





Patent Citation Network Analysis Based on Improved Main Path Analysis: Mapping Key Technology Trajectory

Zikui Lu , Yue Ma, and Luona Song 

University of Posts and Telecommunications, Beijing 10036, China

Abstract. Nowadays, more and more people realize the importance of patent for innovation activities. Patent citation network analysis is one of the most important methods for patent measurement, patent mining, and core patent identification. In nowadays, finding technology trajectories and analyzing major technologies in patent networks are intensively used in technological competition. Main path analysis (MPA) is a famous directed graph-based method to extract main paths in certain networks, such as a citation network. However, the accuracy of main path identification may be distracted due to a large volume of wrong references when using MPA in patent citation networks solely. To tackle this challenge and extract reasonable main paths from patent citation network, in this paper, we combined the classic MPA with the PageRank algorithm and we tested this new combined method on authorized patent datasets. The results show that the improved method achieved better performance in average cited frequency and other indicators of core patents comparing with traditional MPA.

Keywords: Citation network · Main path analysis · PageRank

1 Introduction

For any researchers who want to engage in researching and developing new innovative technologies in a scientific domain, the first step is to understand the evolution of core technologies and to find the key point. Therefore, we need to extract core technology among countless documents before we start to study. Patent documents, recording innovation completely and detailly, are regarded as the foundation of technology, products, and processes. According to the World Intellectual Property Organization, 90% of the world's technologies are registered in patent, so make use of patent information effectively will contribute greatly to master technology and creation. Additionally, analyzing core patents is the most efficient way to use patent information. Patent documents contain references and a large number of relationships of citation constitute a patent citation network, which can be analyzed to find core patents. In general, innovations are based on the existing research, as well as the citation network demonstrates an inheritance relationship according to the idea of technological evolution, thus, every patent can be regarded as a piece of a fragment of knowledge. The main paths in the network are the backbone of technological progress, which include more core patents.

The study of references, as one of the means to trace patterns of technological advancements, has been growingly used among scientific publications. As a result, many methods have been proposed to exact and represent the evolution patterns of ideas, including algorithmic historiography, knowledgeflow, and influence trajectories [1]. However, those means are considered too complex to understand. Thus, new ways have been designed to find critical subnetworks as a summary of core patents. Main path analysis (MPA), originally proposed by Hummon and Doreian [2], has been applied in various domains to extract main paths. Compared with historical and descriptive methods used before, it enriches the engineering perspective for technological research and makes a significant contribution to the identification of core patents and extraction of the mainstream [3].

Although applied by many studies of citation networks, MPA is not always a proper method to exploit the main paths of patent citation networks. Through experiments, we have discovered the number of patent references can reach hundreds or even thousands incredibly, which will cause the main paths to concentrate on those patents because MPA includes a greedy algorithm. Once the citation network contains nodes with large references, the result will be not accurate enough through MPA.

PageRank, an algorithm proposed by Google to rank webpages, can evaluate the quality of webpages. It assigns relative importance and authority scores to each webpage through hyperlinks, which can be regarded as the basis for webpage ranking. Absolutely, the algorithm is still applied in most search engines. Inspired by that, we use the idea of PageRank to calculate the importance of the patents, then traverse paths of citation network to extract main paths. The results show obvious improvements in the average cited frequency and coverage ability of core patents founded by our methods.

2 Related Work

2.1 Main Path Analysis

The goal of the MPA is to find main paths. A main path of citation network is a connected subnetwork of important citation arcs connecting a number of origin nodes (typically, historical files) to destination nodes (typically, recent publications) [4]. Citation arcs, if the arcs sit on many paths between origin nodes and destination nodes, are considered as have evolutionary importance to a certain filed, thus it is appended to the main paths of the corresponding citation network. Briefly speaking, main paths approximately demonstrate development trajectories among the major ideas.

First step of the MPA is to weight citation arcs based on some distinctive features. In terms of knowledge transfer, the weight of citation arc (u, v) is measured by the number of search paths passing through (u, v) between a set of sources and destinations. Regularly, Search Path Count (SPC) and Search Path Node Pair (SPNP) are used to weight citation arcs [5].

