Assessing Patents based on Their Structural Significance in Patent Citation Network

Chung-Huei Kuan[1,3], Dar-Zen Chen[2,3]
[1] Graduate Institute of Patent, National Taiwan University of Science and Technology, Taipei, Taiwan R.O.C.
[2] Dept. of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan R.O.C.
[3] Center for Research in Econometric Theory and Applications, National Taiwan University, Taipei, Taiwan R.O.C.

Abstract—**This study proposes a patent assessment method based on a patent's structural role within a patent citation network. The proposed method includes two major steps: (1) assigning a weight to each citation of the patent citation network according to its traversal count within the network, and (2) obtaining a *pivotalness* value for each patent by summing the weights of its citations. A patent's pivotalness value is, therefore, the patent's traversal count within the network. If a citation may be deemed as a flow of knowledge or a step of technology evolution from the cited to the citing, the pivotalness value reflects a patent's significance in knowledge dissemination or technology evolution within the field. To observe this measure, this study selects for empirical analysis patents in the field of carbon dioxide capture, storage, recovery, delivery, and regeneration and collects a total of 34,083 US utility patents issued between 1976/1/1 and 2017/3/31 by the United States Patent and Trademark Office database.**

I. Introduction

To map technology landscape and to monitor technology evolution using patents, analysts often have to deal with thousands or tens of thousands of patents. Manually reviewing these patents is obviously a daunting task and analysts usually seek to identify the most representative ones, as starting points or as an epitome of the large set of patents, so that their intended analytic work may be conducted in a more efficient and orderly manner.

The field *patent bibliometrics* [16] provides a wealth of tools and indicators for evaluating patents from various perspectives. Instead of providing a detailed literature review here, interested readers may get a glimpse of the abundance of patent indicators from Ernst [5].

This study is intrigued by a branch of patent bibliometric methodologies called *main path analysis* (MPA). MPA is proposed by Hummon and Doreian [8] originally to determine the major development trajectory in a scientific field by identifying a series of citations in a citation network of the field's research articles. MPA has been applied to, for example, detecting technological changes and knowledge transformation [12, 13, 14], literature review for a field [2, 3, 4, 7, 10, 11], and finding trajectories of technological development [6, 18, 19], to name just a few. MPA is widely accepted so that the popular social network analysis software Pajek [15, 17] has built-in MPA capability.

MPA is traditionally for observing knowledge dissemination or technological development within the network. When observing knowledge dissemination, the arcs denote the passage of knowledge, and MPA may provide a characteristic course of knowledge flow. When observing technological development, the arcs reflect technology relatedness from earlier to later articles/patents, and MPA may offer a representative trajectory of technology evolution [6].

This study is intrigued by MPA not only because it may reduce a large network of patents to a handful of representative ones, but also because it provides us a hint in assessing patents from a different perspective, i.e., their roles in disseminating knowledge or in carrying the torch of technology evolution for a field.

II. Main Path Analysis

MPA generally involves three major steps: (1) a citation network is constructed whose nodes denote articles/patents with directional arcs from the cited to the citing, (2) a weight for each arc is assigned based on its traversal count, and (3) a series of connected arcs across the network is determined as a representative trajectory, referred to as the *main path*, of the citation network.

Fig. 1 illustrates two fictitious citation networks (A) and (B), where the nodes are numbered from 1 to 13 and the weights are shown beside the arcs. These weights are assigned using the algorithm *search path link count* (SPLC) [8]. For example, the weight of the arc 5→7 of the network (A) is 15, as there are five preceding nodes (1 to 5) and each will traverse the arc 5→7 once to reach one of the three sink nodes (9 to 11). Similarly, the weight of the arc 5→7 of the network (B) is 25 as, for the same five preceding nodes (1 to 5), there are now five sink nodes (9 to 13). As another example, the weight of the arc 8→12 of the network (B) is 12 because each of the four preceding nodes (1 to 4) traverses the arc 8→12 twice, one through the arc 6→8 and the other through the arc 5→7, and the other four preceding nodes (5 to 8) traverse the arc 8→12 once to reach the sink node 12.

