IMT School for Advanced Studies, Lucca

Lucca, Italy

Network Analysis of International Patent Data: Technology Cohort, Temporal Dynamics and the Role of Trade

PhD Program in Management Science

XXX Cycle

By

Yuan Gao

2018

The dissertation of Yuan Gao is under evaluation.

Program Coordinator: Prof. Massimo Riccaboni, IMT School for Advanced Studies Lucca

Supervisor: Prof. Massimo Riccaboni, IMT School for Advanced Studies Lucca

Supervisor: Prof. Raja Kali, University of Arkansas

Tutor: Dr. Zhen Zhu, University of Greenwich

The dissertation of Yuan Gao is being reviewed by:

Dr. Luis Correa Da Rocha, University of Greenwich

Prof. Marco Duenas, Universidad de Bogot Jorge Tadeo Lozano

IMT School for Advanced Studies, Lucca

2018

Contents

List of Figures

List of Tables

Acknowledgements

My cordial acknowledgement to the people below:

1. Prof. Massimo Riccaboni for his guidance, advice and coordination during my PhD programme. Prof. Riccaboni is also a co-author of the published work based on Chapter 1 in my thesis[1], and the upcoming publication based on Chapter 2 of my thesis[2].
2. Dr. Zhen Zhu for his tutoring and follow-up through my thesis work. I appreciate Zhen for his time spent giving me valuable advice in research and career. Dr. Zhu is also a co-author of the published work based on Chapter 1 in my thesis, and the upcoming publication based on Chapter 2 of my thesis.
3. Prof. Raja Kali for his advice and support in my thesis work, especially during my visiting period at his department. Prof. Kali is also a co-author of the upcoming publication based on Chapter 2 of my thesis.
4. Dr. Andrea morescalchi for his tutoring and feedback for Chapter 3 of my thesis.

[1] Yuan Gao, Zhen Zhu, and Massimo Riccaboni. Consistency and trends of technological innovations: A network approach to the international patent classification data. In *International Workshop on Complex Networks and their Applications*, pages 744-756. Springer, 2017.

[2] Yuan Gao, Zhen Zhu, Raja Kali, and Massimo Riccaboni. Community evolution in patent networks: Technological change and network dynamics. In *Applied Network Science*, forthcoming.

Vita

2004 Bachelor's degree
Electronics Engineering
Fudan University, Shanghai, China

2014 Master's degree
Industrial Engineering
Purdue University, Indiana, USA

Publications

1. Yuan Gao, Zhen Zhu, and Massimo Riccaboni. Consistency and trends of technological innovations: A network approach to the international patent classification data. In *International Workshop on Complex Networks and their Applications*, pages 744-756. Springer, 2017.
2. Yuan Gao, Zhen Zhu, Raja Kali, and Massimo Riccaboni. Community evolution in patent networks: Technological change and network dynamics. In *Applied Network Science*, forthcoming.

Presentations

1. Yuan Gao, "From Family Members to Followers: A Network Approach to Forward Citations on Patent Family Level," at *Economics PhD Workshop in collaboration with KU Leuven*, Lucca, Italy, 2017.
2. Yuan Gao, "Consistency and Trends of Technological Innovations: A Network Approach to the International Patent Classification Data," at *6th Annual CIRANO-Sam M. Walton College of Business Workshop on Networks in Trade and Finance*, Montreal, Canada, 2017.

Abstract

Access to online electronic database and international patent data harmonization has enabled more researchers to work with patent data. This thesis joins the more recent studies to apply network methodologies to patent data analysis in order to understand the multi-variable interactions in innovation, knowledge flows, and technological trends.

The first chapter focuses on technology cohort in the patent family networks and citation networks to investigate how technologies are integrated in utility invention and the patterns over time. The consistent technological groups found in this analysis are compared to the authority-defined system and provide a more complete coverage.

Chapter 2 focuses on community evolution over time. The stabilized Louvain method is adopted to improve consistency and stability of community detection. The majority mapping algorithm is incorporated for community tracking across time slices. A new method is developed to identify sets of central nodes. A case study of patent filed by applicants in Germany is used to demonstrate and verify the method and the results.

