to scrape the information from “google patents[3]”. The more useful data to apply the present methodology is the USPTO (United States Patent and Trademark Office) database which can be accessed through google. To perform an optimal query that can show all the related work in regard of fuel cell technology, this study used the keywords defined by the research done by Ho et al. (2014). They applied a citation analysis to fuel cell technologies from a literature review point of view (published articles in journals). For their search query they used the following keywords: “fuel cell technology,” “fuel cell*,” “fuel-cell,” “solid oxide,” “molten carbonate,” “phosphoric acid,” “proton exchange membrane” and “direct methanol.” By using these same keyword inputs as well, it helps provide the framework to compare the results obtained from two different perspectives (patents and published articles), and draw conclusion out of it.

3.2 Network Construction

After collecting the datasets for patents with the keywords of interest, it is necessary to start to build the network of the patent citation. In this point each patent data is a non-weighted, non-directed network, each patent represents a node, and backward citations act as links between each one of the nodes. To minimize noise data caused by patent, it is essential to remove the patents that neither has a citing element or is cited. The objective is to leave the network only with the patents that participate in the mapping process of the network and provides information in the overall study.

The building of the network requires the first step as analyzing the citing relationships found in the dataset, and afterwards building the citation network. These two steps are explained as follows.

To begin the construction of a network, it is necessary to identify the following elements: nodes, links, starting and ending points. For this study each node represents a different patent; the links are the citing relations.

A stating point or also known as “source” is a node that has been cited, but cites no other nodes. For most of the cases it is considered as a node being part of the oldest (an early invention). However, this case is not always necessarily true, because in some cases there are inventions that are completely new in every sense, not based in any existing technology present at that time. On the other hand, an ending point, also known as “sink” is a node that cites other nodes, but is not cited. In this case, it’s not true to state that these nodes are the newest chronologically speaking, since there are a lot of inventions based their work on other technologies, but afterwards do not provide any new or useful knowledge to continue working on.

In a citation network it is possible to have several sources nodes and several sink ones. Also, in the data collected it will be important to identify isolated nodes, since those patents do not interact with other inventions nor provide any useful information for the current study and

[2] WebHarvy: https://www.webharvy.com/

[3] Google offers an access to the databases more important in the world. Currently it hosts data from USPTO, EPO and WIPO. This arrangement started in 2010 and it’s an easy access to information (https://patents.google.com/).

will be removed from the dataset.

It is important to mention two constraints that any citation has to follow. First, a citation relation is not allowed to point forward in time, e.g. a patent of 2011 is not allowed to cite a patent of 2012. The second constraint is that a citation network must be acyclic; this means that it is not allowed to have a citation from patent A to patent B and a citation from patent B to patent A.

Another concept to incorporate in the construction of network is "Transitive Reduction" as shown in Figure 1 which introduces the distinction between what is referred to as essential and non-essential citations relations. A citation relation is considered essential if, apart from a relation between patent A and patent B there are no other paths in the citation network connecting these two nodes. In the other words, it will be considered as non-essential if beside the direct path between patent A and patent B, there exists other paths such as a citation from patent A to patent C and from patent C to patent B. Transitive reduction of a citation is obtained by removing all non-essential citations from the network (van Eck & Waltman, 2014; Clough et al., 2015).

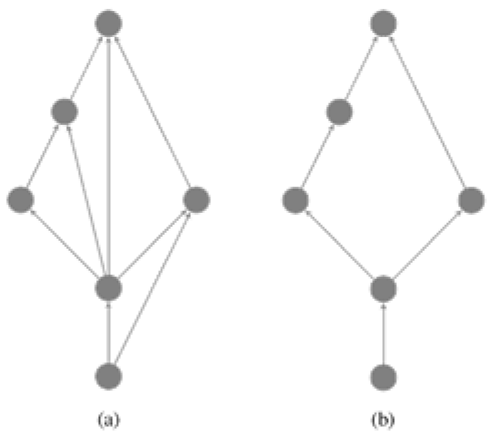

Figure 1: Illustration of transitive reduction. (a) original citation. (b) a citation network after transitive reduction. (van Eck & Waltman, 2014)

Once analyzing the citation relations in the dataset, it's time to start building the citation network. This can be done depending on the size of the network to build. The easiest way is through a software tool, in this research it will be Gephi[4] and Pajek[5], which are ideal for building type of networks in a practical way.

3.3 Main Path Analysis

As it was established beforehand, in a citation network the links among each node are conduits of knowledge flows. Consequently a sequence of conduits that link a given node to an end node is called "search path" (Liu & Lu, 2012). The Main Path Analysis technique identifies the main path, which is the most significant search path among the whole citation network constructed in section 錯誤! 找不到參照來源。. It's vital to clarify that there can be several search paths; however, by defining criteria of selection (for example a condition to satisfy) it's possible to limit to the most relevant ones.