SPLC is not the only weight assignment algorithm. There are also algorithms such as *search path count* (SPC) [15] and *search path node pair* (SPNP) [8]. The assigned weights by these algorithms are all related to the traversal counts of the arcs, except that these traversal counts are derived in different manners.

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 106-2221-E-011-105-, MOST 107-3017-F-002-004-.

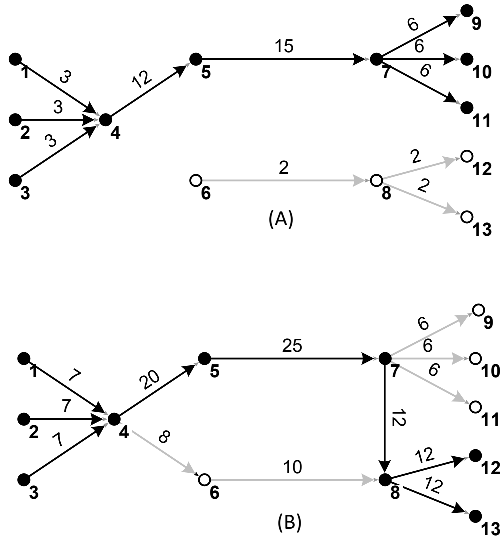

Fig. 1. Two fictitious citation networks

There are also different methods to determine the main path. The *global search* method [9] selects one of the paths from source to sink nodes having the greatest combined weight (i.e., the sum of all weights for the arcs on the path). The *local search* method starts from the source nodes, selects the arc(s) from these nodes with the greatest weight(s), and works forward for the next search until a sink node is reached [8]. The local search method can also work backward from sink nodes until a source node is reached. The *key route* method [9] determines one or more main paths by first locating the arc(s) having the greatest weight and then tracing both backward and forward until source and sink nodes are reached. The respective main paths for networks (A) and (B) of Fig. 1 using the global search method are those dark arcs connecting the black nodes. For example, the main path in the network (B) involves intermediate nodes 4, 5, 7, and 8, and the combined weight along the path is 76 (=7+20+25+12+12).

III. Patent's Structural Significance

As shown in Fig. 1, an arc would have a greater weight based upon the weight assignment algorithm such as SPC, SPLC, or SPNP, if the arc has greater structural connectivity [8], meaning that it can be reached from more preceding nodes and/or it may lead to more succeeding nodes. Taking the arcs 5→7 and 6→8 of the network (B) as an example, they have an identical number of preceding nodes, but the arc 5→7 has a greater weight because it may reach more succeeding nodes.

Then, in addition to using these weights to derive the main path, each node may be assigned a value related to the weights of its incident and/or outgoing arcs. Using the network (B) of Fig. 1 as an example, the nodes' combined incident weights (i.e., by summing the weights of all incident arcs), combined outgoing weights (i.e., by summing the weights of all outgoing arcs), and both incident and outgoing combined weights are listed in Table I. The nodes that are on the main path, except the source and sink nodes, have their weights shown against a grey background.

TABLE I. Weights for the Nodes of Fictitious Network (B)

	1	2	3	4	5	6	7	8	9	10	11	12	13
Incident	0	0	0	21	20	8	25	22	6	6	6	12	12
Outgoing	7	7	7	28	25	10	30	24	0	0	0	0	0
Both	7	7	7	49	45	18	55	46	6	6	6	12	12

This study chooses to use the combined weights of a node's outgoing arcs and the value is referred to as the node's *pivotalness* value. This value is related to the total traversal counts of the node. For example, node 7 has a combined outgoing weight, therefore a pivotalness value, 30 (=6·3+12), because it is traversed 18 times through the arcs 7→9, 7→10, 7→11, and 12 times through the arc 7→8.

