STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Patent Main Path Analysis with Arc Weights Adjusted by Classification Similarity[1]

Chung-Huei Kuan[*], Cheng-Wei Chiu[**], John S. Liu[***], Mu-Hsuan Huang[****] and Dar-Zen Chen[*****]

[*] *maxkuan@mail.ntust.edu.tw*
Graduate Institute of Patent, National Taiwan University of Science and Technology, No. 43 Sec. 4 Keelung Rd., Taipei, 10607 (Taiwan, R.O.C.)
Center for Research in Econometric Theory and Applications, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, 10617 (Taiwan, R.O.C)

[**] *yamapfans1991@hotmail.com*
Graduate Institute of Patent, National Taiwan University of Science and Technology, No. 43 Sec. 4 Keelung Rd., Taipei, 10607 (Taiwan, R.O.C.)

[***] *johnliu@mail.ntust.edu.tw*
Graduate Institute of Technology Management, National Taiwan University of Science and Technology, No. 43 Sec. 4 Keelung Rd., Taipei, 10607 (Taiwan, R.O.C.)

[****] *mhhuang@ntu.edu.tw*
Department of Library and Information Science, Center for Research in Econometric Theory and Applications, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, 10617 (Taiwan, R.O.C)

[*****] *dzchen@ntu.edu.tw*
Department of Mechanical Engineering and Institute of Industrial Engineering, Center for Research in Econometric Theory and Applications, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, 10617 (Taiwan, R.O.C)

1. Introduction

Main path analysis (MPA) was proposed to determine a representative trajectory or *main path* of development in a scientific field by identifying a chain of nodes and arcs in a citation network from the field's research articles (Hummon & Doreian, 1989). Using MPA, a large and cluttered citation network may be reduced to and epitomized by a handful of articles and citations in the main path.

Various applications arise out of MPA to observe knowledge dissemination and technology evolution, such as, detecting technological changes and knowledge transformation (Lucio-Arias & Leydesdorff, 2008; Martinelli, 2012; Mina, Ramlogan, Tampubolon, et al., 2007), reviewing literature (Bhupatiraju, Nomaler, Triulzi, et al., 2012; Calero-Medina & Noyons, 2008; Colicchia & Strozzi, 2012; Harris, Beatty, Lecy, et al., 2011; Liu, Lu, Lu, et al., 2013; Lu, Hsieh, & Liu, 2016), and mapping technological development (Fontana, Nuvolari, &

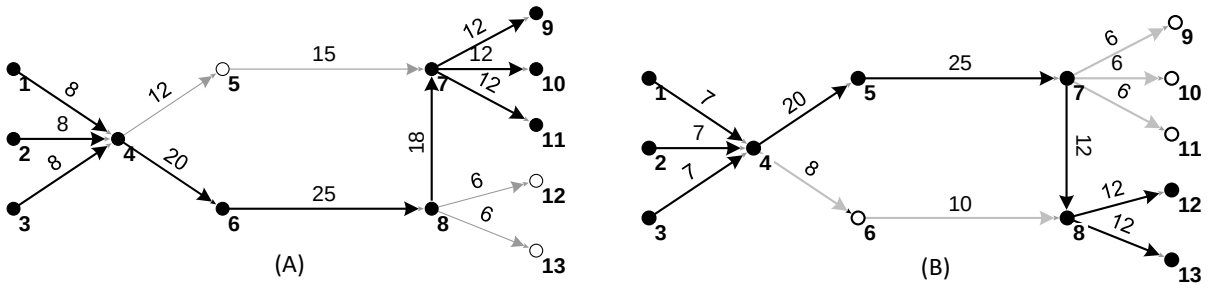
[1] This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 107L900204) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by the Ministry of Science and Technology (MOST), Taiwan, under Grant No. MOST 107-3017-F-002-004-.