SPC measures the importance of the arcs in the network by counting the number of times the connection between two adjacent nodes traversed by all paths of the corresponding network. Let $N^-(m)$ denotes the path from the source point s to the point m , and $N^+(n)$ denotes path from node n to destination t , then for any path from s to t passes

arc (m, n) , the value of the arc π is [6]:

$$\pi = \sigma \cdot (m, n) \cdot \tau, \quad (1)$$

where σ denotes all paths from s to m , τ denotes all paths from n to t . The mathematical expression of the weight $N(m, n)$ of the citation arc is shown in Eq. 2:

$$N(m, n) = N^-(m) \times N^+(n) \quad (2)$$

$$N^-(m) = \begin{cases} 1 & m = s \\ \sum_{n:nRm} N^-(m) & \text{others} \end{cases} \quad (3)$$

$$N^+(m) = \begin{cases} 1 & m = t \\ \sum_{n:nRm} N^+(v) & \text{others} \end{cases} \quad (4)$$

where nRm denotes m cites n .

After obtaining the weighting directed graph, for extracting main paths, MPA then use deep first method and greedy algorithm to get paths consisting of arcs with highest traversal weight from sources and destinations.

Since SPC is the most typical algorithm for weighting network, in the comparative experiment we also use SPC to weight arcs.

2.2 PageRank

The reason why a page has a link to another one is because the later has more authoritative and reliable content. The basic idea of PageRank to rank webpages is to calculate number of links to the webpages. [7] They proposed a concept of PR value represent importance of one page. Furthermore, the value of each page is related to the number of external links that link to its own and their PR value. Therefore, the method reduces impact of link number on weight. For example, webpage $w1$ is linked by a meaningful webpage and webpage $w2$ is linked by many webpages with poor information, because of the different PR value among webpages link to $w1$ and $w2$, the PR value of $w1$ may large than $w2$. Formula for PageRank is as follow:

$$PR(u) = d \sum_{v \in B_u} \frac{PR(v)}{L(v)} + \frac{1-d}{n}, \quad (5)$$

where B_u denotes a collection of all webpages link to webpage u , v denotes a webpage belonging to B_u , $L(v)$ denotes the number of external links of webpage v , d denotes damping factor, n denotes the number of all the webpage in this network.

3 Datasets

In this paper, we use ‘‘Etching the Insulating Layers’’ and ‘‘Encapsulations’’ in chip manufacturing as keywords to draw the core patent technology trajectories [8] for two patent sets respectively.

3.1 Datasets

As a first step, we build patent datasets respectively. The datasets are extracted from USPTO(United States Patent and Trademark Office) [9], which including all the US patents from July 31,1790 to present as well as full searchable text. When searching related patents, USPTO allows patents to be retrieved by International Patent Classification (IPC). The IPC strategy selects patents by consulting a broad range of secondary sources and searching in International Patent Classification Table [10]. Then, we can confirm query statements “ICL/H01L21/311” and “ICL/H01L21/02 AND ICL/H01L21/56” to retrieval related patents. The partial table of IPC is shown in Table 1.

Table 1. Internet patent classification table

IPC	Description
H01L21/00	Processes or apparatus specially adapted for the manufacture or treatment of semiconductor or solid-state devices or of parts thereof
H01L21/02	Manufacture or treatment of semiconductor devices or of parts thereof
H01L21/311	Etching the insulating layers
H01L21/56	Encapsulations, e.g. encapsulating layers, coatings

Then, we analyze the number of patent references, and remove the patents with unreasonable citations from datasets. When it comes to patents related to “Etching the Insulating Layers”, we obtain 5496 patents with application time span from 1975 to 2020. By exploiting distribution of patent references we remove patents with number of citations more than 300, the citations of those patents are incredible, which can cause unnecessary calculations. Figure 1 shows the distribution of the patent citations, after removing patents with unreasonable references, there are 5448 patents left. In same way, there are 2921 patents build dataset of patents related to “Encapsulations”.