Therefore, a node's pivotalness value reflects its significance in accordance with its structural position within the citation network. A node would have a greater pivotalness value if it is traversed more frequently from its proceeding nodes to succeeding nodes. If a citation may be deemed as a flow of knowledge, or a step of technology evolution, this pivotalness value then may be considered as a measure regarding how *pivotal* a patent is in terms of its role in disseminating knowledge or evolving technology within a field's citation network.

Using the combined outgoing weights has a number of benefits. First, the sink nodes (i.e., nodes 9-13) are automatically filtered as they do not have any outgoing arcs as shown in Table I. This is convenient and reasonable as the corresponding patents are not yet cited and their importance in terms of knowledge dissemination or technology evolution is yet to determine. The combined incident weights and the total combined weights do not have this benefit. Second, the pivotalness values allow analysts to differentiate the patents on the main path. Using nodes 7 and 8 of the fictitious network (B) as an example, their respective pivotalness values are 30 and 24, suggesting that the patent associated with node 7 is more pivotal than that corresponding to node 8, even though they are both located on the main path and considered as equally representative by MPA. Third, patents not on the main path are overlooked by the traditional MPA. The proposed assessment method not only may fill the blank but also may identify some snub patents that are left out by the few main path(s).

Some may suggest using mean weights instead of the combined weights as the pivotalness values. However, this is not a viable approach. Again using nodes 7 and 8 of the fictitious network (B) as an example, their mean outgoing weights are 7.5 (=30/4) and 12, respectively, and node 8, which has a smaller mean weight, would be mistaken as the more pivotal one.

IV. Data and Analysis

If the above-described assessment method seems plausible, a number of questions follows: (1) Are the patents having the greatest pivotalness values (or, *pivotal* patents) always the ones present on the main path or are there some pivotal patents actually missed by the MPA, and (2) are the pivotal patents also the ones with the largest citation counts.

Therefore, this study selects the field of carbon dioxide capture, storage, recovery, delivery, and regeneration for empirical study in answering these questions and collects a total of 34,083 US utility patents issued between 1976/1/1 and 2017/3/31 by the United States Patent and Trademark Office database. These patents contain at least one specific keyword[1] in at least one of the Title, Abstract, Specification, and Claims fields and are assigned at least one particular Cooperative Patent Classification symbol prefix[2].

A patent citation network is established using the empirical data and the software Pajek. Arc weights are assigned using the SPLC algorithm.

For the 34,083 patents, there are 29,838 patents citing or being cited by at least one other patent, thereby generating total 155,076 citations. Patents' citation counts are well known to have a skewed distribution, and this case study is no exception, as shown in Table II. On the average, each patent is cited 4.55 times and one patent, US4,440,871, has the greatest citation count 251.

It is not hard to imagine that, for the 155,076 citations, the distribution of their weights should also be a skewed one, as shown in Table III. Because the empirical case involves a huge network, the citations weigh could be very high. On the average, each citation has a mean weight 709,957 and one citation has the greatest weight up to 637,705,800!

The main path is then determined using the global search method. The global search method is used because it seems more reasonable than the local search method and simpler than the key route method, which usually produces multiple main paths and thus makes comparisons more complicated. The derived main path is shown in Fig. 2. As illustrated, a huge and cluttered network involving 34,083 nodes and 155,076 arcs is reduced down to as few as 27 nodes and 26 arcs.

To compare the most representative patents identified by the main path, the pivotalness value, and the citation count, the 27 patents on the main path are listed along with the 27 patents having the topmost pivotalness values and citation counts, respectively, in separate columns of Table IV. To facilitate comparison, the patents are sorted in ascending order of their patent numbers. Patents appearing on the main path and also having the topmost pivotalness values are shown against a grey background, whereas those on the main path and also having topmost citation counts are shown in reversed color.

