Journal of Informetrics 12 (2018) 259–270

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Missing links: Timing characteristics and their implications for capturing contemporaneous technological developments

Chung-Huei Kuan[a], Mu-Hsuan Huang[b], Dar-Zen Chen[c,*]

[a] Graduate Institute of Patent, National Taiwan University of Science and Technology, No. 43, Sec. 4, Keelung Rd., Taipei, 10607, Taiwan
[b] Department of Library and Information Science, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, 10617, Taiwan
[c] Department of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei, 10617, Taiwan

ARTICLE INFO

Article history:
Received 2 August 2017
Received in revised form 15 January 2018
Accepted 15 January 2018
Available online 22 February 2018

Keywords:
Missing link
Citation
Bibliographic coupling
Patent citation network
Main path analysis

ABSTRACT

A *missing link* in this study refers to a pair of patents whose relatedness is not manifested by one citing the other but implied by their strong bibliographic coupling. By analyzing empirical data, this study discovers that the occurrence of missing links is not coincidental but arises systematically; patent pairs with missing links usually have highly overlapped application processes, whereas those with direct citations more frequently have successive or less overlapped application processes. The missing links thus may capture relatedness between patents that direct citations fail to detect. By applying main path analysis to a network containing 34,083 patents, 155,076 citations, and 9,213 missing links designed to simulate direct citations, this study further finds that the missing links—accounting for only approximately 5% of all connections—identify patents embodying contemporaneous technological developments, which may evade detection if only direct citations are considered.

1. Introduction

Investigating relatedness between the technological content of patents is central to the study of technological development. Such study may include mapping the technology landscape and monitoring technology evolution, evaluating the interaction between and impact of science and technology, observing technology spillover into other geographical areas or industry segments, and shaping patent owners' competitive goals and merger and acquisition strategies.

Direct citation (DC) (Hall, Jaffe, & Trajtenberg, 2005; Jaffe, Fogarty, & Banks, 1998; Trajtenberg, 1990) is perhaps the most widely applied tool in detecting patent relatedness. Taking its application in technology spillover as an example, a cited patent is regarded as containing the pieces of technical information that the citing patent is constructed upon; these pieces of technical information are deemed to flow from the cited to the citing patent holder's affiliated geographical areas (Figueiredo, Guimarães, & Woodward, 2015; Li, 2014; Murata, Nakajima, Okamoto, & Tamura, 2014) or industry segments (Karvonen & Kässi, 2013; Kim, Lee, & Sohn, 2016; Noailly & Shestalova, 2017).

Another popular relatedness tool is bibliographic coupling (BC) (Kessler, 1963), in which two patents are bibliographically coupled if they respectively cite at least one common document. The two patents' degree of relatedness is measured by the total number of commonly cited documents, referred to as the bibliographic coupling strength (BCS).

* Corresponding author.
E-mail addresses: maxkuan@mail.ntust.edu.tw (C.-H. Kuan), mhhuang@ntu.edu.tw (M.-H. Huang), dzchen@ntu.edu.tw (D.-Z. Chen).

https://doi.org/10.1016/j.joi.2018.01.005
1751-1577/

Both DC and BC may be utilized individually or with other relatedness tools. For example, Leydesdorff, Kushnir, and Rafols (2014) integrated DC with patent classification codes, and Nakamura, Suzuki, Sakata, and Kajikawa (2015) combined DC and co-word analysis. Kuusi and Meyer (2007) employed BC alone to cluster some related patents and identify an emerging technological paradigm in the field of carbon nanotubes. Lo (2007) also employed only BC to identify technological connections between major research organizations in the field of genetic engineering. Conversely, Von Wartburg, Teichert, and Rost (2005) combined DC and BC in a multistage analysis to reveal the technological change. Chen, Huang, Chen, and Lin (2012) used both DC and BC to construct citation networks among smart grid patents and observed the evolution of clusters of patents through a number of overlapping snapshots. Park, Jeong, Yoon, and Mortara (2015) used BC and patent text semantic analysis to locate potential research and development collaboration partners in the field of fuel cell membrane electrode assembly technology.

When employing both DC and BC to capture patent relatedness, situations may occur in which one tool indicates relatedness whereas another suggests otherwise. If one patent cites another patent, the two patents are said to form a DC pair, and if they are bibliographically coupled, they are said to constitute a BC pair. Then, if both DC and BC reflect patent relatedness, as shown in the aforementioned studies, it is interesting to notice that two patents often form a BC pair but not a DC pair.

This study was triggered by this type of BC-but-no-DC pair, particularly when a pair has high BCS, a situation which is referred to as a *missing link* (ML). Two patents form an ML pair if (1) they do not cite each other, (2) they are bibliographically coupled, and (3) they have high BCS. Therefore, an ML can only occur between two bibliographically coupled patents, but two bibliographically coupled patents do not always have an ML unless conditions (1) and (3) are satisfied. In other words, for an ML pair, their relatedness is not explicitly manifested by one directly citing the other, but strongly implied by the high BCS.

An ML pair example, two US utility patents, US8,622,222 and US8,623,202, were filed by the same company, one in January 2011 and the other in October 2012. Both were granted in January 2014 by different examiners. The two patents do not cite each other but have exceptionally high BCS of 1039 (US8,622,222 cited 1063 and US8,623,202 cited 1072 domestic and foreign patents and published applications). Both patents concern membrane bioreactor technologies and it is clear, even without examining their content, that they should be highly related. Similarly, US8,585,882 and US9,586,842, both concerning water treatment technologies, were filed by different companies, one in December 2008 and the other in December 2015. They were granted by different examiners in November 2013 and March 2017, respectively. Again, the two patents do not cite each other but have high BCS of 465 (US8,585,882 cited 472 and US9,586,842 cited 535 domestic and foreign patents and published applications); their relatedness is clearly reflected by their high BCS.

The usefulness of the ML is thus that it may be utilized to discover patent relatedness that escapes detection using DC. Chen, Huang, Hsieh, and Lin (2011) considered ML pairs as "missing citations" and used them together with DC pairs to construct comprehensive clusters of patents. Yeh, Sung, Yang, Tsai, and Chen (2013), in addition to supplementing the ML pairs in a patent citation network (PCN), further considered DC pairs with BCS less than a threshold as unreliable and removed them from the PCN.

Based on this literature review, the present study intends to contribute to the discussion of patent relatedness by utilizing empirical data to address the following issues: (1) why MLs occur and whether they are simply coincidences, (2) what useful information may be derived from the relatedness captured by MLs if their occurrence is not coincidental, and (3) how MLs may be utilized to capture this useful information.