Verspagen, 2009; Park & Magee, 2017; Verspagen, 2007). MPA is well accepted as a viable approach that the acclaimed social network analysis application Pajek (Batagelj & Mrvar, 1998; De Nooy, Mrvar, & Batagelj, 2011) has built-in MPA capabilities.

The conduction of MPA generally involves three steps. First, a directed citation network is constructed where nodes denote articles/patents with directional arcs originated from the cited to the citing. Then, each arc is assigned a weight based on its traversal count across the network. Finally, a series of connected nodes and arcs is determined as the main path of the citation network.

Figure 1 provides two fictitious citation networks (A) and (B), where the nodes are numbered from 1 to 13 and the arc weights, shown besides the arcs, are assigned using the algorithm *search path link count* (SPLC) (Hummon & Doreian, 1989). The two networks differ only in the direction of the arc between nodes 7 and 8. According to SPLC, for example, the weight of the arc 5→7 of the network (A) is 15, as the arc has five preceding nodes (1 to 5) and each preceding nodes will traverse the arc three times to reach the sink nodes (9 to 11). As another example, the weight of the arc 8→12 of the network (B) is 12 because four of its preceding nodes (1 to 4) traverse the arc twice (one following the arc 6→8 and the other following the arc 5→7), and the other four preceding nodes (5 to 8) traverse the arc once, to reach the sink node 12. The main paths for networks (A) and (B) of Figure 1 determined using the *global search* method (Liu & Lu, 2012) are those dark arcs connecting the black nodes. The global search method selects one of the paths from source to sink nodes having the greatest combined weight. For example, the main path in the network (B) involves the source nodes 1 to 3, the intermediate nodes 4, 5, 7, 8, and the sink nodes 12, 13, and the combined weight along the path is 76 (=7+20+25+12+12).

Figure 1: Two fictitious citation networks.



There are variations of MPA, depending on the algorithms used to assign arc weights, and methods employed to select the main path. For weight assignment algorithms, other than SPLC, there are also *search path count* (SPC) (Batagelj & Mrvar, 1998), and *search path node pair* (SPNP) (Hummon & Doreian, 1989) etc. As to the selection of main path, the *local search* method begins from the source nodes, and works forward iteratively by choosing the arc(s) from these nodes with the greatest weight(s) until a sink node is reached (Hummon & Doreian, 1989). The local search method can also start from sink nodes and work backward until a source node is reached. The *key-route* method (Liu and Lu, 2012) determines one or more main paths by first locating the arc(s) having the greatest weight and tracing both backward and forward until source and sink nodes are reached.

Prior researches have reported that the aforementioned weight assignment algorithms all produce comparable main paths (Batagelj, 2003; Verspagen, 2007). This is because that an arc

would have a greater or less weight by these algorithms is solely dependent on its structural connectivity within the network (Hummon & Doreian, 1989). An arc would have a greater weight because it may be reached from more preceding nodes and/or it may lead to more succeeding nodes, not because the corresponding citation reflects a more significant flow of knowledge or a greater relatedness between its cited and citing ones.

These algorithms treat all arcs equally but, in real life, not all citations are created equal. Taking patent citations as an example, there is continuous debate about whether applicant-submitted citations and examiner-identified citations should be treated equally (Hegde and Sampat, 2009; Cotropia, Lemley, and Sampat, 2013; Park, Jeong, and Yoon, 2017).

Based on the above review, this study intends to contribute to the discussion of MPA by treating citations not equally, but according to the degree of relatedness between the cited and the citing, and empirical patent data are utilized to observe and compare the main paths resulted from the traditional, all-citation-equal approach and the proposed approach of this study.

2. Methodology

This study proposes to modify the traditional MPA that, when assigning arc weights, each traversal of an arc is counted not by one, but by a value reflecting a degree of relatedness between the cited and the citing of the corresponding citation. In this way, this study expects that the lineage of knowledge dissemination or technological development may be more clearly captured.