In the last chapter, I apply the methods developed from the first two chapters to the pharmaceutical sector. Motivated by the needs to encourage pharmaceutical R&D and promote globalized innovation, I focus on the effects of being central with a favorable balance in international market on being central in global pharmaceutical R&D collaboration. The descriptive results and a further regression analysis shows positive effects from export "Coreness" on co-invention centrality. Despite causality is not fully resolved, our results show a fun-

damental relationship between knowledge production and trade in pharmaceuticals.

Chapter 1

Introduction

1.1 Background

As a way to protect intellectual property and reward inventions with commercialized value, the patent system is known to generate informative indicators for analysis in technological innovation. The availability of online electronic database and the efforts by worldwide patent authorities to consolidate and harmonize patent data at international level has made it possible for researchers to work with much more accessible patent data. Since the Sixties many researchers have used patent data to measure patent quality, their economic value and possible impact on technological developments and economy. Among the earlier works, Zvi Griliches investigated the relationship of productivity growth to R&D expenditure in industry (Gri98), William S. Comanor, and F. M. Scherer used patent statistics to measure technical change (CS69). Later researchers like Mariagrazia Squicciarini and Hlne Dernis established an extended system of patent quality measures (SDC13a), and the global economic crisis brought up more attention to the role of innovation in economy, such as the analysis by Angela Hausman and Wesley J. Johnston (HJ14).

However, most of the well-recognized conventional indicators are straightforward measures, such as the number of patent applications

and publications, time needed from filing to grant, number of different technology classification codes involved, forward and backward citation counts, etc. Such indicators may be used to track technological changes and innovation, but when considered alone, will fall short due to simplicity and lack of context, resulting in bias and sometimes contradicting conclusions as shown in literature.

From 2000 onwards, more researchers started to apply network methodologies to patent data analysis in order to understand the multi-variable interactions in innovation, knowledge flows, and technological trends. The majority of these studies were based on individual patents, but no investigation has been done from a cohort perspective.

In the light of this, this study takes the initiative to conduct patent data analysis by applying network methods in the context of patent families, defined as a set of patent filings taken in various countries to claim the same priority and to protect the same invention. The patent family dataset represents patents deemed to have higher economic value in the international market. It also eliminates the home bias of regional patent authorities.

The patent database consists of multiple dimensions, including time, geography (location of applicants and inventors) and technology sectors. This thesis constructs networks on different dimensions and implements different methods for community detection, temporal tracking and core identification. The research involves both descriptive and quantitative analysis, and the contributions are both methodological and practical, to the literature of technological innovation.

1.2 Overview

1.2.1 Chapter 1

The first chapter focuses on technology clustering in the patent family networks and citation networks to investigate how technologies from different sectors are integrated in utility invention and the changes of such integration patterns over time. Results of the network approach are

compared to the International Patent Classification system (IPC) which has a authority-defined reference list for each subclass. on the OECD Triadic Patent Family database, this study constructs a cohort network based on the grouping of technologies in the same patent families, as classified by the IPC, and a citation network based on citations between subclasses of patent families citing each other. In both networks, the nodes are the IPC subclasses.

I present a systematic network approach which obtains naturally formed network communities identified using a Lumped Markov Chain method proposed by Carlo Piccardi (Pic11). I define the intra-community closeness weighted by community-level persistence probability as an indicator named coreness to measure the persistent centralness of a node in the network. This indicator not only highlights the most central technologies in each time slice, but also helps solve the problem of tracking unidentified communities continuously over time. Community cores with highest coreness throughout all the years can be traced across time slices, making it possible to investigate two important community characteristics: consistency (sets of nodes that are persistently found in the same community with the cores) and changing trends (sets of nodes that are less persistent, but show a continuous occurrence over time). Compared to the authority-defined IPC references, the consistent node sets from my analysis have a more accurate and complete coverage.

The results are verified against several other metrics, including simple patent counts and patent text similarity from a recent research (ACG18). The proposed method can be generalized as a network-based approach to study endogenous community properties of an exogenously devised classification system. The application of this method may improve accuracy and efficiency of the IPC search platform by providing an improved reference list for the searched subclass. It can also help detect the emergence of new technological focus. However, the changing trends can be fairly fluctuant due to marginality.