2. Data

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. These patents contain at least one specific keyword[1] in at least one relevant field (i.e., Title, Abstract, Specification, or Claims) and at least one specific Cooperative Patent Classification symbol prefix.[2]

Among the 34,083 patents, there are 155,076 DC and 1,609,549 BC pairs. From their sheer volume, BC appears much noisier than DC. The BC pairs have a significantly skewed BCS distribution with a mean (μ) of 2.74, a standard deviation (σ) of 15.66, and a maximum of 1,123. Among the BC pairs, 72.55% (1,167,794) have the smallest BCS of 1, again suggesting that BC is relatively noisy. Among the 1,609,549 BC pairs, 75,700 are also DC pairs, and these pairs have much higher mean BCS (9.56) than the overall average. Therefore, a simultaneous DC and BC relationship indeed reflects a greater degree of relatedness between patents.

A design decision of this study is the use of a threshold to determine MLs. Swanson (1971) and Jarneving (2007) indicated that only BC pairs having BCS more than a threshold are truly related. Chen et al. (2011) used the mean BCS for pairs having a simultaneous DC and BC relationship to define a threshold, whereas Yeh et al. (2013) used the mean BCS of BC pairs without DC as a threshold. This study employs a much more conservative threshold, equal to the mean BCS plus two times the

[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 B63 B 35$, C01 B 3$, C01B31/20, C01 B 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 1
BCS distribution for the 9,213 ML pairs.

Range	35–100	101–200	201–300	301–400	401–500	501–600	601–700	701–800	801–900	901–1000	1001–1200
Pairs	5686	1987	1001	221	139	54	38	32	27	25	3
%	61.72	21.57	10.87	2.40	1.51	0.59	0.41	0.35	0.29	0.27	0.03

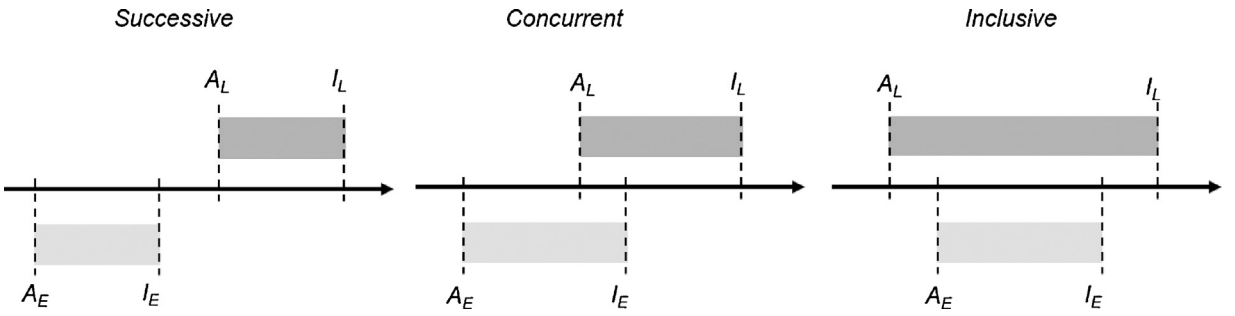

Fig. 1. Three types of timing relationships.

standard deviation ($\mu + 2\sigma = 34$). Therefore, this study considers only BC pairs that have BCS greater than 34 and do not cite each other as having an ML. Only 9,213 ML pairs are identified. The BCS distribution of these ML pairs is summarized in Table 1.

The conservative threshold was chosen so that the empirical analysis would not be overwhelmed by a large number of irrelevant BC pairs, as is evident from the presented statistics. On the other hand, a fixed value for the threshold may be questionable. As will be explained later, the fixed value nonetheless allows the observation of the different impacts of ML pairs at different stages of technological development.

Do MLs occur by coincidence? That the applicant and examiner of a first patent, after citing numerous documents that are commonly cited by a second patent, accidentally fail to cite the second patent is certainly possible. However, what intrigues the authors of this study is that there may be some type of systematic factor behind a pair of patents that causes the formation of an ML pair.

Chen et al. (2011) did notice that a majority (65.52%) of their ML pairs have a first patent's application date earlier than the second's issue date. The authors then briefly argued that it was impossible for the applicant of the first patent at its time of filing to cite the second patent because the second patent was not available for citation yet. This argument is not persuasive enough because, in one respect, whether the DC pairs in their study also behave similarly was not explored and, in another respect, whether the examiner of the first patent was handicapped in a similar manner was not discussed either. However, Chen et al. (2011) provided a hint that some insight may be obtained by observing and comparing the timing characteristics of ML and DC pairs.

3. Methodology

3.1. Timing characteristics of MLs

Each DC or ML pair has one of three types of timing relationships between their application processes, as shown in Fig. 1.[3] Application dates are denoted as A_E and A_L and issue dates as I_E and I_L, with subscripts E and L representing the pair's *earlier* and *later* issued patents, respectively. Researchers tend to regard patents filed on prior dates as earlier patents, but those issued at first are available for citation first. In the following, the earlier and later issued patents are shortened to "the earlier" and "the later," and their respective applicants/examiners are shortened to the "earlier" and "later applicants/examiners."

For the *successive* type of timing relationship, the later's application processes (the darker bar) trail behind those of the earlier, and the earlier is certainly available for citation by the later applicants and examiners. The *concurrent* type is where the application processes of the earlier and the later partially overlap. It is possible, but uncommon, that the later examiners may cite the earlier under a special condition when the earlier (1) claim identical inventions and (2) are issued before the later examiners finish searching for prior documents.[4] The *inclusive* type is a special concurrent type wherein the application processes of the earlier are fully started and completed within the time period covered by those of the later. It is very unlikely, though not impossible, that the later applicants and examiners cite the earlier in this type.

Table 2
Summary of timing relationships among DC and ML pairs.

	Successive		Concurrent		Inclusive	
Timing characteristics	$A_E < I_E < A_L < I_L$		$A_E \leq A_L \leq I_E \leq I_L$		$A_L \leq A_E < I_E \leq I_L$	
Later applicant citing earlier	Possible		Unlikely		Unlikely	
Later examiner citing earlier	Possible		Possible but rare		Unlikely	
ML pairs	2214	24.03%	4196	45.54%	2803	30.42%
DC pairs	169916	90.04%	16003	8.48%	2802	1.48%

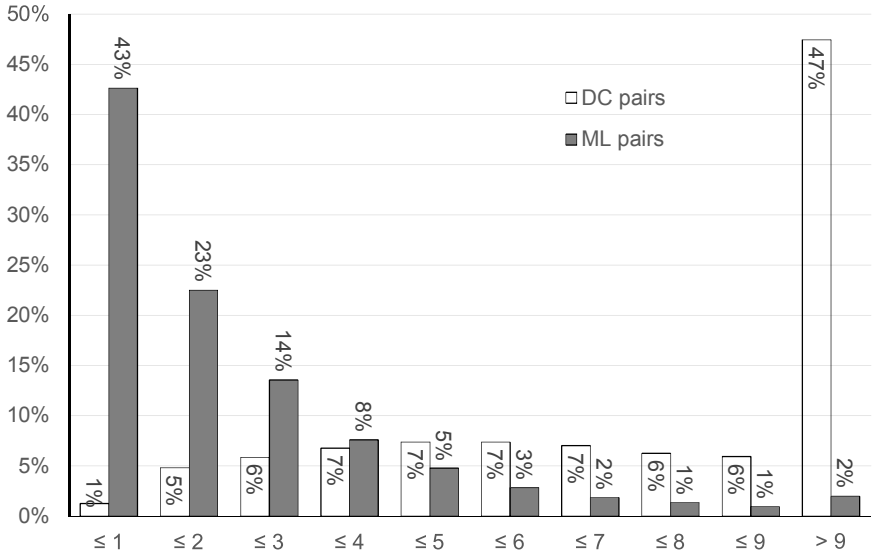

Fig. 2. Distributions of time spans, in years, of the DC and ML pairs.