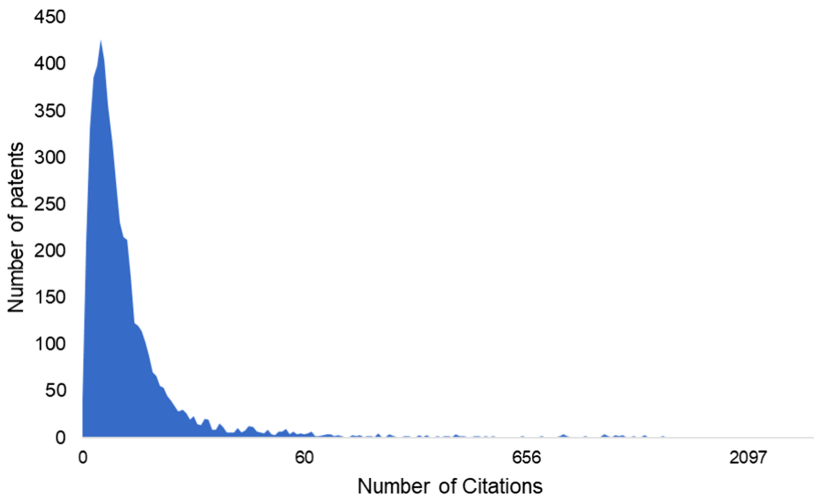


Fig. 1. Distribution of the citations of patents related to “Etching the Insulating Layers”

3.2 Improved Method

In the present article, we use PageRank [10] to improve weighting algorithm for MPA and give patent nodes value directly, instead of using the method of giving value of arcs, which helps to reduce the effects of unreasonable references to main paths.

Supposing there are two citation networks like Fig. 2. Left network shows relationship of single reference, node *A* references node *B* (knowledge flow from *B* to *A*) and node *B* references node *C*. According to SPC, it's obvious that weight of arcs *BA* and *CB* are both 1. Comparingly, right network demonstrates node *B* cites many other nodes. Although relationship of *BA* in both networks are same, with the number of node *B* references increases the value of *BA* will raise correspondingly according to MPA. The number of paths from source nodes to the destination nodes pass *BA* are increasing. So, if node *B* has unreasonable references, the mistake references can amplify the weight of arc *BA*. Due to the main paths is the path consist of largest citation arcs, the arc *BA* in later network has greater probability be selected to the main path and node *B* be regarded as core patent.

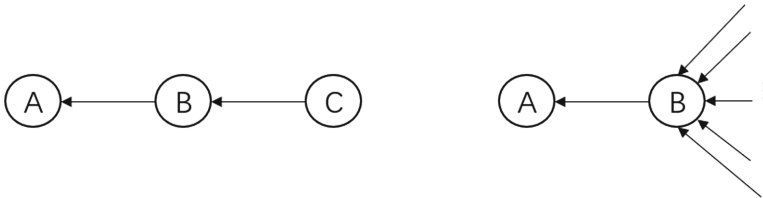


Fig. 2. Two cases of patent citation

In fact, a network composed of thousands of nodes is much more complicated than the case of graph with only two or three nodes. Normally, about five thousand patents and their citations can build a network with even millions of paths, while, any operation for nodes will cause huge effect on the network, especially for weight of arcs. Table 2 shows the impact of changes of references in patent node ‘3951843’ on the network.