TABLE II. EMPIRICAL DATA'S DISTRIBUTION OF PATENT CITATIONS

	<1	<10	<100
No. of Patents	20,201	5, 038	13
Percentage	40.73%	85.22%	99.96%

[1] The keyword search command was '(carbon or dioxide$ or co2) AND (storage$ or captur$ or recover$ or deliver$ or regenerat$),' where '$' is the wildcard character.

[2] These CPC symbol prefixes are B63B 35$, C01B 3$, C01B31/20, C01B 21/22, C02F 1$, C07C 7/10, F01N 3/10, F25J 3/02, B01J 20$, B01D 53$, and B01D 11, where '$' is the wildcard character.

TABLE III. EMPIRICAL DATA'S DISTRIBUTION OF CITATION WEIGHTS

	<100	<1,000	<10,000	<100,000	<1,000,000
No. of Citations	68,659	95,791	118,800	136,108	147,694
Percentage	44.02%	61.41%	76.17%	87.26%	94.69%

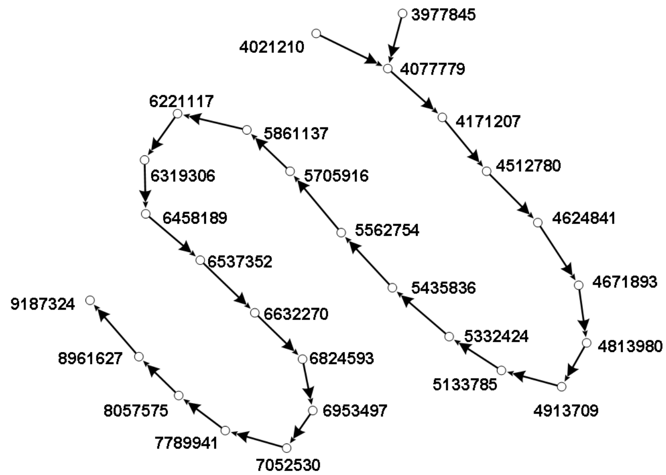

Fig. 2. Main path of the empirical data

TABLE IV. LIST OF THE MOST REPRESENTATIVE PATENTS

Global Main Path	Topmost Pivotalness	Topmost Citations
3977845	3977845	3986849
4021210	3986849	4094777
4077779	4021210	4101631
4171207	4171207	4310440
4512780	4466946	4440871
4624841	4578089	4522894
4671893	4701187	4567029
4813980	4705541	4711645
4913709	4737167	5057483
5133785	4914218	5354547
5332424	4952219	5409522
5435836	5073356	5458857
5562754	5096470	5505766
5705916	5133785	5597771
5861137	5234472	5858314
6221117	5248322	5861137
6319306	5294246	5938800
6458189	5326550	5997594
6537352	5415682	6063161
6632270	5417742	6077620
6824593	5435836	6221117
6953497	5503658	6333016
7052530	5571309	6719828
7789941	5669958	6878358
8057575	5779768	6890497
8961627	6231644	6953494
9187324	6610124	7132090

As revealed from Table IV, few patents on the main path are identified using the pivotalness values and citation counts, and there is only one patent, US3,986,849, that is overlapped between those determined by pivotalness values and by citation counts.

In other words, each of the three ways of identifying representative patents has its merit and captures a separate aspect of a patent's quality. For those on the main path, these patents are not necessarily the ones having the greatest citation counts or traversal counts; it is their structural connectivity that matters. As to citation count, it reflects a patent's "local" property, just between the patent and its immediate successors. These patents, despite being highly cited within a certain section or stage of the field's development, may lose their shine when they are evaluated in a broad context against all patents of the field. Pivotal value seems to be a measure sitting in between. It is more "global" than the citation count as it is related to a patent's traversal, directly and indirectly, by those preceding and to those succeeding it. Nevertheless, it is not as "global" as the main path because it still is limited to those, directly and indirectly, preceding and succeeding it.