There are three major categories of approaches to measure patent relatedness. The citation-based approaches involve mechanisms such as bibliographic coupling and co-citations. There are also text-based approaches analysing patent specification texts through, for example, co-word analysis. This study adopts the third type of approaches that involve the classification symbols respectively assigned to the cited and citing patents.

The advantage of using classification symbols is that every patent includes one or more classification symbols readily available from its bibliometric data. These symbols are assigned during the patent's application process according to the patent's technical content and a standard hierarchical scheme such as International Patent Classification (IPC), Cooperative Patent Classification (CPC). Patents' classification symbols are a valuable source of information as they are determined by professional and experienced personnel, and are deemed representative of the patent's technical content.

Taking two U.S. utility patents, 3989811 and 8221527, as examples, the former is assigned with a set of CPC symbols {B01D 53/1456, B01D 53/526, C10L 3/104, C10L 3/103, C01B 17/0408}, and the latter with {B01D 53/0407, B01D 53/22, B01D 53/62, C07C 1/12, B01D 53/1493, C07C 1/12, C07C 9/04, Y02P 20/152, B01D 2253/10, B01D 2253/202, B01D 2253/206, B01D 2253/302, B01D 2253/306, B01D 2253/3425, B01D 2257/504, B01D 2258/06, B01D 2259/4508, Y02C 10/04, Y02C 10/06, Y02C 10/08, Y02C 10/10, Y02C 20/20}. To see how related or similar these two patents are, the two sets of symbols are compared.

This study chooses to use the Jaccard similarity coefficient (Jaccard, 1901) to compare the two sets of symbols using their prefixes. There are totally about 260,000 different CPC symbols and using prefixes means that the symbols are compared at a higher or more abstract level in the taxonomy of technologies. For the two sample patents, the assignment frequencies of the first

four digits (i.e., the so-called sub-class or third-level symbols) of their symbols are listed in the first two rows of Table 1.

As can be seen from Table 1, symbol prefixes may occur multiple times in the same patent and, the more frequently a prefix occurs, the patent's technological content is considered to be more focused in the corresponding technological concept. The traditional Jaccard coefficient, however, ignores this phenomenon and counts each prefix just once. To obviate such a shortcoming, this study employs the generalized version of Jaccard coefficient which, for two vectors $x=\{x_1, x_2, \ldots, x_n\}$ and $y=\{y_1, y_2, \ldots y_n\}$, their Jaccard coefficient $J(x, y)$ is calculated as Eq. (1).

$$J(x,y) = \frac{\sum_i \min(x_i,y_i)}{\sum_i \max(x_i,y_i)}, \quad 1 \le i \le n. \tag{1}$$

Therefore, using the last two rows of Table 1, the two sample patents have a measured similarity 2/25. In contrast, for ordinary Jaccard coefficient using binary vectors, the similarity would be 1/6 as there are total 6 different symbols and one is commonly shared between the two sets.

Table 1. Numbers of CPC sub-class prefixes assigned to two sample U.S. patents.

	B01D	**C10L**	**C10B**	**C07C**	**Y02C**	**Y02P**
3989811	2	2	1	0	0	0
8221527	13	0	0	3	5	1
$\min(x_i, y_i)$	2	0	0	0	0	0
$\max(x_i, y_i)$	13	2	1	3	5	1

3. Data and Analysis

This study selects the field of carbon dioxide capture and storage (CCS) for empirical comparison of the main paths derived from the traditional and proposed approaches. There is a readily available CCS patent search strategy published by World Intellectual Property Organization (WIPO) in an alternative energy patent landscape report (WIPO, 2009). Then, based on the WIPO strategy, United States Patent and Trademark Office (USPTO) database were searched for utility patents issued between 1976/1/1 and 2015/12/31 having at least 'CO2,' 'carbon dioxide,' or 'CO.sub.2' in the Title, at least one additional keyword[2] in either Title or Abstract, and at least one particular classification symbol prefix[3]. The search result was then reviewed to remove those obviously not related to CCS and the patents that do not cite or are not cited by any other patents were also filtered. A final set of 675 patents were left.

