Patent Analysis for China's Epidemic Prevention Products

---- Based on the Perspective of Network

Course: CASA0005 Geographic Information Systems and Science

Programme: Msc Spatial Data Science and Visualisation

Department: Centre for Advanced Spatial Analysis

Repository with Data and Scripts

https://github.com/Xudong-zhang1994/GIS-Assignment-Reproducible-Analysis.git

Patent Analysis for China’s Epidemic Prevention Products
---- Based on the Perspective of Network

Introduction

At the beginning of 2020, COVID-19 started to spread all over the world. In support of scholars studying the level of innovation in products designed to contain the spread of the epidemic, China's State Intellectual Property Office has published patents of China's epidemic prevention products (EPP) from 1993 to 2019. This dataset involves the medicine, medical device, data analysis, and other fields patent.

Patents analytics have been regarded as an effective method to detect the level of innovation in one region. Due to the invention of ‘cutting-edge’ technologies, researchers started to applied various methods, such as machine learning, modelling for spatial statistical analysis, to discover the knowledge of patents from the different research fields.

This paper is mainly to find a way to reflect the possible connection among patent knowledge in the form of network and project this network on the city level. This study proposed an approach to measure the number of potential knowledge connections between two cities based on the International Patent Classification (IPC). This was done by applying an analytical framework, which is proposed by Peter. J. Taylor and his colleagues(Taylor, 2001), and a potential city network has been built based on the interconnection of knowledge of patents. This city network could also be used to find the cities that have a large number of potential knowledge connections and detect the relationships among these cities by using social network analysis methods.

Literature Review

Globalization and technological progress lead to increasingly close network connections in various fields. Scholars started to focus on studying the interrelations between different subjects from the perspective of network, such as Global Production Network (GPNs), Global Value Chain (GVCs), Global Service Network, Global City Network, Global Innovation Network, Global Trade Network, and other network forms(Castells, 1996; Taylor, 2001). As for global city network, King regarded that Friedmann and

Wolffs proposed a ‘global network of cities’. (King, 1990) It was the start of the contemporary study of world cities (Taylor, 2001). From then on, many concepts about “world cities” were raised. The notions and analytical frameworks have prompted precise analysis of world cities, quantitatively, or qualitatively.

Since the 1980s, relevant research on innovation has embedded geographical elements, and "new regionalism" has played an important role in the analysis framework of the formation mechanism of innovation activities. Boschma R A applied quantitative methods and theory from evolutionary economics to reveal the mechanism of distance and knowledge at different stages(Ron A. Boschma and Koen Frenken, 2006). In recent years, scholars tried to classify different types of industry based on their knowledge. Asheim et al. (Asheim, Coenen and Vang, 2007) divided industries into three types: science-based, engineering-based, and art-based industry. Plum et al. From the perspective of knowledge base, discussed the difference of knowledge network characteristics between the two knowledge bases: analytical knowledge base and synthetic knowledge.(Plum and Hassink, 2011) Their works showed that research on innovation network has been focusing on more detailed knowledge to improve the accuracy of the network.

Some scholars also studied the impact of internal and external relations of enterprises on innovation performance, based on a specific industry. Taking the pharmaceutical industry as an example, they found that functional proximity is more important to the innovation of the biomedical industry than geographical proximity(Coenen, Moodysson and Asheim, 2004). However, other scholars believe that the interactions among localized organizations are the internal mechanism of cluster innovation(Bagchi-Sen, Smith and Hall, 2004).

Overall, scholars analysed various types of data, such as patent data, trade data, the number of multinational companies and their overseas subsidiaries (branches), to find the characteristics of knowledge flow, the form of agglomeration in the different scales of space, the evolution of the network, and the influence of different factors on innovation. For patent analysis, many scholars tried to mine the knowledge of patents, by using the methods of clustering and deep learning, to get series of knowledge characteristics of patents and predict future inventions in this field. But these studies lacked cities elements(de la Paz-Marín, Campoy-Muñoz and Hervás-Martínez, 2012; Jun, Park and Jang, 2014). City network based on knowledge characteristics needs to explore more in the future.