1.2.2 Chapter 2

In this chapter, I focus on the patent family cohort network and continue with the efforts of analyzing community evolution over time. I adopt the stabilized Louvain method for network community detection to improve consistency and stability. Thomas Aynaud and Jean-Loup Guillaume proposed this more stable modification to the classic Louvain modularity optimization method by changing the initial network partition of each single time slice to the partition found in the previous time slice (AG10a).

For community tracking, I incorporate the community mapping algorithm from the work by Qinna Wang and Eric Fleury to identify modular overlaps, which are groups of nodes or sub-communities shared by several communities (WF10a). The tracked communities are the ones containing the largest overlap. I have also developed a new method to identify the central nodes based on the temporal evolution of the network structure itself: For each node found in a given community at the initial time slice, the higher occurrence it has in the given community's tracked communities in the following time slices, the more central the node is.

A case study of patent filed by applicants in Germany is used to demonstrate and verify the application of the method and the results. The results are verified against industry literature and robustness checks. Methodologically, our method contributes to the literature of temporal networks analysis with a new approach. In terms of application, it is the first of such in patent data analysis. Compared to the non-network metrics and other conventional network measures, my method has found similar results and is more efficient in showing which nodes are the most central and investigating the evolution of the community containing certain technologies of interest.

1.2.3 Chapter 3

In this chapter, I apply the methods and experience from the first two chapters to find the relationships between the innovation collaboration

network and the world input-output network, and focus on the pharmaceuticals sector, a R&D-driven sector with increasing innovation activities worldwide in the recent decades.

Literature suggests that cross-national knowledge diffusion can boost overall globalization of innovation, and international R&D collaboration in patenting is one of the ways of such diffusion. Researchers have found innovation as a positive factor for firm-level and national exporting (Gre90; Wak98a; And99; RL02). This paper analyzes whether the positive relationship exists the other way around: How does export affect innovation? I construct a cross-national inventor collaboration network based on pharmaceutical patent co-inventions, and use the World Input-Output Database as the international trade network. Through community detection and centrality calculation for both networks, I demonstrate how countries are grouped into communities in either network and how the groupings change over time. Then I obtain a panel data set involving each country's "Coreness" in the co-invention network and the input-output network. Together with other relevant variables, I perform regression analysis. The result shows positive effects of being central with a favorable balance in international market on being central in global pharmaceutical R&D collaboration, but the effects of drug prices remain inconclusive. This analysis may provide evidence for decision makers to use trade strategies to enhance national R&D impact and contribute to globalized innovation.

Chapter 2

Consistency and Trends of Technological Innovations: A Network Approach to the International Patent Classification Data

Abstract

Classifying patents by the technology areas they pertain is important to enable information search and facilitate policy analysis and socio-economic studies. Based on the OECD Triadic Patent Family database, this study constructs a cohort network based on the grouping of IPC subclasses in the same patent families, and a citation network based on citations between subclasses of patent families citing each other. This paper presents a systematic analysis approach which obtains naturally formed network clusters identified using a Lumped Markov Chain method, extracts community keys traceable over time, and investigates two important community characteristics: consistency and changing trends. The results are verified against several other methods, including a recent research mea-

suring patent text similarity. The proposed method contributes to the literature a network-based approach to study the endogenous community properties of an exogenously devised classification system. The application of this method may improve accuracy and efficiency of the IPC search platform and help detect the emergence of new technologies.

2.1 Introduction

As a form of intellectual property, a patent grants the inventor and/or owner of an invention a set of exclusive property rights to legally prevent others from commercially exploiting the invention without the patent owners permission. Patent search allows inventors and entrepreneurs to avoid duplicate efforts and explore promising opportunities, and facilitates researchers and policy-makers to monitor technology development activities. Such searches are based on patent attributes, among which one of the most important is the technology field classification. A good classification system has to categorize patents based on the technologies their pertain accurately and efficiently, with an updated reflection of the technological development trends.

This study integrates the OECD triadic family dataset with individual patent information and citation data, along with additional datasets from national patent authorities to construct a family-cohort network based on patent family grouping, and a citation network based on citation connections across families. Using citation linkages to study scientific and technological landscapes and changes over time is a popular approach for the analysis of both patents and scientific publications(BKB05; BBK08). In our analysis we combine citation-based and patent family data to study the patent landscape over time through the lens of network analysis.