A citation network is constructed using these patents and the network is then fed into Pajek. For the proposed approach, the conduction of MPA using Pajek is identical to the traditional approach, except that each arc is assigned an initial weight equal to the generalized Jaccard coefficient between the corresponding cited and citing patents' classification symbols. The final weight of each arc is automatically obtained by Pajek and is equal to the arc's traversal count times its generalized Jaccard coefficient.

[2] These keywords include `captur*, storage*, recover*, deliver*, regenerat*, sorb*, adsor*, absorb*, solv*, membrane*, where '*' is the wildcard character.
[3] These symbol prefixes are B63B 35/*, B01D 53/*, B01D 11/*, B01J 20/*, C01B 3/*, C01B 31/20, C01B 31/22, C07C 7/*, C02F 1/*, F01N 3/*, F25J 3/*, where '*' is the wildcard character.

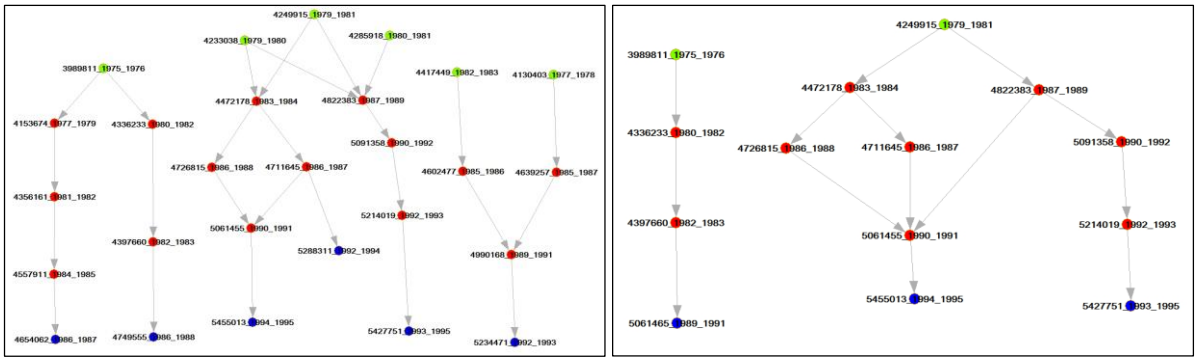
In obtaining the initial weights of the arcs, a number of analysis decisions are made. First, the CPC symbols of these patents are used, for CPC is the classification scheme currently adopted by USPTO. Then, the so-called main-group or fourth-level prefixes of the CPC symbols (i.e., the part of a symbol to the left of '/') are used for calculating generalized Jaccard coefficient. For example, a complete symbol 'B01D 53/526' has the fourth-level prefix 'B01D 53.' The fourth level provides an adequate granularity (there are about 7,000 CPC fourth-level symbols) as they are not too coarse and not too fine either. In addition, the SPLC algorithm is applied to determine the final arc weights, and the main paths are determined using the Key-route method.

To compare the traditional and modified MPAs, main paths for patents up to 1995, 2005, and 2015 are obtained, respectively, and displayed side-by-side in Figures 2 to 4 below.

A common theme between these pairs of main paths is that, by incorporating similarity measures into arc weights, the main paths from the proposed approach seems to be better at capturing technology development lineages than those of the traditional approach, which are relatively more scattered and deviant.

For the development of the CCS technologies up to 1995, as shown in Figure 2, the modified main path retains some major segments of the traditional main paths, such as 3989811→4336223→4397660, 4249915→4472178/4711645→4726815→5061455→545503, 4249915→4822383→5091358→5214019→5427715, while leaving out the others. It is speculated that, at this stage, the citation network is rather limited and the structural connectivity of the arcs is not significant enough to establish clear lineages, whereas proposed approach seems to be capable of overcoming this shortcoming.

Figure 2: Traditional (left) and modified (right) main paths up to 1995.