Research Hypothesis

The purpose of this paper is to try to establish a city network based on the potential connection between knowledge of patents. The establishment of this network requires the following assumptions:

H1: Patent classification is hierarchical, which can be regarded as a pyramid model. It is easier to innovate by integrating knowledge from the same domain than by integrating knowledge from different domains. For example, in the case of epidemic prevention products, the process of inventing pharmaceutical patent may be more inclined to use the methods from chemical-pharmaceutical than fixed constructions.

H2: The knowledge coded by the complete IPC number has its unique characteristics. For a patent, a complete IPC number contains relatively clear information of one area and unique knowledge background of the patent invention. Thus, a complete IPC number could be regarded as a distinct symbol of one piece of knowledge. But at the same time, it still has a classified function so that it could be in contact with other patents that contain the same patent classification number. This requires selection in a specific context. This paper considers that each complete IPC number of a patent reflects the uniqueness of knowledge.

H3: The knowledge of one patent cannot be automatically connected with the knowledge of other patents. The process of connection requires some infrastructures, such as the Internet, interpersonal networks. Intermediaries are also needed, such as individuals who want to make inventions. At this point, knowledge of different patents will be directed to the intermediary when the intermediary uses a 'search engine' to search them. After the intermediary receives different kinds of patent knowledge, it can be considered that a practical contact of different kinds of knowledge has finished. But the potential connection of knowledge of patent does not require the participation of an intermediary.

H4: After the invention of one patent, the knowledge in a patent may exist forever, no matter the status of a patent is, valid or invalid. Changes in the environment may affect the utility of knowledge of the patent, but they will not change its original measurement function based on the original context. This may result that the amount of knowledge in one region is much higher than that in other regions in one period.

H5: If the above hypothesis is true, potential connections between the same area of knowledge of patent can have a product effect.

Fig.1. showed the mechanism of how I define the potential connection of patent knowledge between two cities. In a patent, the complete IPC (IPC1 and IPC4 in Fig.1) can be regarded as the carrier of the unique knowledge of the patent. And a potential connection is between IPC1 and IPC4. When an intermediary performs a query on these pieces of knowledge (IPC1 and IPC4) for a specific purpose, it completes a practical connection between these two pieces of knowledge, thus forming a real network connection (as shown in the figure below). This paper mainly discusses the projection of the potential connection between patent knowledge on the city level, so the intermediary agency can be omitted in this process.