In both networks, the nodes are subclass-level codes of the International Patent Classification (IPC) extracted from each patent's classification information. Assigning a patent with IPC codes that correctly and thoroughly describe its nature is not only important in properly recognizing the novelty values, but also determines the efficiency and costs of patent filing and searching. Our analysis has found that naturally

formed network clusters contain more information than the IPC-defined references can explain. This study also develops a measurement based on the closeness centrality inside a cluster and the cluster's persistence to evaluate each node's "coreness." This indicator has been used to establish a systematic method for community tracking and analysis over time.

2.2 Literature

The IPC system was established out of the Strasbourg Agreement of 1971 to provide a hierarchical technology classification system of language independent symbols for patents and utility models (WIP17a) and is now used by more than 100 countries as the major or only classifying method. The current IPC system is structured in four levels: The top level includes eight "Sections" corresponding to very coarse technical fields, each subdivided into "Classes" at the second level, and then "Subclasses" on the third level, and finally the finest level "Main groups" and "Subgroups." The 2016.01 edition IPC scheme includes 8 Sections, 130 Classes and 639 Subclasses. In this research the subclass codes with the first 4 digits are used as Squicciarini and co-workers have done in their 2013 report on OECD patent quality indicators to measure patent scope(SDC13b).

Despite the wide adoption and continuous updating, the discussion over IPC's limitations due to lack of precision and easiness of use has been going on for decades. Harris et al pointed out that the IPC classifies an invention according to its function whereas the USPC not only classifies based on the function but also on the industry, anticipated use, intended effect, outcome, and structure (HAS10). Lai and Wu proposed a new classification system based on co-citations to replace IPC or USPC and improve categorization accuracy (LW05). When using the IPC system to determine the patentability of an invention, the maximum applicable protection for a filing, or to understand the innovative activities in a certain field, non-patent-experts often face challenges of ensuring accuracy and thoroughness of the search results. The current IPC online platform provides two searching strategies: navigate through the

complete hierarchic IPC scheme, or search for potential matches by keywords using the Catchwords feature. The IPC definitions (WIP16b) contain specific information to facilitate search, including references, classifications notes and indexing codes for hybrid systems. But the massive documentation written in no plain language makes ordinary users prone to mis-classification, under- or over-classification, causing prolonged filing processes and costs for revisions required by the patent authority offices. Users from different application fields have attempted to develop customized search strategies and tools to work with the IPC scheme (Fog07; MJ05; VHAA+12; YRBY05; YP04). Most of them are based on text or images in the patent documents using text mining, bibliometric and semantics methods, and are usually applicable to a specific technological field only.

We introduce a systematic strategy to assist search with semi-automatic suggestions and reduce the reliance on domain knowledge or exhaustive understanding of the IPC scheme itself. Our method does not rely on semantics or manipulation of the patent documents. The network perspective brings additional information not conveyed by the IPC definitions or naive counting.g.

2.3 Data

The data used in this analysis is mainly retrieved from the February, 2016 edition OECD patent database (OEC17b), specifically, the Triadic Patent Families (TPF) database, the REGPAT database, and the Citations database. To complete the information not included in the OECD datasets, we have also referred to data from the U.S. National Bureau of Economic Research (NBER), distributed as a result of Lai and his colleagues' work (LDY+11). The JPO data is not included due to scarce matches with OECD data in the available time period. IPC scheme of edition 2016.01 is used to decode the IPC codes (WIP16a).

The consolidated dataset consists of two parts: the triadic family dataset and the citation dataset. The former is based on OECD TPF database, where a patent family is defined as a set of patents taken at USPTO, EPO,

PCT and JPO that share one or more priorities (OEC09; DK04). A family may contain multiple patents, and each patent may be assigned with multiple IPC codes, some of which might be duplicate among patents. Patent years information comes from different databases, with complexities such as one patent having several priority filings at different times among which an earlier priority was granted later. In this study we consistently define the earliest first priority year among all patents under a family as the family Application Year. The consolidated TPF dataset covers a time span from 1964 to 2014.