Then, for CCS technologies developed up to 2005, the citation network grows larger and a number of technologies, such as physical absorption and membrane separation, are manifested in Figure 3. On the other hand, not only that the major segments of the traditional main paths are retained by the modified main path, an additional and independent trajectory of the technology liquid absorption is revealed, which is absent by the traditional MPA.

Finally, as CCS technologies developed up to 2015 as illustrated in Figure 4, the traditional and modified main paths look rather similar. It is speculated that the citation network has become sizable at this stage and the traversal counts of the arcs are so great that the limited Jaccard coefficients (always between 0 and 1) would not produce noticeable difference. However, it still can be seen that some branches of the traditional main paths are pruned from the modified main paths, and the modified main paths appear to be more focused.

Figure 3: Traditional (left) and modified (right) main paths up to 2005.

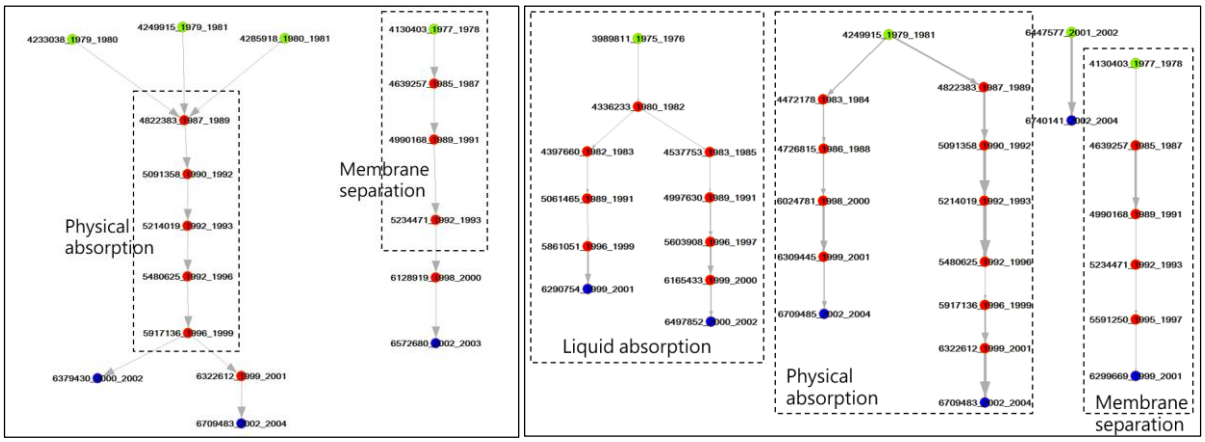
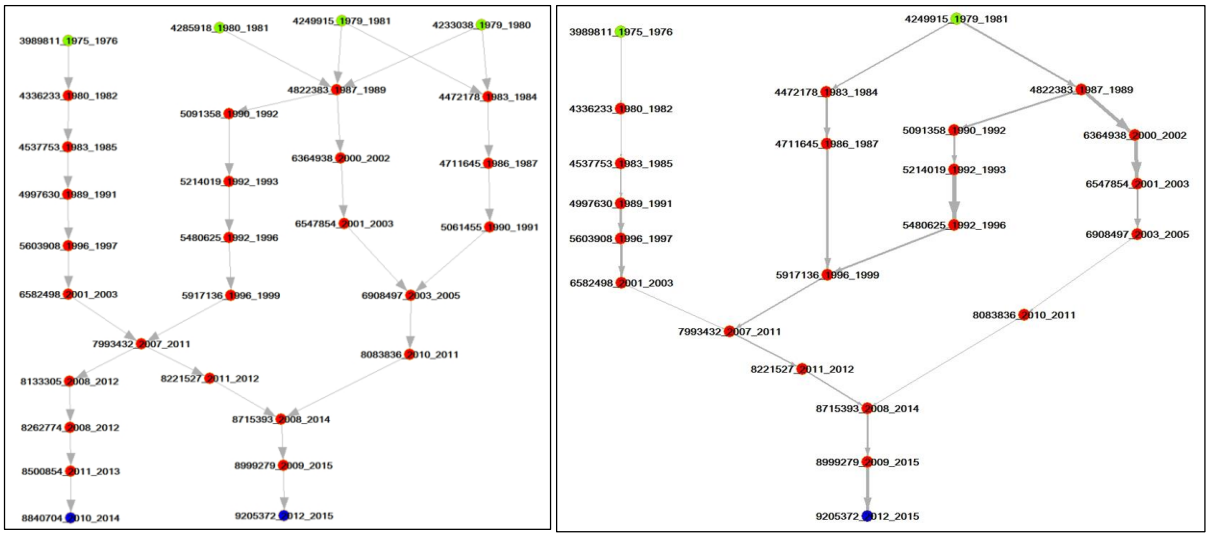