The numbers and percentages of the three types of timing relationships among the DC and ML pairs in this study are summarized in Table 2, along with the three types' timing characteristics and whether the later applicants/examiners may cite the earlier.

As few as 1.48% of DC pairs belong to the inclusive type, where the later applicants and examiners are highly unlikely to cite the earlier, in contrast to a much higher 30.42% of ML pairs. Similarly, 8.48% of DC pairs belong to the concurrent type—where the later applicants are highly unlikely to cite the earlier but the later examiners do have an opportunity to cite the earlier—in contrast to an even more significant 45.54% of ML pairs. In total, three-quarters (75.97% = 45.54% + 30.42%) of ML pairs belong to the concurrent and inclusive types, whereas only approximately 10% (9.96% = 8.48% + 1.48%) of DC pairs belong to these two types.

The sharp contrast between DC and ML pairs from these observations suggests that ML pairs cannot be ascribed simply to later applicants' sometimes purposely withholding knowledge of the earlier patent, or the earlier accidentally evading the detection of the later examiners. What is revealed in Table 2 is that MLs tend to occur when patent applications undergo concurrent or inclusive application processes.

Such a tendency can be further observed from the *time spans* of the DC and ML pairs. The time span of a DC or ML pair is defined as the difference between its two patents' issue dates (i.e., $I_L - I_E$). The distributions of the time spans for all DC and ML pairs are displayed in Fig. 2, where the time spans have been calculated in days and converted into 365-day years. The DC and ML pairs' time spans have completely different distributions; 66% (43% + 23%) of ML pairs have a time span within 2 years, whereas a comparable proportion (66% = 47% + 6% + 6% + 7%) of DC pairs have time spans of more than 6 years. The time span may be alternatively defined as the difference between application dates, but the application time spans in the present study have similar distributions and therefore are omitted for brevity.

The observations prompted by Fig. 2 and Table 2 imply that MLs are not coincidences and may identify patent relatedness when DC is less likely. As such, MLs should not be omitted when conducting patent citation analysis; some crucial patent relatedness may be systematically ignored when only DCs are considered.

[3] These three types are similar to the "before," "during," and "overlap" temporal association predicates of a query language (cf. Russell & Norvig, 1995).
[4] The patent systems of many nations have such a regulation. For example, the US Patent Act specifies this scenario in §102(a)(2).

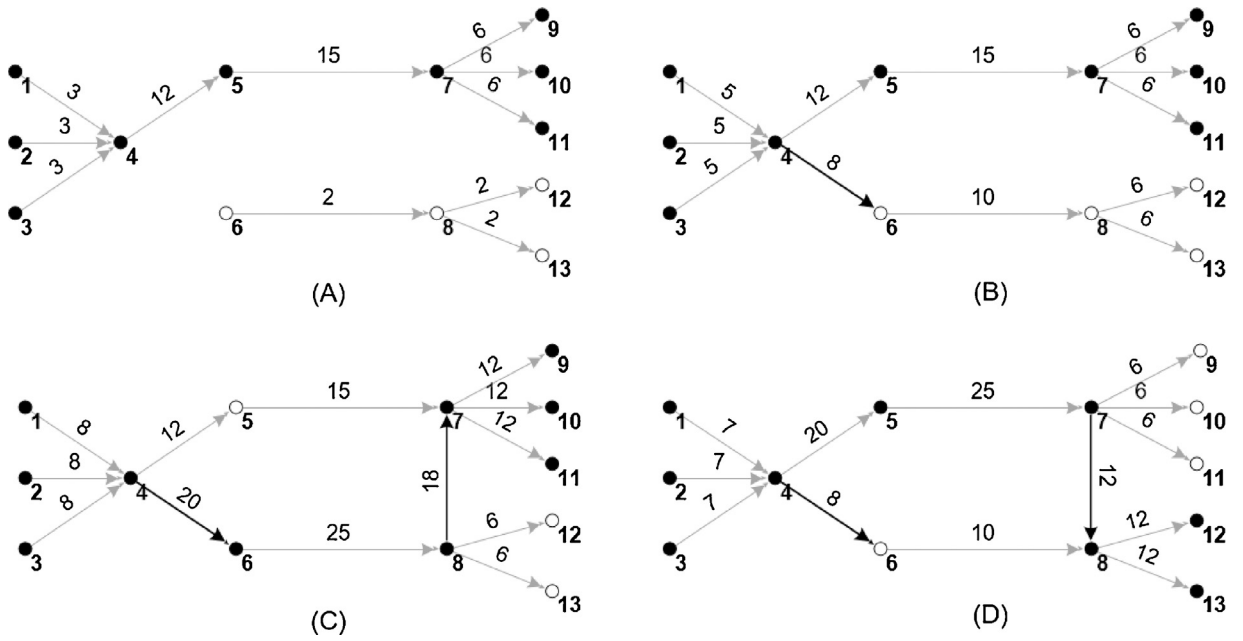

Fig. 3. Fictitious citation networks with arc weights assigned using SPLC.

3.2. MPA

This study employs main path analysis (MPA) for further analysis. The purpose of using MPA is twofold: (1) to obtain a representative set from the 34,083 patents, 155,076 DCs, and 9,213 MLs so that a more feasible and efficient examination may be conducted, and (2) to determine what useful information may be captured by MLs but missed by DCs in the context of technological development.

MPA was first proposed by Hummon and Dereian (1989) to determine the major development trajectory of a scientific field by identifying the most significant chains of DCs in a citation network of scientific articles. Since its inception, researchers have applied MPA to, for example, the detection of technological changes and knowledge transformation (Lucio-Arias & Leydesdorff, 2008; Martinelli, 2012; Mina, Ramlogan, Tampubolon, & Metcalfe, 2007), reviews of literature (Bhupatiraju, Nomaler, Triulzi, & Verspagen, 2012; Calero-Medina & Noyons, 2008; Colicchia & Strozzi, 2012; Harris, Beatty, Lecy, Cyr, & Shapiro, 2011; Liu, Lu, Lu, & Lin, 2013; Lu, Hsieh, & Liu, 2016), and mapping of trajectories of technological development (Fontana, Nuvolari, & Verspagen, 2009; Park & Magee, 2017; Verspagen, 2007). MPA is now so popular that the acclaimed social network analysis application Pajek (Batagelj & Mrvar, 1998; De Nooy, Mrvar, & Batagelj, 2011) incorporates various MPA analytic methods.