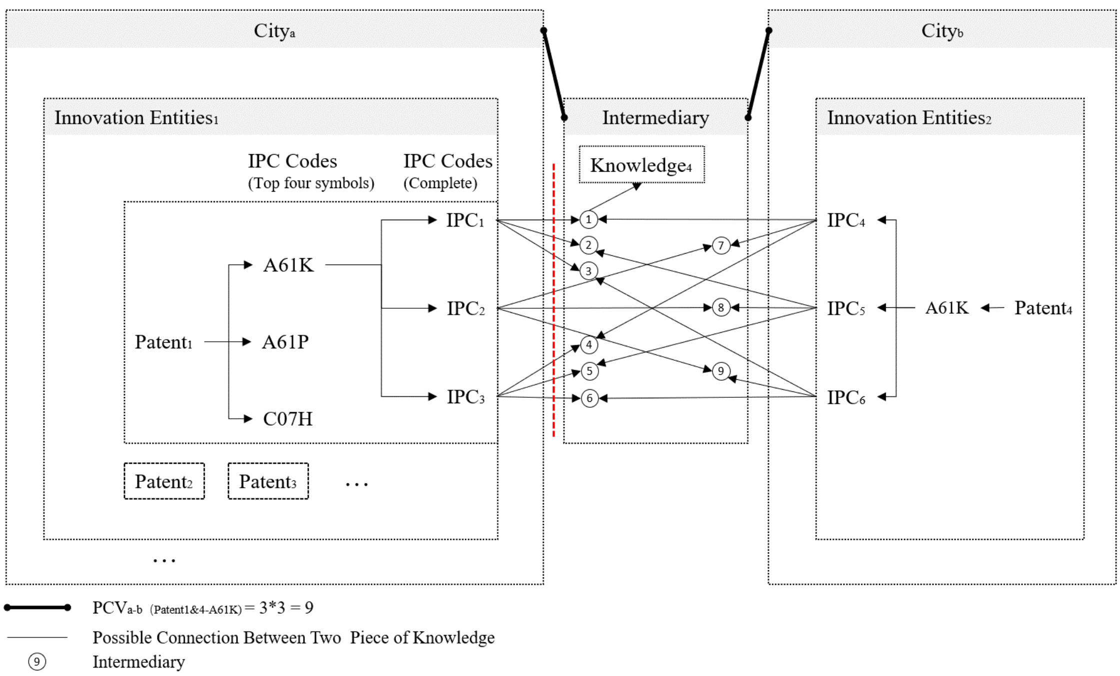

Fig.1. Principles of potential connection of patent knowledge

Data

This data set of patent contains information such as name, abstract, IPC number, disclosure date, applicants, and their detailed address. In terms of patent classification structure, each patent has one or more IPC numbers, which indicates that one patent may use one or more areas of knowledge. (Table 1) The structure of the complete IPC number contains five levels, which are level I-V. Taking C07H19/00 as an example, the first character ‘C’ belongs to level I, which means that the patent belongs to chemistry and metallurgy. C07 belongs to level II, which refers to organic chemistry. C07H belongs to level III, referring to ‘sugars; derivatives thereof; nucleosides; nucleic acid’. C07h19/00 belongs to level IV, and it refers to ‘compounds containing a hetero ring sharing one ring hetero atom with a saccharide radical; Nucleosides; Mononucleotides; Anhydro derivatives thereof’. This patent number does not refer to level V, which is more detailed.

Table 1
Examples of IPC Codes

Publication Codes	IPC Codes
CN103052631A	C07D309/10; C07H19/00
CN103845316A	A61K31/196; A61K31/215; A61K31/27; A61P31/12; A61P31/16; C07C279/16; C07C279/24; C07C277/08
CN102964267A	C07C233/47; C07C231/12; C07C279/16; C07C277/08; C07C255/45; C07C253/30; C07C323/61; C07C319/12; C07C271/24; C07C269/06; C07C311/20; C07C303/40; C07C309/30; C07C303/32; C07F9/44; C07F9/40; A61K31/664; A61K31/662; A61K31/277; A61K31/27; A61K31/245; A61K31/19; A61P31/16

From Fig.2., between 1993 and 2003, no more than five patents were applied for EPP each year. From 2003 to 2005, the annual number of patent applications for EPP increased rapidly, but from 2005 to 2009, the rate of patent applications declined. From 2010 to 2012, the annual number of applications for EPP was relatively stable. Until 2013, it reached the level of 2005. After 2013, the number of annual applications grew faster than any other stage.

The number of patents of EPP in all provinces or cities was counted and the top 15 provinces or cities are shown in Fig.4. Guangdong has the largest number of patents and

Beijing is the second one. Among them, the number of patents in examination or valid in Guangdong province and invalid in Beijing are the most in the three categories respectively (Fig.3.).