Figure 4: Traditional (left) and modified (right) main paths up to 2015.



4. Summary

Hoping that a main path may better reflect the lineage of technology development, this study modifies the traditional MPA by counting each traversal of an arc using a similarity measure based on generalized Jaccard coefficient. According to above observations, the modified MPA indeed better identifies development lineages in an earlier stage when the citation network is limited in size, as manifested in Figure 2. The modified MPA may also identify additional trajectories of separate technology development in an intermediate stage, which may be overlooked by the traditional MPA, as shown in Figure 3.

However, it seems that the modified MPA may be handicapped when applied to a sizeable citation network. The modified MPA assigns arc weights equal to an arc's traversal count times its generalized Jaccard coefficient. Then, at a later stage and for a large citation network, the traversal counts of the arcs become so great that the limited generalized Jaccard coefficients cannot provide much differentiation, as revealed in Figure 4.

Therefore, this study believes that the modified MPA proposed has its merit but, for a large citation network, how patent relatedness should be combined with the traversal count into the arc weight could be further investigated.

In addition, this study has chosen a limited set of patents in a specific field for empirical study and the generalized Jaccard coefficient for measuring patent relatedness. There are various other approaches such as cosine similarity. How these approaches may behave differently in different fields may be interesting to observe.

References

Batagelj, V., & Mrvar, A. (1998). Pajek - Program for Large Network Analysis. *Connections*, *21*(2), 47-57.

Bhupatiraju, S., Nomaler, O., 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.

Calero-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.

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.

Cotropia, C. A., Lemley, M. A., & Sampat, B. (2013). Do applicant patent citations matter? *Research Policy*, *42*(4), 844-854.

De Nooy, W., Mrvar, A., & Batagelj, V. (2011). Exploratory social network analysis with Pajek (Vol. 27). Cambridge University Press.

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.

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.

Hegde, D., & Sampat, B. (2009). Examiner citations, applicant citations, and the private value of patents. *Economics Letters*, *105*(3), 287-289.

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

Jaccard, P. (1901). Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, *37*(140), 241–272.

Jaffe, A. B., Fogarty, M. S., & Banks, B. A. (1998). Evidence from patents and patent citations on the impact of NASA and other federal labs on commercial innovation. *The Journal of Industrial Economics*, *46*(2), 183-205.

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 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). Data envelopment analysis 1978-2010: A citation-based literature survey. *OMEGA: The International Journal of Management Science*, *41*(1), 3-15.

Lu, L. Y., Hsieh, C. H., & Liu, J. S. (2016). Development trajectory and research themes of foresight. *Technological Forecasting and Social Change*, *112*, 347-356.

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

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.

Park, I., Jeong, Y., & Yoon, B. (2017). Analyzing the value of technology based on the differences of patent citations between applicants and examiners. *Scientometrics*, 1-27.

Park, H., & Magee, C. L. (2017). Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. *PloS one*, *12*(1), e0170895.

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.

WIPO (2009). Patent-based Technology Analysis Report - Alternative Energy Technology. Retrieved July 18, 2016 from: http://www.wipo.int/edocs/plrdocs/en/landscape_alternative_energy.pdf.