MPA generally involves three major steps. First, a citation network is constructed where nodes denote documents and directional arcs connect DC pairs from the cited to the citing. Then, a weight for each arc is determined according to the arc's traversal count. Finally, 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. 3 illustrates four fictitious citation networks (A) to (D), where the nodes are numbered from 1 to 13 and the weights are displayed beside the arcs. These weights are calculated using the algorithm *search path link count* (SPLC) (Hummon & Dereian, 1989). For example, the weight of the arc $5 \rightarrow 7$ in network (A) is 15, because SPLC counts the number of traversals of the arc $5 \rightarrow 7$ from all preceding nodes (1–5) to the sink nodes (9–11). Similarly, the weight of the arc $5 \rightarrow 7$ in network (D) is 25 because there are five preceding nodes (1–5) and each will traverse the link $5 \rightarrow 7$ once to reach one of the sink nodes (9–13). Other algorithms can be used, such as *search path count* (SPC) (Batagelj & Mrvar, 1998) and *search path node pair* (SPNP) (Hummon & Dereian, 1989). No matter the algorithm, an arc has greater weight if it has greater structural connectivity (Hummon & Dereian, 1989) or if it can be reached from more preceding nodes and/or it may lead to more succeeding nodes.

Various methods also exist for determining the main path. The *global search* method (Liu and Lu, 2012) selects the path from source to sink nodes having the greatest total weight. The *local search* method begins 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 (Hummon & Dereian, 1989). The local search method can also commence 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 locating the arc(s) having the greatest weight first and tracing both backward and forward until source and sink nodes are reached. The respective main paths for networks (A) to (D) of Fig. 3 are those connecting the black nodes.

MPA is traditionally applied to a network of DCs for observing knowledge dissemination or technological development within the network. When observing knowledge dissemination, the arcs denote the continuation of knowledge, and the main path is a representative course of knowledge flow. When observing technological development, the arcs reflect technology relatedness from earlier to later documents/patents, and the main path is a representative trajectory of technology evolution (Fontana et al., 2009).

This study extends MPA to a network involving not only DCs but also MLs so that both the explicit relatedness manifested by the DCs and the latent relatedness captured by the MLs are considered in deriving the trajectory of technology evolution. To incorporate MLs into the network so that they are applicable to MPA, each ML pair is represented by an arc and, as ML does not have a direction, the arc is arranged so that it originates from a node of the pair's lower-numbered patent to a node of the pair's higher-numbered patent.

In this method, each ML is treated as a fictitious DC. As detailed in the previous section, MLs occur mostly between patents with concurrent and inclusive application processes where the later applicants and examiners may fail to cite the earlier patent. Therefore, simulating each ML as a DC is equivalent to assuming that there would have been a DC by the later applicants and examiners if they had been able to cite the earlier patent. In this sense, the proposed arrangement of arc direction is not unreasonable because it is highly unlikely that a higher-numbered patent is cited by a lower-numbered one. In fact, all 155,076 DC pairs in this study have their lower-numbered patent cited by the higher-numbered patent.

Such a network is no longer a conventional citation network, but a heterogeneous network containing two types of arcs: those denoting DCs and those denoting MLs.

4. Analysis and result

Two networks are constructed using the empirical data: a conventional PCN with 34,083 nodes connected by 155,076 arcs, one for each DC pair (hereafter, DC arcs); and a heterogeneous network with an additional 9,213 arcs, one for each ML pair (hereafter, ML arcs). Arc weights are assigned using the SPLC algorithm, as prior research has reported that the SPC, SPLC, and SPNP algorithms all produce comparable results (cf. Batagelj, 2003; Verspagen, 2007). Then, two main paths are derived from the networks using the global search method. The path obtained from the PCN is referred to as the *original main path* (OMP), whereas that obtained from the heterogeneous network is referred to as the *heterogeneous main path* (HMP). The global search method is used because it is 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 OMP and HMP together discover 67 patents, 33 DCs, and 42 MLs. These patents and their nodes are numbered from 1 to 67 according to the ascending order of their patent numbers and, therefore, their issue dates. Subsequently, for example, node 20 and patent 20 are used interchangeably. The lower-numbered nodes therefore denote earlier issued patents with lower patent numbers. The patents are listed in Appendix A along with their application and issue dates. The OMP and HMP are displayed in Fig. 4 using nodes and arcs of different styles for easy comparison. The black, white, and gray nodes denote patents recognized by both the OMP and HMP, the OMP only, and the HMP only, respectively. The solid gray and dashed black arcs represent the DC and ML arcs, respectively.

The 67 patents, 33 DCs, and 42 MLs account for only a small portion of data from the case study. However, they act as epitomes to the original PCN and the heterogeneous network. As described in Section 3.2, MPA has an advantage for this purpose because it can filter out irrelevant nodes and arcs, and preserve those most structurally important ones in the main path. As such, the DC noise problem (Jaffe et al., 1998; Park, Jeong, and Yoon, 2017) is of little concern. Conversely, the BC noise problem (Jarneving, 2007; Swanson, 1971) is resolved first by the choice of a conservative threshold and further by the use of MPA.

The OMP identifies 27 patents and 26 DCs. Among the 27 patents, 19 patents are also identified by the HMP. The OMP is therefore denoted by the chain of 26 solid gray arcs connecting 19 black nodes and 8 white nodes. In other words, a majority of OMP patents are included in the HMP. However, for the 26 DCs, only 9 are included in the HMP.

The HMP identifies 59 patents, 16 DCs, and 42 MLs. Other than the 19 patents and 9 DCs already discovered by the OMP, the HMP identifies additional 40 patents and 7 DCs. The HMP is denoted by the chain of 16 solid gray arcs and 42 dashed black arcs connecting 19 black and 40 gray nodes. The 42 MLs have BCS ranging from 39 to 256, with a mean of 101 and standard deviation of 57. Compared with the values in Table 1, their BCS is not particularly high because, for an ML arc to be included in the HMP, it is its structural connectivity rather than its BCS that matters.

The heterogeneous network has 9,213 ML arcs and 155,076 DCs arcs. Despite the small proportion of arcs that are ML arcs (5.61%), these arcs dominate the resulting HMP (42 ML arcs vs. 16 DC arcs). The ML arcs clearly occupy key locations within the heterogeneous network so that their strong structural connectivity allows them to be selected into the HMP.

The HMP does not, however, deviate entirely from the OMP. Each time the HMP deviates from the OMP, it almost always returns to intersect the OMP; at nodes 8, 19, and 25, for example. With these nodes and their patents as demarcations, their respective issue dates define five time windows, each involving a number of patents, DCs, and MLs. These issue dates are included in Fig. 4, and relevant data for these time windows are summarized in Table 3. In the following, an ML or DC is said to *occur* within a window if the pair's time span from the earliest application date, $min(A_E, A_L)$, to the latest issue date, $max(I_E, I_L)$, falls within the window entirely (i.e., window's start date $\leq min(A_E, A_L) < max(I_E, I_L) \leq$ window's end date) or partially by extending across the window's end date (i.e., window's start date $\leq min(A_E, A_L) \leq$ window's end date $< max(I_E, I_L)$).