Table 2. Changes in network when the number of patent citations changes

Number of citations	Number of paths in network	Weight of related arc
8	3125682	225
6	3125464	175
5	3125355	150
3	3125028	75
1	3124810	25

Therefore, to prevent unreasonable references from distracting the main paths, weighting algorithm should be improved. In order to weighting more reasonable, before we calculate value of patents, we assume each node that not be referenced has same initial value I , and the other nodes has initial value i in the network [11]. When a patent has N citations, it divides its value into N parts, and gives a piece of part to every patent it references. In this way, patent will give lower values to other nodes cited once it has many references, comparingly, patent will transfer all value itself to the node cited if it has just one citation. Then we can calculate the real value of each patent in the network, which we call PR value. The expression of calculating PR value is as follow:

$$PR(\text{patent}) = \begin{cases} \sum_{v \in U} \frac{h(v)}{L(v)} & \text{indegree of the patent not equals to } 0 \\ I & \text{indegree of the patent equals to } 0 \end{cases}, \quad (6)$$

where PR denotes PR value of current patent, v denotes patent that cites current patent, U denotes a collection of patents, $h(v)$ refers to the initial value of a patent. While, in our method, we no longer assign weight to arc between nodes, but give value to patent nodes directly, which means that, when calculating the weight of the patents in the network, we have to build network according to relationship of reference. When extracting main paths according to direction of knowledge flow. The Algorithm of SPC and improved weighting algorithm is shown in Table 3 and Table 4 respectively.

Table 3. Algorithm of SPC

Algorithmic 1 Method of Search Path Count

Input: *oldGraph*, *startNodes*[], *paths*[]

Output: *newGraph*

Function *WeightingGraph(oldGraph,startNodes[],paths[])*

- 1: **for** each *path* \in *paths*[] **do**
- 2: **for** each $i, j \in$ *path*[] **do**
- 3: *OldGraph*[i][j]['weight'] \leftarrow *oldGraph*[i][j]['weight']+1
- 4: **end for**
- 5: **end for**
- 6:**end function**

Table 4. Algorithm of improved method**Algorithmic 2** Improved Weighting Method**Input:** *oldGraph* , *startNodes*[], *endNodes*[], *n*[], *noded*[], *I*, *i***Output:** *newGraph*

```

1: function CoPagerank(oldGraph,startNodes[],endNodes[],noded[],n())
2:   for each node  $\in$  n[] do
3:     if node  $\in$  startNodes[] then
4:       oldGraph[node][‘pr’]  $\leftarrow$  I
5:     end if
6:     NAL  $\leftarrow$  len(list(oldGraph[node]))
7:     nodeA  $\leftarrow$  list(oldGraph[node]))
8:     for nextnode in [nodeA] do
9:       if (node,nextnode)  $\in$  noded then
10:        do nothing
11:       else
12:         oldGraph[nextnode][‘pr’]  $\leftarrow$  oldGraph[nextnode][‘pr’]+i/NAL
13:         noded.append(node,nextnode)
14:       end if
15:     end for
16:     for nextnode in nodeA do
17:       if nextnode not in endNodes[] then
18:         CoPagerank(oldGraph,startNodes[],endNodes[],noded[],n())
19:       end if
20:     end for
21:   end for
22: end function

```

After getting directed graph, we follow the idea of MPA and traverse all paths to find the paths with largest PR value, that is main paths.

3.3 Evaluation

For better measure importance of the main paths, we propose some indicators. A patent citation network often contains more subnetworks, which reflect different branches in the development of this domain. When the degree values of all nodes of a subnet are greater than or equal to K value, the subnet is called K -core graph. Because the cited frequency of patents is equal to out-degree value of corresponding nodes in citation network, the K value here is considered as out-degree value. The larger the K value is, the larger the average out-degree value of the subnet is, and the more nodes the subnet can affect, so the subnet is more important to the citation network. Consequently, for a collection of nodes, calculating the number of nodes in different K -core graphs of a network can reflect the importance of the collection to it. We can measure importance of main paths that extracted by MPA and method we improved respectively through this

way. Since out-degree value can reflect cited situation of node in network, so average out-degree value of main paths also reasonable indicator for measure importance of main paths.

K-core decomposition is regular method help find core subgraph (the largest connected subgraph) in network [12]. In K-core graph, all the edges including nodes whose degree value less than k also been removed from network. 1-core net represents original graph, and $(K + 1)$ -core net is subgraph of K-core graph. We can also observe changes by measuring graph density, average degree, number of connected blocks, average clustering coefficient and other indicators besides number of nodes and edges.