The citation dataset combines the OECD Citations database with the consolidated TPF dataset. One citing patent could cite multiple patents, vice versa. Mapping patent and family IDs, we have a master dataset containing pairs of triadic families connected by citations relationships as long as any of the citing family's subordinate patents cites any of the cited family's the subordinate patents.

2.4 Network Construction

2.4.1 Family-Cohort Network

The family-cohort network directly stems from the consolidated TPF dataset. It is a network expression of the way different technology classifications are grouped into the same family. Only the families with more than one unique subclass codes are used, and duplicates within each family are disregarded. It is noteworthy that some earlier classifications might have been redefined or restructured in the current IPC system. The result is a symmetric 639 x 639 matrix in which the value of each element is the number of shared families between the two subclasses indexed by the row number and the column number. The diagonal values are all zeros since only unique subclasses are considered. The sum of all the matrix elements, i.e., the total number of shared families between every two different subclasses, is 9,700,490. In this family-cohort network, the more edges between two nodes, the more different families they belong to are in common.

2.4.2 Citation Network

The citation network represents how IPC subclasses cite each other across patent families. All the IPC subclasses of each citing and cited patent are included. In other words, self-citing is included as found by Hall et al that self-citations are generally more valuable than citations from external patent in terms of market value (HJT05). The network matrix is also 639 x 639, but asymmetric and the element values sum up to 13,211,260, representing the aggregated citation connections from each IPC subclass of every patent in any citing families to each IPC subclass of every patent in any cited families.

2.4.3 Temporal Networks

The 2 networks mentioned above are aggregated from all the available years. in order to capture the changes over time, we split the data by year. The family-cohort dataset is simply divided by the application year of each family. For the citation dataset, it is divided by the cited family application year over a 5-year forward period to consider forward citations within 5 years. Forward citation stands for the citations a patent receives. It's been widely recognized as an indicator of a given patent's technological influences on subsequent technology development, and, to a certain extent, the invention's economic values (THJ97; HSV03; HJT05; SDC13b). The 5-year window is an empirical setting based on the statistical distribution of foward citation quantities over time: the majority (11.62%) of the citations are from patents filed 2 years after the cited patent has been filed, followed by 3 years (11.04%). Also, Squicciarini et al reported that it typically takes 18 months from patent application to publication, and showed that the distribution of forward citation numbers is extremely similar between citation lags of 5 years and 7 years (SDC13b).

In order to build a network with sufficient connections, only the years with above 0.1% of all patent family applications are used: 1978 to 2013.

2.5 Network Analysis Methods

2.5.1 Community Identification

The network analysis is carried out in 3 stages: endogenous community identification, community tracking over time, and the analysis of the structure of the communities.

To identify the naturally formed network clusters, we use the approach proposed by Piccardi (Pic11). In his paper, Piccardi used a measure called Persistence Probability to evaluate the quality of a cluster. The Persistence Probability is calculated from an approximate lumped Markov chain model of the random walker (i.e., a reduced-order Markov chain in which the communities of the original network become nodes) derived from the original high-order Markov chain model. Persistence Probability can be used as a threshold to find out the finest partition in a given network satisfying the desired quality.

In an N–node network corresponding to an N–state Markov chain, let w_{ij} be weight of the edge from node i to node j, then the probability for the random walker to transition from i to j is

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} \tag{2.1}$$

Let π_i be the probability of the random walker being at node i, and C_c be a candidate cluster in the given network, the Persistence Probability of C_c, defined as U_{cc} can be calculated as

$$U_{cc} = \frac{\sum_{i,j \in C_c} \pi_i p_{ij}}{\sum_{i \in C_c} \pi_i} \tag{2.2}$$

The problem of community identification can be formulated as: max q subject to $U_{cc} \geq \alpha$, c = 1,2,...,q, where q is the number of communities in the entire network. When this method is applied to the family-cohort network, nodes in the same α-community (or cluster) with Persistence Probability U_{cc} are IPC subclasses which tend to gather in the same patent family at a certain likelihood. Larger Persistence Probability indicates that the random walker is more likely to circulate inside the community

than visiting another community. In the citation network, nodes in the same cluster represent the IPC subclasses which tend to cite each other at a certain Persistence Probability.