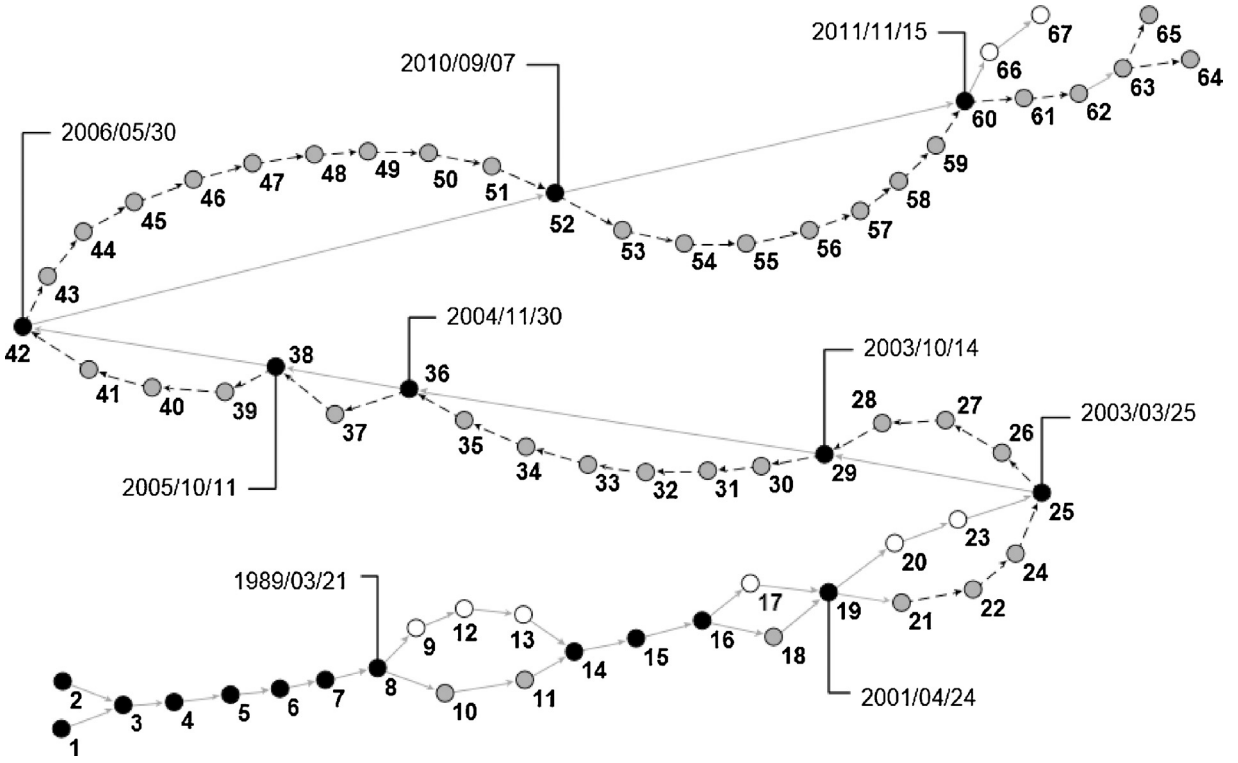

Fig. 4. Integrated display of the OMP and HMP.

Table 3
Relevant data on the five time windows.

#	Time window	HMP seg.	MLs	DCs	MLs/DCs	Avg. Doc.Cited
1	1976/01/01–1989/03/21	$1 \rightarrow \ldots \rightarrow 8$	11	5885	0.19%	8.00
2	1989/03/22–2001/04/24	$8 \rightarrow \ldots \rightarrow 19$	1513	118005	1.28%	15.69
3	2001/04/25–2003/03/25	$19 \rightarrow \ldots \rightarrow 25$	985	19472	5.06%	20.02
4	2003/03/26–2011/11/15	$25 \rightarrow \ldots \rightarrow 60$	5200	42594	12.21%	30.52
5	2011/11/16–2017/03/31	$60 \rightarrow \ldots \rightarrow 67$	1504	2766	54.37%	40.83

Window 1 should correspond to an early stage of the field's development, and 11 MLs and 5,885 DCs occur within this window. These MLs are too few to considerably affect the main path, and the HMP and OMP share a segment within this window (i.e., nodes 1–8 connected by solid gray arcs). This scenario is similar to, in a simplified manner, adding a black arc $4 \rightarrow 6$ to the network (A) of Fig. 3 and deriving the same main path in the resulting network (B). There are so few MLs because patents issued at this early stage have few documents to cite, and patent pairs are less likely to have BCS high enough to qualify as ML pairs. For example, US3,977,845 and US4,813,980, i.e., nodes 1 and 8, citing only 7 and 14 documents, respectively, can never be part of an ML pair according to the threshold, 34. The average number of cited documents for patents issued within this time window is only 8.00.

More patents were issued within window 2, so significantly more DCs (118,005) and MLs (1,513) occur within this window. Even though the MLs still amount to a small portion (1.28%), they force the HMP to follow alternate routes $8 \rightarrow 9 \rightarrow 12 \rightarrow 13 \rightarrow 14$ and $16 \rightarrow 17 \rightarrow 19$. However, the HMP segment within this window is connected by only DC arcs, meaning that the ML arcs only strengthen the existing DC arcs to form new routes in the HMP, but the ML arcs themselves are not dense enough to form part of the alternate routes.

Window 3 has denser MLs (5.06%), and the corresponding ML arcs, in addition to strengthening existing DC arcs, create a new route. This effect is similar to adding black arcs $4 \rightarrow 6$ and $8 \rightarrow 7$ to the network (A) of Fig. 3, so that a new route ($4 \rightarrow 6 \rightarrow 8 \rightarrow 7$) is obtained in the resulting network (C). The application processes of patents from both the original and new routes are illustrated in Fig. 5. Because these patents were issued within a short 2-year period, their application processes were mostly concurrent or inclusive. The application processes of patents in the original route (i.e., nodes 19, 20, 23, and 25, connected by DC arcs) overlap less than those of patents in the new route (i.e., nodes 19, 21, 22, 24, and 25, connected by ML arcs except the DC arc $19 \rightarrow 21$). This is consistent with what is suggested from the earlier section.