Graph density of a simple graph is defined as the ratio of the number of edges with the maximum possible edges. Since patent citation network has no mutual references, the more nodes the network has, the lower density of the net will have. Because the new node cannot refer to all previous nodes in citation network. For directed simple graphs, the density is:

$$D = \frac{|E|}{2\binom{|V|}{2}} = \frac{|E|}{|V||V-1|} \quad (7)$$

where E is the number of edges, V is number of nodes.

Six Degrees of Separation is a conjecture in the field of mathematics, which point out that it spends only 6 steps to connect any two people in the world on average. In other words, if each individual build link to their friends, link relationship will cover all the individuals after 6 times traversal. Based on the Six Degrees of Separation, we can measure the coverage ability of core patents to judge the importance. In patent citation network, if node A refers to node B , then node B has an impact on node A . We can consider that node B connects to node A . Because core patents have more influence than other patents, so they can connect to most patent nodes faster than others. Under the same conditions, we can measure importance of two sets of core patents by comparing the number of other patent nodes the core patents node can connect to in citation network.

4 Experiment Result

First of all, we build patent citation networks according to the related patents of “Etching the Insulating Layers” and “encapsulation” technologies. Among them, the former graph has 40605 nodes and 77426 edges; the later graph has 21344 nodes and 45865 edges. “Etching the Insulating Layers” related patent references network is shown in Fig. 3. Different color nodes denote different out-degree value.

Secondly, we calculate main paths through our method and MPA respectively and we will analyze those patents below.

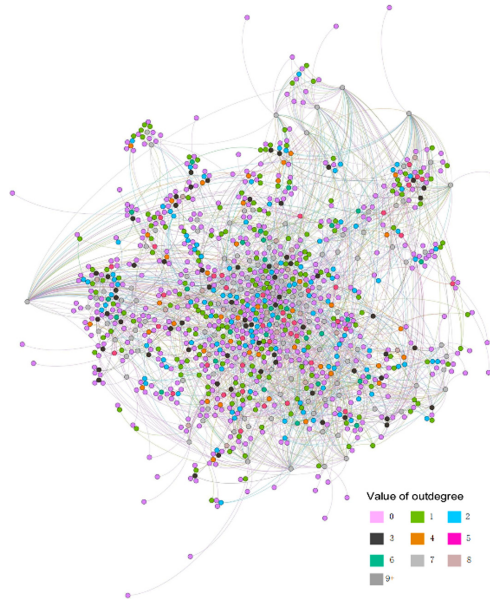


Fig. 3. Citation network of patents related to “Etching the Insulating Layers” (part)

Figure 4 shows the main paths of patents related to “Etching the Insulating Layers” found by our method and MPA.

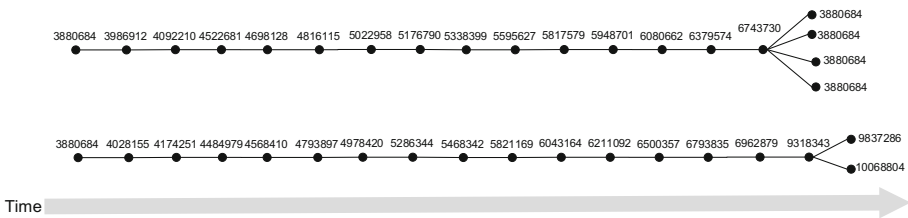


Fig. 4. Main path calculated by MPA and our method, the above is extracted by our method, the below is extracted by MPA. The patent on the left is granted before the patent on right

4.1 K-core Graph

We decompose graph through K-core decomposition according to set $k = 10, 15, 20, 25, 30$, and then analyze connected subgraphs by calculating features mentioned above, preparing for the experiment of counting number of patents existing in k-core graph. Table 5 shows the relevant data of the k-core graph of the patents related to “Etching the Insulating Layers”.