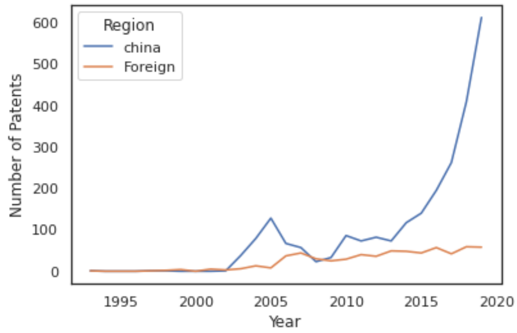

Fig.2. Trends of number of patents

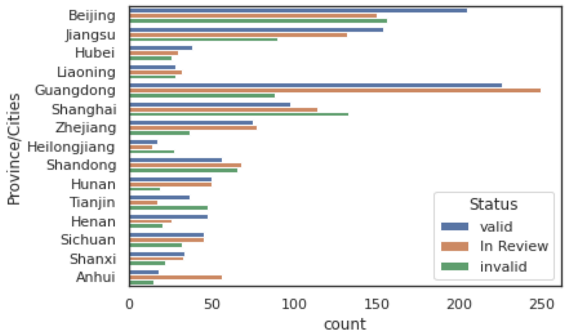

Fig.3. Validity of patents in Top 15 provinces

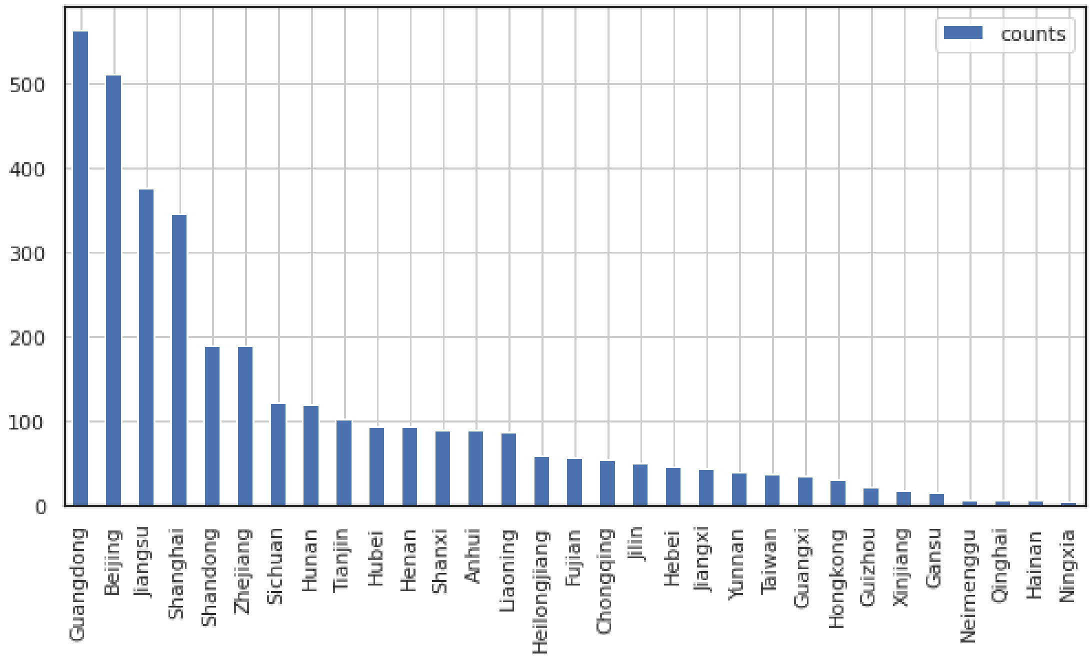

Fig.4. Number of patents in each province(cities).

Methods

This article selects level III as the criteria to count the number of IPC, this is because the number of level III symbols (such as CO7H) can reflect the scale of knowledge in one category, and it can also be used to establish the connection among the different area of knowledge conveniently. The formula is as follows. The number of IPC of city j in field m (such as CO7H) is K_{jm}, and i means the patent i. x_{ijm} means that patent i, whose applicant's address is in city j, has x_{ijm} IPC(s) in field m.

$$K_{jm} = \sum_{i=1}^{n} x_{ijm}$$

Through this formula, the following matrix (K) can be obtained:

Table 2
Matrix of K_{jm}

City j	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	$\boldsymbol{m_3}$
$\boldsymbol{j_1}$	K_{11}	K_{12}	K_{13}
$\boldsymbol{j_2}$	K_{21}	K_{22}	K_{23}
$\boldsymbol{j_3}$	K_{31}	K_{32}	K_{33}
$\boldsymbol{j_4}$	K_{41}	K_{42}	K_{43}

Furthermore, the complete IPC of a city in all fields of Level III can be calculated as C_j, which can be called the reserve value of knowledge in city j. This number is a simple measure of the rank of cities.

$$C_j = \sum_{m} K_{jm}$$

The value of potential connection between cities in field m can be calculated by PCV_{a-b}. PCV_{a-b} represents the potential knowledge connection between city a and b, which is obtained by calculating $K_{am} * K_{bm}$ on each pair of cities. For the hypothesis that the formula is true, please refer to the hypothesis section of this paper.