Window 4 has even denser MLs (12.21%). The OMP within this window includes a series of seven nodes (i.e., nodes 25, 29, 36, 38, 42, 52, and 60). These nodes are not only retained by the HMP but also positioned as junctions. Using these nodes

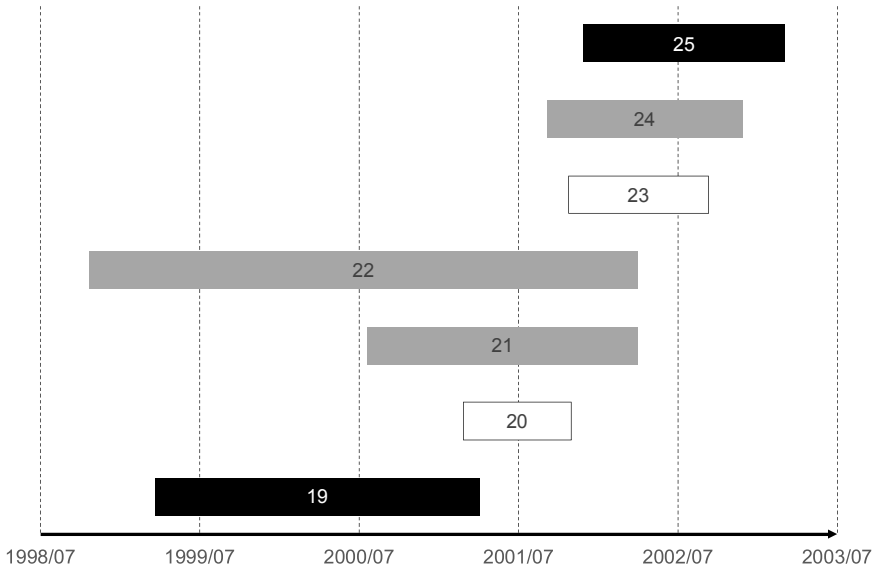

Fig. 5. Application processes of patents 19–25.

and their patents as demarcations, window 5 may be further divided into six subwindows delimited by the issue dates of these patents, also shown in Fig. 4. Except for those defined by patents 42 and 52, the subwindows have short periods of less than or approximately 1 year. Within each subwindow, the HMP always follows a separate route consisting exclusively of ML arcs.

Using the subwindow defined by patents 52 and 60 as an example, most patents on the HMP segment within this subwindow have an inclusive relationship with an earlier patent and/or a later patent, as shown in Fig. 6. For example, the application process of patent 52 was covered by that of patent 53, whose application process was covered by that of patent 54, and so on. The applicant/examiners of these patents were highly unlikely to cite the earlier patent or to be cited by the later patent. These patents therefore may be omitted by the OMP due to the absence of a DC, yet they are discovered by the HMP with the help of MLs.

Patents 53–59 should, on the one hand, be highly related to each other and also to the DC pair of patents 52 and 60 because they are all cascaded by MLs of high BCS. On the other hand, in addition to their relatedness, these patents embody technological developments contemporaneous with those of the DC pair since they were issued between the issuance of the DC pair of patents 52 and 60, and their application processes mostly overlapped. As technology evolves, there may be

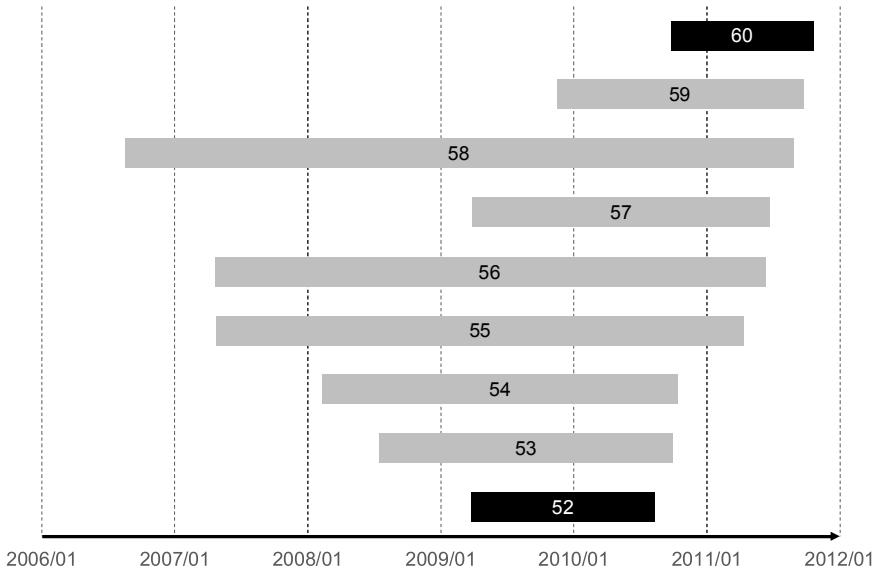

Fig. 6. Application processes of patents 52–60.

Table 4
Relevant data for patents 38–42.

#	Grant date	Applicant	Title
38	2005/10/11	IdaTech, LLC	Hydrogen purification devices, components and fuel processing systems containing the same
39	2005/11/22	The University of Chicago	Autothermal hydrodesulfurizing reforming method and catalyst
40	2006/02/27	IdaTech, LLC	Integrated fuel cell system
41	2006/02/28	IdaTech, LLC	Steam reforming method and apparatus incorporating a hydrocarbon feedstock
42	2006/05/30	IdaTech, LLC	Hydrogen purification membranes, components and fuel processing systems containing the same

times when developments of related technologies occur concurrently. If these contemporaneous developments are filed for patent protection, their application processes mostly overlap. Their relatedness then may elude detection by DCs but would still be caught by MLs. MLs therefore appear to fill some "gap" between a DC pair by chaining these highly related patents that reflect contemporaneous developments.

MLs are densest in window 5 (54.37%). The numerous MLs allow the HMP to extend independently from the OMP, and, unlike what is observed in window 4, it does not appear that the OMP and HMP will intersect in the future. A similar and simplified scenario can be observed in Fig. 3, where adding black arcs $4 \rightarrow 6$ and $7 \rightarrow 8$ to the network (A) causes the main path to develop toward a different set of sink nodes, as shown in the network (D). So many MLs are present in window 5 because patents issued at this late stage had more documents available for citation, and patent pairs are more likely to pass the BCS threshold (34) and qualify as ML pairs. For example, US8,057,575 and US8,696,772, i.e., nodes 60 and 65, cited 238 and 121 documents, respectively. The average number of cited documents for patents issued within this window is 40.83.

This study further conducts a number of experiments. Three HMPs are derived after the weights of all ML arcs are manually reduced to 0.5, 0.1, and 0.01 of their SPLC-derived original values whereas the weights of the DC arcs remain the same. The resulting HMPs are exactly the same as that shown in Fig. 4. In other words, these MLs are confirmed as having structural connectivity so great that downplaying their significance to a large degree does not affect the main path.

By observing and comparing the OMP and HMP within the five time windows, this study finds that the sporadic ML arcs of windows 1 and 2 have little influence and the resulting HMP does not differ significantly from the OMP. Then, as the density of ML arcs increases in windows 3 and 4, these ML arcs join to form new routes with stronger structural connectivity than the OMP's original routes. In other words, the HMP obtains a more comprehensive trajectory when MLs fill the gaps, i.e., latent patent relatedness, left open by DCs. Then, as observed in window 5 when the density of ML arcs continues to rise, the HMP may deviate entirely from the OMP. However, it should be kept in mind that patents issued within window 5 cite more than 40 documents on average and easily form ML pairs according to the chosen threshold. The validity of these ML pairs may be dubious.