The first step is to translate the citation network into a weighted network, with the weight of each link indicating the significance of the link. This is done through a concept known as "traversal counts", and this acts like a "significance index" for the network. The second step applies a search algorithm based on the traversal count mentioned before, with the objective to identify the key components of the main path. These significance paths increase exponentially with the total number of nodes in a citation network (Hung et al., 2014).

Traversal counts measure the times a citation link has been traversed, if one exhausts the search from a set of start nodes to another set of end nodes in the built network. If a citation occupies a route through which much knowledge flows, it has to have a certain importance in the knowledge-dissemination process. Mainly there are three types of traversal counts that can be used in networks. The one to use in this research is Search Path Count (SPC) based on all possible search paths emanating from a start node. SPC is recognized for its simplicity and versatility to represent and calculate the weights for each link in the network. It is also highly recommended by different authors who work on main path analysis research such as Liu & Lu (2012), and Hung et al. (2014).

A citation link's SPC is the number of times the link is traversed if one exhausts the search from all the sources to all the sinks in a citation network. A link with a large SPC indicates that it sits on a large number of diffusion paths making it important in the evolution of the field.

The main path analysis procedure is based on the simple idea to find a main path, beginning to search from all nodes that cite no others and picks the link with the largest traversal count emanating from these nodes as the start link (Liu & Lu, 2012). The exact procedure is summarized as follows:

1. Find the link with the largest traversal count from all possible links emanating from all the sources. Assign the beginning node of this link as the start point of the main path. Take the end node of the link as the start point for

---

[4] Gephi: interactive visualization and exploration platform for all kind of networks and complex system, dynamic and hierarchical graphs (https://gephi.org/).

[5] Pajek: free software developed for analysis and visualization of large networks (http://mrvar.fdv.uni-lj.si/pajek/).

the next step. If there is a tie, take all the tied links into consideration.
2. Find the link with the largest traversal count emanating from the current point(s). Take the node(s) of the link(s) as the start point(s) for the next step. If there are ties, take all the tied links into consideration.
3. Continue step 2 until all the paths hit a sink.

4. CASE STUDY – FUEL CELL TECHNO LOGY

As mentioned beforehand the methodology described will be applied in a specific case study, which is the fuel cell technology. The reason behind choosing this topic as an example is due to its history on constant evolution through the years. There have been a lot of investments in terms of time, money and research to improve and provide better approaches in the fuel cell. The basic concept for fuel cell is a device that generates electricity by a chemical reaction, and the advantage is that it generates little pollution (much of the hydrogen and oxygen used in generating electricity ultimately combine to form a harmless by-product, namely water).

The constant evolution of the technology takes representation in the following factors such as efficiency of the cells, sizes, components, and electrodes designs. The five main fuel cell technologies known are as mentioned in section 3.1 - solid oxide, molten carbonate, phosphoric acid, proton exchange membrane and direct methanol. Each type of fuel cell has advantages and drawbacks compared to the others, and none is cheap and efficient enough to widely replace traditional ways of generating power, such as coal-fired, hydroelectric or even nuclear power plants.

By inputting the keywords through google patents website and using WebHarvy as a tool to extract the patent data, a collective database to be used in the research is obtained. The raw data obtained at first hand is a total dataset of 10,330 inputs, and these contain the founding of each one of the keywords used. After having this first result it is necessary to preprocess the data to eliminate duplicates, redundancy, and data that is missing vital information (for example patent number). This leaves the dataset with a total amount of 4,508 inputs which will be used for the following analysis. The detail of the dataset is shown in Table 1

To obtain the main path analysis, it is necessary to obtain the traversal counts that can be used to weigh each citation relation in the network of patents of fuel cell technologies. As described in section 3.3, it will be essential to calculate the search path count. To better handle the magnitudes of the values in each relation, the weights are normalized (values only between 0 and 1). That means each weight will be the number of paths divided by the total number of paths among sources and sinks. This operation is done via Pajek software, and the results are summarized in 錯誤! 找不到參照來源。2.

Table 1: Number of effective patents for different keywords in fuel cell technology

Technology	Data Points
Direct Methanol Cell	863
Molten Carbonate Cell	945
Phosphoric Acid Cell	691
Proton Exchange Membrane	852
Solid Oxide Cell	1157

From Table 2, the total of relations is 5,943. The values of SPC are clustered into 10 groups, from which 99.16% of the arcs have values of 0.0226 or less. As stated before, the higher the value is, the more important the link is. There are two links (which are included in the highest cluster) that are very important in the development of the fuel cell technologies over the years and it only represents 0.0337% of all the links. These crucial vertices are US5356731 A (Molten carbonate fuel cell with sintered LiCoO2 electrode 1994), US4710436 A (Molten carbonate fuel cell and method of manufacturing electrolyte plate thereof 1987) and US5858567 A (Fuel cells employing integrated fluid management platelet technology 1999).