$$PCV_{a-b} = \sum_{m} K_{am} * K_{bm} \qquad a \neq b$$

After the formula is calculated, the matrix of value of potential connection between cities is formed as follows:

Table 3
Matrix of PCV_{a-b}

City j	$\boldsymbol{j_1}$	$\boldsymbol{j_2}$	$\boldsymbol{j_3}$
$\boldsymbol{j_1}$	0	PCV_{1-2}	PCV_{1-3}
$\boldsymbol{j_2}$	PCV_{2-1}	0	PCV_{2-3}
$\boldsymbol{j_3}$	PCV_{3-1}	PCV_{3-2}	0

Results

After the calculation, the matrix of PCV was imported into Gephi. I select Top 25% of cities in terms of connectivity. The sizes of nodes were classified according to the weighted degree of nodes (the scale of potential knowledge connection between two cities), and the same criteria was used to classify colours of nodes. Thus, I obtained a visualization diagram of city potential connections of knowledge of EPP patents (Fig.5). It can be seen that most of the cities in network are cities in China. Beijing, Shanghai, and Guangzhou are at the core of potential connectivity, the level of scale in these cities is similar. Beijing has strong potential connections with Tianjin, Nanjing, and other cities. Most potential connections among cities reflect a "core- periphery" pattern.

Every IPC number in a patent could be categorized into the main classification or other classifications based on the specification of each patent. So here I regard that this brings me an opportunity to establish the network between the main IPC number and other IPCs, to detect the real connection among the knowledge. To facilitate category, I chose level III of IPC. By importing Gephi, node size and colour are still classified according to node degree and modularity, respectively (Fig.6). It can be found that knowledge of EPP patents presents an aggregation form, and there are cases where specific fields are at the core of the overall knowledge area, such as A61K, A61B, etc.

Also, this paper counts the number of PCV that is not zero and visualizes them on a map (Fig.7). The cities with the highest counts are mainly distributed in the southeast of China, such as Shanghai, Guangzhou, Nanjing, and other cities, where the economy is much better than other regions in China.