The most crucial finding is that observed from window 4, when various endeavors are engaged to drive technology forward simultaneously. Using DCs alone may miss critical details because patents embodying these contemporaneous technological developments can be left unnoticed. MLs may then be employed to uncover these details, providing a more complete picture of how technology evolves.

To further demonstrate the utility the MLs, a segment of HMP involving patents 38–42 is adopted as an example and relevant data about these patents are summarized in Table 4.These patents are all related to the production and purification of hydrogen gas from carbon-containing materials for a fuel-cell system. Four of them are filed by the same company, share at least one common inventor, and actually are members of the same patent family. Their relatedness is without a doubt but, due to their highly overlapped application processes, OMP can only identify two of them, patents 38 and 42, that span apart enough to have a DC. The HMP not only identifies the other two related patents from the same company but also, most interestingly, includes one applied by the University of Chicago. This patent 39 discloses a method and catalyst to reform a sulfur-containing carbonaceous fuel into a hydrogen-containing gas. This highly-related and concurrently developed technology from academia would be left unnoticed if only DCs are considered but is exposed with the help of MLs.

5. Conclusion and discussion

This study addresses the questions: (1) why do MLs occur, and are they simply coincidences, (2) what useful information may be captured by MLs, and (3) how may MLs be utilized to capture this useful information.

This study discovers that ML pairs often undergo highly overlapped application processes and the applicants/examiners of such ML pairs are inherently unlikely to cite each other. DC pairs more frequently have successive or less overlapped application processes, and their applicants/examiners are not handicapped in citing each other. Therefore, MLs are fostered out of a systematic context where DCs are most likely to be absent.

This study proposes a method of utilizing MLs for the investigation of technological development by extending MPA to a heterogeneous network involving not only DCs but also MLs so that both the explicit relatedness manifested by DCs and latent relatedness captured by MLs are considered to derive a trajectory of technology evolution. In doing so, each ML is treated as a fictitious DC and represented in the PCN by an arc of artificially assigned direction. Based on such a heterogeneous network, this study finds that MLs may capture concurrent efforts in developing related technologies embodied in contemporaneous patents.

It should be emphasized that ML should not be considered as a better tool than DC. What is discovered by this study suggests that ML should be treated as a supplement to, not replacement for, DC. Both DC and ML have their merits. Similarly, both OMP and HMP have their advantages. The OMP is simpler to obtain and may provide a general view to the technology development, as some key patents are preserved in the HMP as well. The HMP, on the other hand, is more difficult to derive but may fill in additional details or may provide alternate routes, as illustrated in Fig. 4, within the general view offered by DCs. In other words, the HMP may offer details that OMP is unable to provide. However, OMP is required as a backdrop so as to see the difference.

The usefulness of ML therefore lies in two respects. First, ML may discover patent relatedness that is not readily apparent by utilizing DCs only. In this respect, utilizing MLs and DCs together should be able to construct clusters of patents (or assignees) of greater accuracy than those constructed by DCs alone. Second, ML may particularly identify patents for contemporaneous technological development that DCs may fail to detect. In this respect, utilizing DCs and MLs should allow analysts to acquire a more thorough understanding of the evolution of technology. Overall, analysts should not omit MLs when conducting various types of patent citation analysis; otherwise, some crucial patent relatedness may be systematically ignored.

When discovering contemporaneous patents, the HMP introduces a sequential order among these patents; however, this sequential order should not be taken at face value. For example, patents 52–57 all precede patent 58 along the HMP as shown in Fig. 4, yet analysts should not jump to the conclusion that patent 58 is evolved from these patents because patent 58 actually has the earliest filing date, as shown in Fig. 6.

It may be argued that a fixed BCS threshold is not appropriate. Indeed, using a fixed threshold may exclude some critical BC pairs in an early stage of technology evolution, as observed in windows 1 and 2, and may include less important ML pairs in later stages, as observed in window 5. However, this study is aimed at the investigation of the characteristics of MLs, and a fixed threshold enables efficient observations as different amounts of MLs are supplemented.

This study may be extended along several directions. First, some effort should be invested in the determination of an appropriate BCS threshold for qualifying BCs as MLs. Then, based on the improved threshold, it may be investigated whether the trajectory of technological development obtained by the proposed MPA-over-heterogeneous-network method is indeed more accurate than the conventional method. Furthermore, in addition to including MLs in a PCN, the DCs with low BCS may also be considered irrelevant and removed from the PCN, as suggested by Yeh et al. (2013). Other possible extensions may include investigating MLs that are determined by co-citation rather than BCs, comparing the patent relatedness captured by MLs against that detected by other relatedness tools such as topical similarity, and integrating MLs with other tools to see whether even more accurate patent relatedness may be obtained.

Author contributions

Chung-Huei Kuan: Conceived and designed the analysis, collected the data, contributed data or analysis tools, performed the analysis, wrote the paper.

Mu-Hsuan Huang: Conceived and designed the analysis, contributed data or analysis tools, wrote the paper.

Dar-Zen Chen: Conceived and designed the analysis, collected the data, contributed data or analysis tools, performed the analysis, wrote the paper.

Appendix A.

Node	Pat. Num.	OMP	HMP	App. Date	Issue Date
1	3977845	Y	Y	1975/06/20	1976/08/31
2	4021210	Y	Y	1975/12/19	1977/05/03
3	4077779	Y	Y	1976/10/15	1978/03/07
4	4171207	Y	Y	1978/08/21	1979/10/16
5	4512780	Y	Y	1983/11/08	1985/04/23
6	4624841	Y	Y	1985/01/22	1986/11/25
7	4671893	Y	Y	1986/02/24	1987/06/09
8	4813980	Y	Y	1987/10/16	1989/03/21
9	4913709	Y		1989/02/17	1990/04/03
10	4915711		Y	1989/05/18	1990/04/10
11	5073356		Y	1990/09/20	1991/12/17
12	5133785	Y		1991/02/26	1992/07/28
13	5332424	Y		1993/07/28	1994/07/26
14	5435836	Y	Y	1993/12/23	1995/07/25
15	5562754	Y	Y	1995/06/07	1996/10/08
16	5705916	Y	Y	1996/01/17	1998/01/06
17	5861137	Y		1996/10/30	1999/01/19
18	5997594		Y	1997/10/15	1999/12/07
19	6221117	Y	Y	1999/04/13	2001/04/24
20	6319306	Y		2001/03/19	2001/11/20
21	6375906		Y	2000/08/10	2002/04/23