After obtaining SPC of the network, it is time to apply the main path analysis. It takes into account the highest value of traversal count from the source node to the final sink. This is done through Pajek software and the final flow-over-time network is shown in Figure 2 with a final count of 23 vertices (patents) and 26 relations. It is important to understand the "flow over time" required to order the network based on the years of publication of the patents. The final network is shown in Figure 3.

Table 2: Normalized traversal weights using search path counts in the fuel cell technology

SPC (normalized)	Frequency	Freq %	CumFreq	CumFreq%
(0.0000, 0.0226)	5,893	99.1587	5,893	99.1587
(0.0226, 0.0452)	30	0.5048	5,923	99.6635
(0.0452, 0.0679)	13	0.2187	5,936	99.8822
(0.0679, 0.0905)	2	0.0337	5,938	99.9159
(0.0905, 0.1131)	0	0.0000	5,938	99.9159
(0.1131, 0.1358)	3	0.0505	5,941	99.9663
(0.1358, 0.1584)	0	0.0000	5,941	99.9663
(0.1584, 0.1810)	0	0.0000	5,941	99.9663
(0.1810, 0.2036)	2	0.0337	5,943	100.0000
Total	**5,943**	**100**		

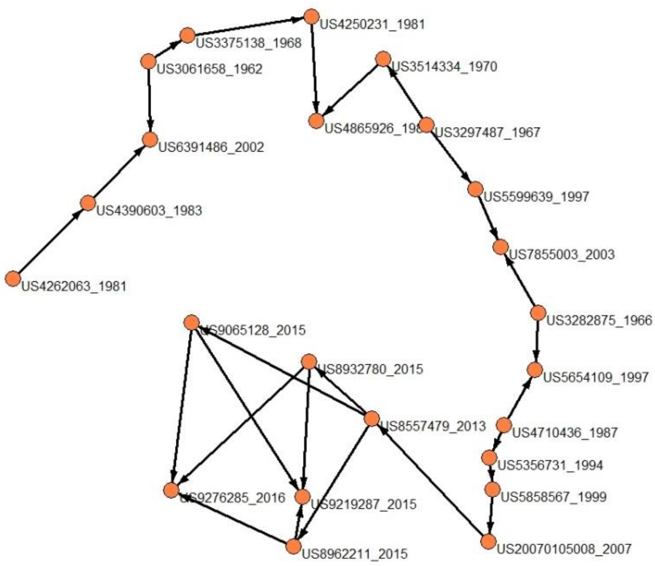

Figure 2: Flow-over-time main path analysis network of fuel cell technology

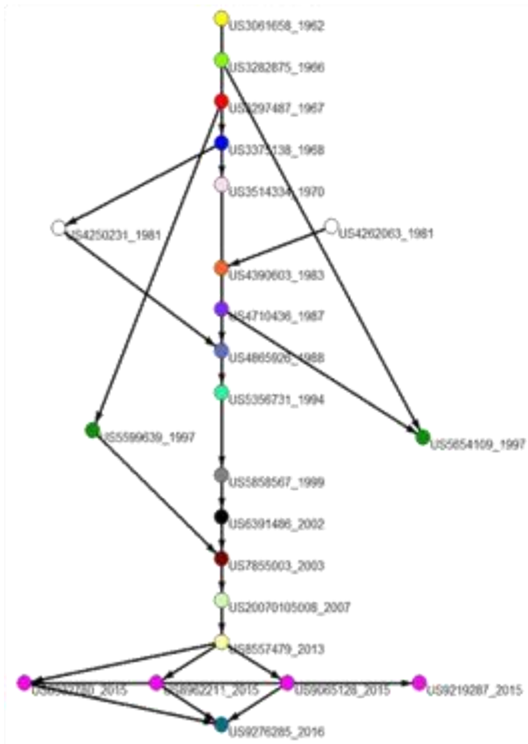

Figure 3: Main path analysis network of fuel cell technology by years

Figure 3 shows the main patents that helped in the development of the fuel cell technology, providing a satellite view of the citation relations. There is a series of observations that can be made based on the main path analysis results. First, this is the backbone of the development of the technology, from which the beginning was done by 2 sources, the first source patent titled "Fuel cell construction" of 1962 (US3061658) and the second one being "Acid electrolyte fuel cell method having improved carbon corrosion protection" in 1981 (US4250231). These two are the most important origins to incite the developments of new applications. These main patents end in 3 sinks. Firstly, one abruptly ends in 1997 with patent US5654109 - "Composite fuel cell membranes", the following two are much more recent, US9219287 - "Fuel cell" in 2015 and the patent US9276285 - "Shaped fuel source and fuel cell" from the present year 2016.