One may question whether the above speculation may continue to hold if the main path is derived using a different algorithm or method. First, prior research has already reported that the SPC, SPLC, and SPNP algorithms all produce comparable main paths (cf. [1][18]). Second, to see for ourselves, another main path is determined using the local search method while the arc weights are still assigned using SPLC algorithm. The result is shown in Table V in the same manner as Table IV.

TABLE V. LIST OF THE MOST REPRESENTATIVE PATENTS

Local Main Path	Topmost Pivotalness	Topmost Citations
3977845	3977845	3986849
4021210	3986849	4094777
4077779	4021210	4101631
4171207	4171207	4310440
4512780	4466946	4440871
4705541	4578089	4474896
4790858	4701187	4522894
4913709	4705541	4567029
5133785	4737167	4711645
5332424	4914218	5057483
5435836	4952219	5354547
5562754	5073356	5409522
5705916	5096470	5458857
5861137	5133785	5505766
6221117	5234472	5597771
6319306	5248322	5858314
6458189	5294246	5861137
6537352	5326550	5938800
6632270	5415682	5997594
6824593	5417742	6063161
6953497	5435836	6077620
7052530	5503658	6221117
7789941	5562754	6333016
8057575	5571309	6719828
8961627	5669958	6878358
9102529	5779768	6890497
9187324	6231644	6953494
9556025	6610124	7132090

The newly found main path by the local search method has 28 nodes. Comparing the two main paths from Tables IV and V, it can be seen that they do not differ much (there are 24 common patents), conforming to the prior studies. Again, only a few patents on the new main path are identified using the pivotalness values and citation counts. The local search method's working from the source nodes towards the sink nodes and selecting the arc(s) with the greatest weight(s) in successively determining the main path is more compatible with how the pivotalness values are determined, and therefore two additional common patents are identified.

V. SUMMARY

This study proposes a patent assessment method which may be conveniently conducted as a side product to the traditional MPA using the readily available software Pajek. After determining a weight to each citation, a patent's pivotalness value may be immediately obtained using Pajek as its "weighted outdegree."

As described above, a patent's pivotalness value is a measure of the patent's traversal count by those preceding and to those succeeding it, directly and indirectly, in the patent citation network. Therefore, the pivotalness value captures a structural property of the patent that may be interpreted as reflecting its significance in disseminating technological knowledge within the field's patent citation network.

This structural property is unique in that it is more "global" than citation counts yet more "local" compared to the main path. Analysts, as such, may use the patents' pivotalness values to obtain a rather different set of representative patents, that may be used to supplement those identified by main path or citation count, and to gain more comprehensive insight into the field's technological evolution.

This assessment method, of course, has disadvantages. On one hand, as revealed by Tables IV and V, the pivotalness value seems to favor the older patents (i.e., those with smaller patent numbers) as they are the ones accumulating a greater number of traversals. This shortcoming may be obviated by separating the patents into distinctive sets based on their, for instance, issue dates so that patents from the same time window may be grouped together, and their pivotalness values may be compared on a comparable ground. In this way, analysts may single out representative patents from different stages of the field's technological evolution..

Another disadvantage is that, for a complex network such as the one adopted by this study, the pivotalness values could get extremely high, as shown in Table III. These insanely high values and their huge value range would lead to some difficulty in their application. For one thing, it would be hard to decide a threshold for determining what the true representative patents are. A way to overcome this shortcoming is to "normalize" pivotalness values by dividing them using the highest pivotalness value of the patents so that all values fall between 0 and 1 and, therefore, they would be easier to process.

This study may be extended in a number of ways. For example, a more rigorous study on the correlation between patents' pivotalness values and citation counts may be conducted, instead of using the simplified comparison presented in Table IV and V. Another possible extension is to incorporate the degree of similarity between the cited and the citing into the calculation on the pivotalness value. Currently, MPA treats citations uniformly. However, it is well known that not all citations are equal. There are some relevant citations that strongly link the cited and the citing, and there are also irrelevant ones. By incorporating the degree of similarity when assigning weights to the citations, the obtained pivotalness value may be more accurate.