22	6376113		Y	1998/11/12	2002/04/23
23	6458189	Y		2001/11/14	2002/10/01
24	6494937		Y	2001/09/27	2002/12/17
25	6537352	Y	Y	2001/12/19	2003/03/25
26	6562111		Y	2002/02/04	2003/05/13
27	6569227		Y	2002/02/28	2003/05/27
28	6596057		Y	2002/07/15	2003/07/22
29	6632270	Y	Y	2003/02/20	2003/10/14
30	6641625		Y	2000/05/02	2003/11/04
31	6719831		Y	2003/05/05	2004/04/13
32	6719832		Y	2003/05/15	2004/04/13
33	6723156		Y	2003/05/05	2004/04/20
34	6767389		Y	2003/07/21	2004/07/27
35	6783741		Y	2001/04/20	2004/08/31
36	6824593	Y	Y	2003/12/05	2004/11/30
37	6869707		Y	2002/04/19	2005/03/22
38	6953497	Y	Y	2004/03/16	2005/10/11
39	6967063		Y	2001/05/18	2005/11/22
40	6994927		Y	2005/03/18	2006/02/07
41	7005113		Y	2002/04/19	2006/02/28
42	7052530	Y	Y	2004/11/15	2006/05/30
43	7135048		Y	2000/08/10	2006/11/14
44	7195663		Y	2006/05/25	2007/03/27
45	7368194		Y	2005/05/06	2008/05/06
46	7410531		Y	2007/03/20	2008/08/12
47	7470293		Y	2005/03/31	2008/12/30
48	7601302		Y	2005/09/16	2009/10/13
49	7632322		Y	2005/09/13	2009/12/15
50	7682718		Y	2008/05/05	2010/03/23
51	7736596		Y	2009/10/06	2010/06/15
52	7789941	Y	Y	2009/04/20	2010/09/07
53	7819955		Y	2008/08/11	2010/10/26
54	7828864		Y	2008/03/07	2010/11/09
55	7939051		Y	2007/05/21	2011/05/10
56	7972420		Y	2007/05/18	2011/07/05
57	7981172		Y	2009/04/23	2011/07/19
58	8021446		Y	2006/09/13	2011/09/20
59	8038748		Y	2009/12/11	2011/10/18
60	8057575	Y	Y	2010/10/21	2011/11/15
61	8157900		Y	2011/06/09	2012/04/17
62	8257466		Y	2011/11/14	2012/09/04
63	8636828		Y	2012/08/29	2014/01/28
64	8691463		Y	2011/06/24	2014/04/08
65	8696772		Y	2011/06/22	2014/04/15
66	8961627	Y		2011/07/07	2015/02/24
67	9187324	Y		2013/03/14	2015/11/17

References

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

Batagelj, V. (2003) Efficient algorithms for citation network analysis, University of Ljubljana, Institute of Mathematics, Physics and Mechanics, Preprint series, 41, 897.

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.

Chen, D. Z., Huang, M. H., Hsieh, H. C., & Lin, C. P. (2011). Identifying missing relevant patent citation links by using bibliographic coupling in LED illuminating technology. *Journal of Informetrics*, *5*(3), 400–412.

Chen, S. H., Huang, M. H., Chen, D. Z., & Lin, S. Z. (2012). Detecting the temporal gaps of technology fronts: A case study of smart grid field. *Technological Forecasting and Social Change*, *79*(9), 1705–1719.

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.

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

Figueiredo, O., Guimarães, P., & Woodward, D. (2015). Industry localization, distance decay, and knowledge spillovers: Following the patent paper trail. *Journal of Urban Economics*, *89*, 21–31.

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.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics*, 16–38.

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.

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

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.

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informetrics*, *1*(4), 287–307.

Karvonen, M., & Kässi, T. (2013). Patent citations as a tool for analysing the early stages of convergence. *Technological Forecasting and Social Change*, *80*(6), 1094–1107.
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *Journal of the Association for Information Science and Technology*, *14*(1), 10–25.
Kim, D. H., Lee, B. K., & Sohn, S. Y. (2016). Quantifying technology-industry spillover effects based on patent citation network analysis of unmanned aerial vehicle (UAV). *Technological Forecasting and Social Change*, *105*, 140–157.
Kuusi, O., & Meyer, M. (2007). Anticipating technological breakthroughs: Using bibliographic coupling to explore the nanotubes paradigm. *Scientometrics*, *70*(3), 759–777.
Leydesdorff, L., Kushnir, D., & Rafols, I. (2014). Interactive overlay maps for US patent (USPTO) data based on International Patent Classification (IPC). *Scientometrics*, *98*(3), 1583–1599.
Li, Y. A. (2014). Borders and distance in knowledge spillovers: Dying over time or dying with age?—Evidence from patent citations. *European Economic Review*, *71*, 152–172.
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.
Lo, S. C. (2007). Patent coupling analysis of primary organizations in genetic engineering research. *Scientometrics*, *74*(1), 143–151.
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.
Murata, Y., Nakajima, R., Okamoto, R., & Tamura, R. (2014). Localized knowledge spillovers and patent citations: A distance-based approach. *Review of Economics and Statistics*, *96*(5), 967–985.
Nakamura, H., Suzuki, S., Sakata, I., & Kajikawa, Y. (2015). Knowledge combination modeling: The measurement of knowledge similarity between different technological domains. *Technological Forecasting and Social Change*, *94*, 187–201.
Noailly, J., & Shestalova, V. (2017). Knowledge spillovers from renewable energy technologies: Lessons from patent citations. *Environmental Innovation and Societal Transitions*, *22*, 1–14.
Park, H., & Magee, C. L. (2017). Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. *PloS One*, *12*(1), e0170895.
Park, I., Jeong, Y., Yoon, B., & Mortara, L. (2015). Exploring potential R&D collaboration partners through patent analysis based on bibliographic coupling and latent semantic analysis. *Technology Analysis & Strategic Management*, *27*(7), 759–781.
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.
Russell, S., & Norvig, P. (1995). *Artificial Intelligence: a modern approach*. Prentice-Hall.
Swanson, D. (1971). Some unexplained aspects of the cranfield tests of indexing performance factors. *The Library Quarterly*, *41*(3), 223–228.
Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The Rand Journal of Economics*, 172–187.
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.
Von Wartburg, I., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, *34*(10), 1591–1607.
Yeh, H. Y., Sung, Y. S., Yang, H. W., Tsai, W. C., & Chen, D. Z. (2013). The bibliographic coupling approach to filter the cited and uncited patent citations: A case of electric vehicle technology. *Scientometrics*, *94*(1), 75–93.

Chung-Huei Kuan is an assistant professor of the Graduate Institute of Patent at National Taiwan University of Science and Technology, Taipei, Taiwan. His research interests include patent bibliometrics, patent information mining and analysis, and practices in patent prosecution, patent specification drafting, and patent/technology transfer and licensing.

Mu-Hsuan Huang is a distinguished professor of the Department of Library and Information Science at National Taiwan University, Taipei, Taiwan. Her early research focused on information retrieval and information behavior, and turned to bibliometrics, science and technology policy, intellectual property, and patent information for late years. She is also the project investigator of Performance Ranking of Scientific Papers for World Universities (NTU Ranking).

Dar-Zen Chen is a professor of the Department of Mechanical Engineering and Institute of Industrial Engineering at National Taiwan University, Taipei, Taiwan. His research interests include intellectual property management, patentometrics, competitive analysis, robotics, automation, kinematics, and mechanism design. He also leads the Intellectual Property Analysis & Innovative Design Laboratory (IAID).