It is important to mention that there are not many branches before year 2013. The path is almost a straight line, only slightly changes in the years 1981 and 1997. In addition, these always converge back to the straight line of application back to 1962 in which supports even further the importance of this spine in the evolution of fuel cell technology field. Another point to take into account is that, the network has been built upon patents that are still valid until this year, this limitation was considered beforehand.

Another interesting point is that from the 23 nodes, the distribution to the 5 main technologies are: 7 patents belongs to the Proton Exchange Membrane, 4 to the Phosphoric Acid, 3 patents to Molten Carbonate, 2 patents of Direct Methanol and 2 to Solid Oxide. The remaining 5 patents are the most recent ones (between 2013 and 2016) and do not belong to any of the 5 main technologies. This could indicate a rise of a new technology in the fuel cell. Besides that, the last 6 patents in the main path network are owned by the company "Honeywell International Inc.", and done by the same inventor Steven J. Eickhoff. This indicates that in recent years this company has invested in the fuel cell research and treated as a main participating field. On the contrary the company who helped impulse more the field were "Du Pont" and "Hitachi Ltd.", which together owned 6 patents in the early stages of fuel cell development around 1964 to 1982.

5. CONCLUSIONS

This study proposed a three-phase procedure to analyze the trend of patents development, and used fuel cell technology as the keyword for the case study. By collecting and processing patents from the USPTO database, the citations relations were extracted and a network is built in which the interaction between different inventions over the last 80 years in the fuel cell field become visible. Starting with a network of over 42,568 patents and more than 60,000 relations, by applying the main path analysis, it was reduced to a much simpler network with only 23 patents and 5,943 citations. This "spine" or "backbone" helps understand the route in which the field was developed over the years, who were involved and what companies participated in it. Among the five key clusters in fuel cell technologies, Proton Exchange Membrane helped the most in boosting the field with a 30.4% of participation in the main path route, and the runner up goes to the Phosphoric Acid with 17%. Also, 21.7% of the patents in the route are not classified in any of the 5 main technologies and are approved recently, which could mean that there is a new type of application in the fuel cell technology field starting within the last 5 years.

In the present methodology, it has only taken into account "Utility Patents" that are still valid up to these days. By including patents that have already expired, it's possible to extend the view of interactions between citations, obtaining an even better insight on the flow of knowledge. In addition, besides the global Main Path Analysis used in this methodology, another approach is to analyze patents by applying the key-route and make a comparison with this study.

REFERENCES

Altuntas, S., & Dereli, T. (2015). A novel approach based on DEMATEL method and patent citation analysis for prioritizing a portfolio of investment projects. *Expert Systems with Applications,* **42**(3), 1003-1012.

Breitzman, A. F., & Mogee, M. E. (2002). The many applications of patent analysis. *Journal of Information Science,* **28**(3), 187-205.

Breitzman, A., & Thomas, P. (2015). Inventor team size as a predictor of the future citation impact of patents. *Scientometrics,* **103**(2), 631-647.

Cho, T.-S., & Shih, H.-Y. (2011). Patent citation network analysis of core and emerging technologies in Taiwan: 1997–2008. *Scientometrics,* **89**(3), 795-811.

Clough, J. R., Gollings, J., Loach, T. V., & Evans, T. S. (2015). Transitive reduction of citation networks. *Journal of Complex Networks,* **3**(2), 189-203.

Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The Use of Citation Data in Writing the History of Science*. Institute for Scientific Information Inc.

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.

Hummon, N. P., & Dereian, 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. Y., & Tseng, Y. C. (2014). Technological change in lithium iron phosphate battery: the key-route main path analysis. *Scientometrics,* **100**(1), 97-120.

Karki, M. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information,* **19**(4), 269-272.

Lee, C., Cho, Y., Seol, H., & Park, Y. (2012). A stochastic patent citation analysis approach to assessing future technological impacts. *Technological Forecasting and Social Change,* **79**(1), 16-29.

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.

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

Reitzig, M. (2004). The private values of 'thickets' and 'fences': Towards an updated picture of the use of patents across industries. *Economics of Innovation and New Technology,* **13**(5), 457-476.

van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics,* **8**(4), 802-823.

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.

von Wartburg, I., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy,* **34**(10), 1591-1607.