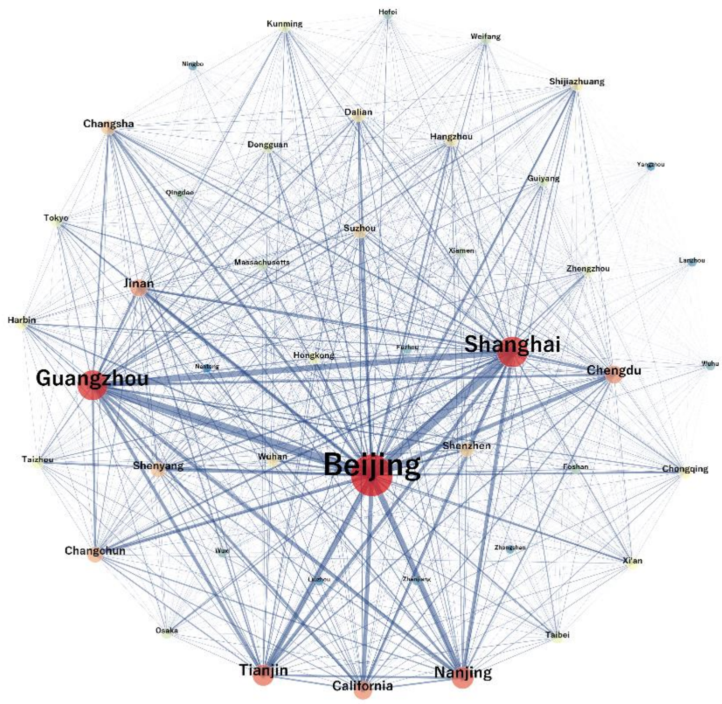

Fig.5. Cities potential connections of knowledge of EPP patents

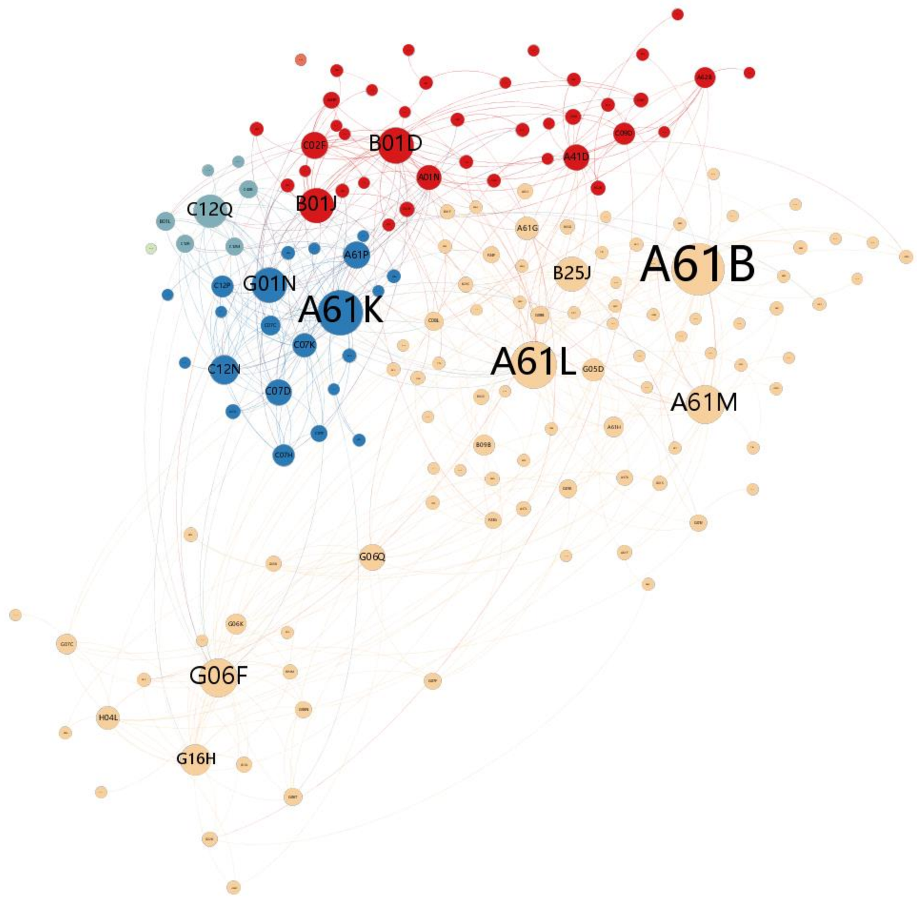

Fig.6. Connections between main IPC number and other IPCs

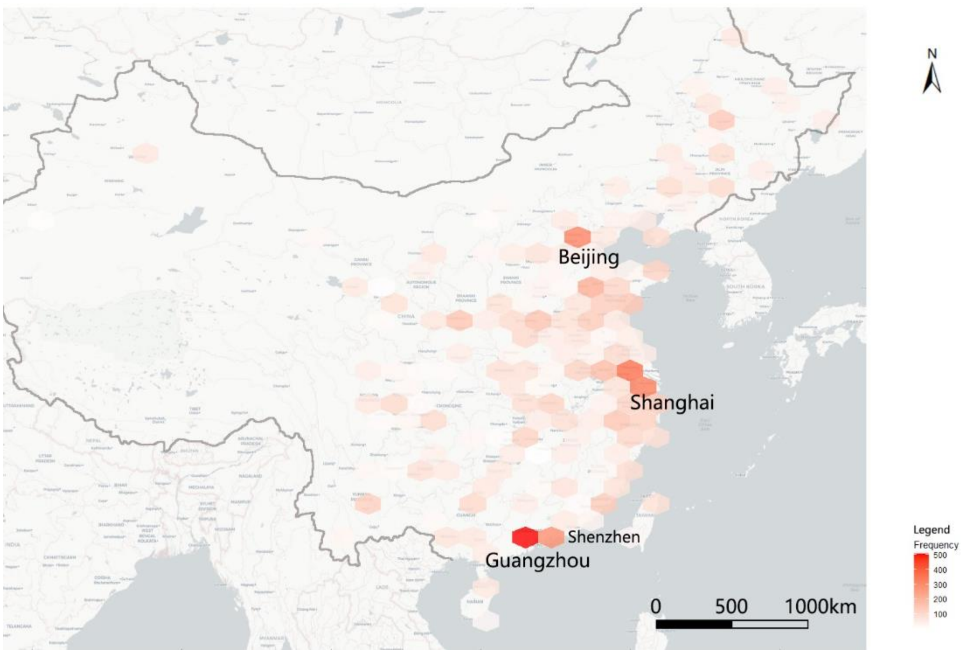

Fig.7. Frequency of potential connectivity in each city

Discussion

The Distribution of Nodes' Degree

In this paper, the network is mainly established by the number of potential connections in the field of EPP patent knowledge, which means that these two networks cover the potential connections in a variety of scales. The node degree value of two networks generally follows the power-law distribution (Fig.8), which means that the two networks are scale-free networks. What's more, each network contains some outliers and most of them have a large amount degree. From the perspective of data, main fields and major cities are in the network, respectively. This can be used to screen out these specific cities for subsequent studies on potential connection networks.

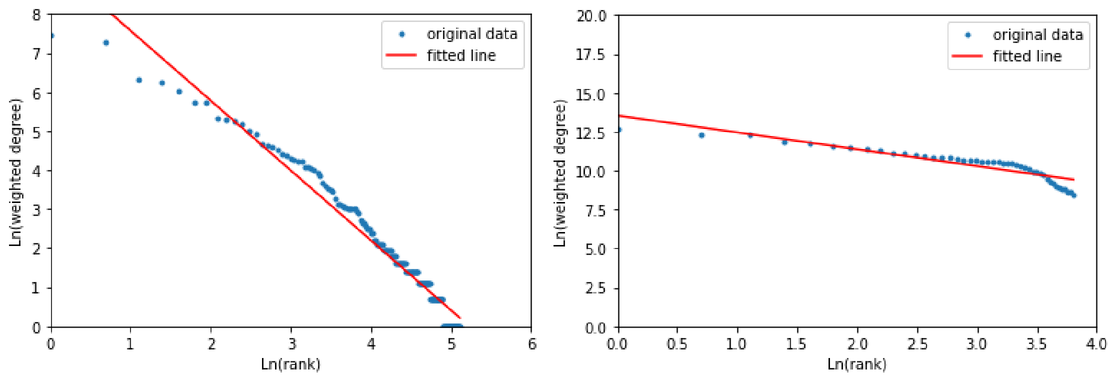

Fig.8. The Distribution of Nodes' Degree

(Left: Patent knowledge network; Right: Potential connection network(top 25% cities))

Table 4
Trend line data

	Patent knowledge network	Potential connection network
Trend Line	y = -1.791 x + 9.367	y = -1.081 x + 13.531
Rsq	0.96	0.84
p-value	1.02 e-119	9.43 e-19

Limitation