Obviously, the numbers of nodes and edges are decreasing sharply as the K value of the network increases, which shows scale of net is becoming smaller and smaller.

Table 5. Data of k-core graph of the patents related to “Etching the Insulating Layers”

K	Nodes	Edges	Density	Connect blocks
K10	2626	13303	0.002	50
K15	1264	5442	0.003	49
K20	720	2780	0.005	37
K25	468	1439	0.007	31
K30	321	679	0.008	27

Additionally, when K value of the network increases, the density of graph becomes greater, which means that nodes remaining in K -core graph with high average cited frequency-cy, so they have higher reference value than nodes being removed. With the K value increasing, the number of connect blocks is decreasing, which means that the technologies represented by remaining connect blocks have more reference value than others.

In next experiment, we set $k = 1, 5, 10, 15, 20, 25, 30$ to decompose patent citation networks of two technologies respectively. Then we calculate number of core patents that existing in different K -core graphs of citation network for two technologies, the core patents were extracted by our method and MPA respectively. The result demonstrates our method has superior performance than MPA. The result as is shown in Fig. 5. In the experiment of patents related to “Etching the Insulating Layers”, the number of core patents that remaining in different subgraphs through our method is greater than that through MPA, and all of the subgraphs with high average cited frequency. In the experiment of patents related to “Encapsulation”, our method also performs better than MPA.

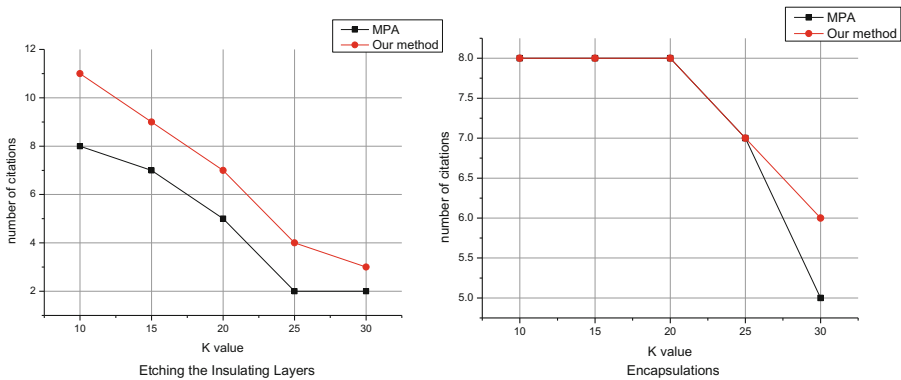


Fig. 5. Situations of core patents in different k-core graphs of two citation networks

4.2 Average Cited Frequency and Coverage

In the following experiments, the results show core patents found by our method are worthier.

We calculate the average cited frequency of core patent sets extracted by the two methods. Result of experiment of patents related to “Etching the Insulating Layers” shows that the value of average cited frequency (average out-degree value) of core patents extracted by our method and MPA are 17.66 and 11.61, as well as, the number in experiment of patents related to “Encapsulations” are 30.9 and 20.61 respectively. A patent with high cited frequency shows this patent is worthy of references [12]. So, a set of patents with high value of average cited frequency is more valuable than the set with lower number. The results prove main paths founded by our method show better performance. The detailed data is shown in Table 6.

Table 6. The results of experiments of two patent citations

feature	Etching the Insulating Layers		Encapsulations	
	MPA	our method	MPA method	our
Average cited frequency	11.61	17.66	20.61	30.90
2-cluster	251	319	688	748
3-cluster	544	795	882	894
K20	5	7	8	8
K30	2	3	5	6

In order to test the influence of core patent nodes, we calculate the number of patent nodes that can be affected by two sets of core patent nodes under the same situation. When the traversal time equals to 1, we calculate the number of patent nodes that reference the core patents, because these patents are directly affected by the core patents. At the same time, we call the collection of affected nodes 1-cluster. When the traversal time equals to 2, we calculate the number of patents that reference the patents that affected by the core patents directly, which are indirectly affected by the core patents. In this case, we call the collection of affected nodes 2-cluster, and so on [13].