References

[1] V. Batagelj, Efficient algorithms for citation network analysis. Ljubljana: Univ. of Ljubljana, Inst. of Mathematics, Physics and Mechanics, Dep. of Theoretical Computer Science, 2003.

[2] S. Bhupatiraju, Ö. Nomaler, G. Triulzi and B. Verspagen, "Knowledge flows – Analyzing the core literature of innovation, entrepreneurship and science and technology studies", Research Policy, vol. 41, no. 7, pp. 1205-1218, 2012.

[3] C. Calero-Medina and E. Noyons, "Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field", Journal of Informetrics, vol. 2, no. 4, pp. 272-279, 2008.

[4] C. Colicchia and F. Strozzi, "Supply chain risk management: a new methodology for a systematic literature review", Supply Chain Management: An International Journal, vol. 17, no. 4, pp. 403-418, 2012.

[5] H. Ernst, "Patent information for strategic technology management", World Patent Information, vol. 25, no. 3, pp. 233-242, 2003.

[6] R. Fontana, A. Nuvolari and B. Verspagen, "Mapping technological trajectories as patent citation networks. An application to data communication standards", Economics of Innovation and New Technology, vol. 18, no. 4, pp. 311-336, 2009.

[7] J. Harris, K. Beatty, J. Lecy, J. Cyr and R. Shapiro, "Mapping the Multidisciplinary Field of Public Health Services and Systems Research", American Journal of Preventive Medicine, vol. 41, no. 1, pp. 105-111, 2011.

[8] N. Hummon and P. Dereian, "Connectivity in a citation network: The development of DNA theory", Social Networks, vol. 11, no. 1, pp. 39-63, 1989.

[9] J. Liu and L. Lu, "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, vol. 63, no. 3, pp. 528-542, 2011.

[10] J. Liu, L. Lu, W. Lu and B. Lin, "Data envelopment analysis 1978–2010: A citation-based literature survey", Omega, vol. 41, no. 1, pp. 3-15, 2013.

[11] L. Lu, C. Hsieh and J. Liu, "Development trajectory and research themes of foresight", Technological Forecasting and Social Change, vol. 112, pp. 347-356, 2016.

[12] D. Lucio-Arias and L. Leydesdorff, "Main-path analysis and path-dependent transitions in HistCite™-based histograms", Journal of the American Society for Information Science and Technology, vol. 59, no. 12, pp. 1948-1962, 2008.

[13] A. Martinelli, "An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry", Research Policy, vol. 41, no. 2, pp. 414-429, 2012.

[14] A. Mina, R. Ramlogan, G. Tampubolon and J. Metcalfe, "Mapping evolutionary trajectories: Applications to the growth and transformation of medical knowledge", Research Policy, vol. 36, no. 5, pp. 789-806, 2007.

[15] A. Mrvar and V. Batagelj, "Analysis and visualization of large networks with program package Pajek", Complex Adaptive Systems Modeling, vol. 4, no. 1, 2016.

[16] F. Narin, “Patent Bibliometrics,” Scientometrics, vol. 30, no. 1, pp. 147–155, 1994.

[17] W. de Nooy, A. Mrvar and V. Batagelj, Exploratory Social Network Analysis with Pajek. New York: Cambridge University Press, 2012.

[18] H. Park and C. Magee, "Tracing Technological Development Trajectories: A Genetic Knowledge Persistence-Based Main Path Approach", PLOS ONE, vol. 12, no. 1, p. e0170895, 2017.

[19] B. Verspagen, "Mapping Technological Trajectories as Patent Citation Networks: A Study on the History of Fuel Cell Research", Advances in Complex Systems, vol. 10, no. 01, pp. 93-115, 2007.