In the process of establishing the potential connection network of patent knowledge of EPP, the weight of each patent data is processed uniformly in this paper(w=1). In fact, the knowledge contained in different patents may have different influences on innovation. For example, among the knowledge fields involved in the pharmaceutical innovation, breakthroughs in the field of the chemical industry may have a higher impact on innovation happened in pharmaceutical industry, compared to other fields. This means that further research should pay more attention to the detailed excavation of the knowledge field and technical content involved in the patent, identify the key fields and technologies of patents, and include such factors in the weight matrix for calculation.

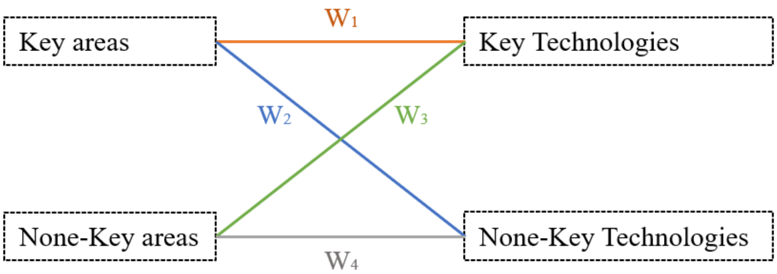

Fig.9. Future research may need to identify key knowledge area and technologies of EPP

Also, the two networks constructed in this paper have a large span of node degree values. Nodes with different degree scales are connected at the same network for analysis, which may obscure the true role of nodes with smaller degree values. For example, the A61K node has a degree of 957, but 35 nodes contain only one degree, such as C11b. Most of them are not belong to category A (level I). So here is a question: aren't these nodes important? Theoretically speaking, cross-field cooperation is more difficult for patent innovation, and it needs to solve more problems in logic, theory, background, and other aspects, which means that the importance of these nodes needs to be determined by deeper mining of patent data, to improve the accuracy of the overall network.