In experiment of patents related to “Etching the Insulating Layers”, the numbers of patents affected by core patents, which extracted by our method, are 319 and 795 respectively. In the experiment of patents related to “Encapsulation”, the numbers are 748 and 894 respectively. All of the results are greater than number in MPA. Figure 6 shows the situation of nodes affected by core patents when traversal time equals to 3, core patents in left figure are extracted by MPA and core patents in right figure are extracted by our method. Affected nodes are marked with red color, unaffected nodes are marked with black color. Obviously, the core patents in the right figure have stronger influence[14].

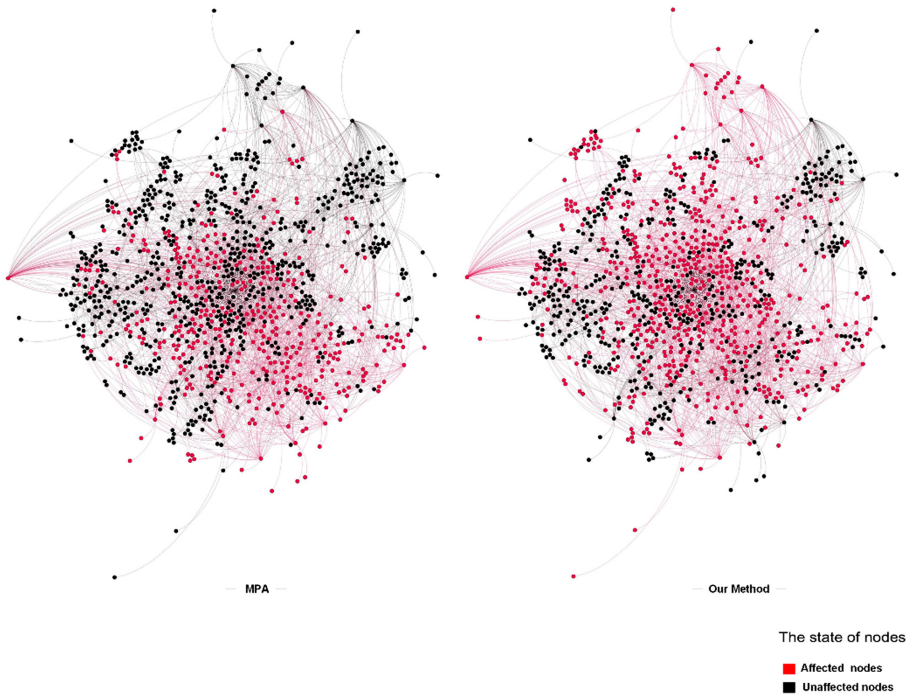


Fig. 6. Situation of 3-cluster of MPA and our method (Color figure online)

4.3 Discussion Main Paths and Related Domains

In main paths of patents related to “Etching the Insulating Layers”, the result contains 18 patents extracted from 5448 patents, whose time span is from 1975 to 2018. And technologies of core patents include “semiconductor preparation method”, “thin film etching process”, “high-density plasms etching” and other methods. In the result of patents related to “encapsulation”, main paths contain 12 patents extracted from 2921 patents and the time span is from 1970 to 2019. Those patents include technologies of “Solder interconnection structure on organic substrates”, “Semiconductor device and method of forming micro interconnect structures”, “BAG packaging” and so on. After consulting relevant literature, those patents can better describe development of the domain respectively. Table 7 demonstrates part of the information of core patents in “Etching the Insulating Layers”.

Table 7. Information of core patents related to “Etching the Insulating Layers”

Patent. no	Issue year	Title	Assignee name
3880684	1975	Process for preparing semiconductor	Mitsubishi Denki Kabushiki Kaisha
4174251	1979	Method of selective gas etching on a silicon nitride layer	ITT Industries, Inc
4568410	1986	Selective plasma etching of silicon nitride in the presence of silicon oxide	Motorola, Inc
4793897	1988	Selective thin film etch process	Applied Materials, Inc.
9318343	2016	Method to improve etch selectivity during silicon nitride spacer etch	Tokyo Electron Limited

5 Conclusion

After experiments, patents extracted by improved method show better performance than MPA in average value of cited frequency and ability of coverage. Based on PageRank, our method reduces the impact of unreasonable references of patents and solves the problems caused by over citations of patents.