What's more, the potential connection network at the city level lacks the discussion of the interactive connection among industrial clusters within the city. In the future, it is important to pay more attention to the local scale.

Conclusion

This paper attempts to establish a city network based on the potential connection of the patent knowledge of EPP to determine the distribution of the core nodes in the network and the community of the nodes. The results show that Beijing, Shanghai, Guangzhou, and other economically developed areas of China have a strong potential connection of patent knowledge. These areas have been at the forefront of patent research and development of EPP since 1993, and have accumulated profound knowledge in corresponding fields. Also, this paper tries to establish a potential network of patent knowledge. The node degrees of the two networks generally present power-law distribution characteristics, indicating that both networks can be regarded as scale-free networks. Future research should pay more attention to the composition of patented technologies and identify important fields and technologies, to improve the accuracy of the network and provide support for effective prediction of potential key knowledge connections between cities. *(2786 words)*

References

Asheim, B., Coenen, L. and Vang, J. (2007) ‘Face-to-Face, Buzz, and Knowledge Bases: Sociospatial Implications for Learning, Innovation, and Innovation Policy’, *Environment and Planning C: Government and Policy*, 25(5), pp. 655–670. doi: 10.1068/c0648.

Bagchi-Sen, S., Smith, H. L. and Hall, L. (2004) ‘The US Biotechnology Industry: Industry Dynamics and Policy’, *Environment and planning. C, Government & policy*, 22(2), pp. 199–216. doi: 10.1068/c0345.

Castells, M. (1996) *The rise of the network society*. Oxford: Blackwell (Information age vol. 1).

Coenen, L., Moodysson, J. and Asheim, B. T. (2004) ‘Nodes, networks and proximities: on the knowledge dynamics of the Medicon Valley biotech cluster’, *European planning studies*, 12(7), pp. 1003–1018. doi: 10.1080/0965431042000267876.

Jun, S., Park, S. S. and Jang, D. S. (2014) ‘Document clustering method using dimension reduction and support vector clustering to overcome sparseness’, *Expert Systems with Applications*, 41(7), pp. 3204–3212. doi: 10.1016/j.eswa.2013.11.018.

King, A. D. (1990) *Global cities : post-imperialism and the internationalization of London / Anthony D. King*. London: Routledge (International library of sociology).

de la Paz-Marín, M., Campoy-Muñoz, P. and Hervás-Martínez, C. (2012) ‘Non-linear multiclassifier model based on Artificial Intelligence to predict research and development performance in European countries’, *Technological Forecasting and Social Change*, 79(9), pp. 1731–1745. doi: 10.1016/j.techfore.2012.06.001.

Plum, O. and Hassink, R. (2011) ‘Comparing knowledge networking in different knowledge bases in Germany’, *Papers in regional science*, 90(2), pp. 355–371.

doi: 10.1111/j.1435-5957.2011.00362.x.

Ron A. Boschma and Koen Frenken (2006) ‘Why is economic geography not an evolutionary science? Towards an evolutionary economic geography’, *Journal of economic geography*, 6(3), pp. 273–302. doi: 10.1093/jeg/lbi022.

Taylor, P. J. (2001) ‘Specification of the world city network’, *Geographical Analysis*, 33(2), pp. 181–194. doi: 10.1111/j.1538-4632.2001.tb00443.x.

Declaration of Authorship

I, Xudong Zhang, confirm that the work presented in this assessment is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Xudong Zhang

Date of signature: 11^{th}, January 2021

Assessment due date:11^{th}, January 2021, 5pm GMT