Certainly, there are still defect in our method, for example, our method doesn't take damping factor into consideration. Researchers cannot consult information endlessly during they creation, and the factor can help us approach the real situation. So, there should be another parameter in the algorithm to represent this constraint. In the following research, we plan to conduct more experiments in more fields to find reasonable damping factor, and improve our method further.

Acknowledgement. I would like to thank professor Mei Song for her important comments on several drafts of this work. I thank associate professor Xiaojuang Wang for the stimulating and meaningful guidance she made as a reviewer. I gratefully acknowledge help from my parents, they always encourage me when I am frustrated, so I can finish this work.

Funding Statement. This work was supported by the National Natural Science Foundation of China (61601053).

References

1. Yi-Ning, T., Shu-Lan, H.: Constructing conceptual trajectory maps to trace the development of research fields. *J. Assoc. Inf. Sci. Technol.* **67**(8), 2016–2031 (2015)
2. Hummon, N.P., Dereian, P.: Connectivity in a citation network: the development of DNA theory. *Soc. Netw.* **11**(1), 39–63 (1989)
3. Kumar, A., Mishra, S., et al.: Link prediction in complex networks based on significance of Higher-Order Path Index (SHOPI). *Phys. A* **545**(1), 1–17 (2020)

4. Xiao-Rui, J., Xiao-Hao, Z., et al.: Main path analysis on cyclic citation networks. *J. Assoc. Inf. Sci. Technol.* **71**(5), 578–595 (2020)
5. Jiang, X., Hai, Z.H.: Forward search path count as an alternative indirect citation impact indicator. *J. Informetr.* **13**(1), 1–28 (2019)
6. Liu, J.S., Lu, L.Y.Y., Ho, M.-C.: A few notes on main path analysis. *Scientometrics* **119**(1), 379–391 (2019). <https://doi.org/10.1007/s11192-019-03034-x>
7. Reinstaller, A., Reschenhofer, P.: Using PageRank in the analysis of technological progress through patents: an illustration for biotechnological inventions. *Scientometrics* **113**(3), 1407–1438 (2017). <https://doi.org/10.1007/s11192-017-2549-x>
8. Mina, A., Ramlogan, R., et al.: Mapping evolutionary trajectories: applications to the growth and transformation of medical knowledge. *Res. Policy* **36**(1), 789–806 (2007)
9. Batagelj, V., Ferligoj, A., Squazzoni, F.: The emergence of a field: a network analysis of research on peer review. *Scientometrics* **113**(1), 503–532 (2017). <https://doi.org/10.1007/s11192-017-2522-8>
10. Epicoco, M.: Knowledge patterns and sources of leadership: mapping the semiconductor miniaturization trajectory. *Res. Policy* **42**(1), 180–195 (2013)
11. Ruas, P., Lamurias, A., Couto, F.M.: Linking chemical and disease entities to ontologies by integrating PageRank with extracted relations from literature. *J. Cheminform.* **12**(1), 1–11 (2020). <https://doi.org/10.1186/s13321-020-00461-4>
12. Su-Fen, Z., Peng, R., et al.: HeteroRWR: a novel algorithm for top- k co-author. *IEICE Trans. Inf. Syst.* **103**(1), 71–84 (2020)
13. Hamed, A.Y., et al.: A genetic algorithm to solve capacity assignment problem in a flow network. *Comput. Mater. Continua* **64**(3), 1579–1586 (2020)
14. Cai, C., Xu, H., et al.: An attention-based friend recommendation model in social network. *Comput. Mater. Continua* **65**(3), 2475–2488 (2020)