2.5.2 Community Tracking

For each year's network, the clusters are identified by sequential numbers which are not compatible with another year because the sequential numbers cannot be anchored to a reference system such as the IPC scheme titles. To understand how a certain endogenous community changes over time requires a method to attach a non-network tag to it. We propose a method to track an endogenous community by its most representative member, referred to as the "*key*" hereinafter.

To do so, we have developed a measure named *Coreness* as the closeness centrality of a given node (IPC subclass) within community, weighted by the Persistence Probability of its community. Utilizing the distance d_{ij} calculated during community identification, Coreness C_{ic} of node i in community c is formulated as

$$C_{ic} = \frac{1}{\frac{\bar{d_{ij}}n}{n-1}} U_{cc} = \frac{n-1}{\bar{d_{ij}}n} U_{cc}, \tag{2.3}$$

where n is the number of nodes in community c. Coreness is an indicator of how centrally connected a subclass at community level considering the network-level community robustness. It can therefore be used to compare all the nodes in a network, whether they are in the same community or not. We apply this measure to all the temporal networks. For consistency, the networks of all years are divided into the same number of endogenous communities. For each year, all the subclasses are ranked by Coreness from high to low. We then calculate the average ranking over time to find out the ideal subclass key candidates.

2.5.3 Community Characteristics Identification

In this study, we are interested in what remains stable for a key's community throughout all the examined years, and the non-incidental changes,

represented by the disappearance or emergence of certain "trending" subclasses in the community.

To improve robustness, we narrow down the range of community members to the most persistent ones. As the number of communities N increases, Persistence Probability drops and communities break down. More persistent members with stronger connections to the community key are more likely to stay as N increases. Based on the clustering results, we set a stability threshold from N=4 to N=12 to cover the more meaningful network divisions where the minimum U_{cc} is larger than 0.15. For each year, only the subclasses remaining in the focal community for all the Ns from 4 to 12 are considered as persistent. We also calculate persistent members' average Coreness rankings over different Ns to rank them in each year.

For consistency, the threshold is set to an occurrence of or above 80% of all times, i.e. a subclass needs to be found in the interested endogenous community in at least 29 years out of the total span of 36 years. For changing trends, an exploring criterion sets an occurrence of no more than 60% of all times, among which at least 3 years are consecutive. So a subclass must appear in the key's community in 3 to 19 years to be considered as a possible technology trend. These criteria help us reduce the noise to focus on the most noteworthy community characteristics..

2.6 Results

2.6.1 Network Communities

We first examine the networks as an aggregate of all years and all families. Using the community identification approach, for the family-cohort network, the maximum cophenetic correlation coefficient C = 0.75437 is found when time horizon T = 1, meaning that only the neighboring nodes with non-zero similarity can be candidates for the same community members. The largest minimum Persistence Probability is min U_{cc} = 0.32664 for community number q from 2 to 6, and above 0.1 for q$\leq$25. In the citation network, the maximum cophenetic correlation coefficient C

= 0.71717 is found with time horizon T = 1.

When split by year, in some years (e.g. 1978 to 1980) the connections in the citation networks are too weak to form any identifiable communities. In other words, the only community is the entire network. This is largely because the citation network contains fewer distinctive IPC subclasses, given that it is a subset of the family-cohort network as a result of matching with the OECD citation dataset. In the first few years, the proportion of patents with incomplete information and thus lower matching rates is higher.

2.6.2 Traceable Endogenous Community Keys

We use the method described in Section 4.2 to calculate Coreness for each of the 639 IPC subclasses from 1978 to 2013. Each temporal network is divided into 8 communities for easier comparison to the 8 IPC sections. No additional Persistence Probability threshold is applied so that all the participating subclasses are covered. We use the average intra-community Coreness ranking order - rather than the actual Coreness values to find out the keys because the keys are defined as the representative of communities regardless of community size or robustness.

We found that the IPC sections do not form a perfect one-to-one mapping with the endogenous communities. For example, subclass A61K (preparations for medical, dental or toilet purposes) is a key subclass for holding the top closeness centrality position inside its endogenous community, and it is also an ideal representative of Section A for having the highest average Coreness ranking in this Section. But in Section D, even the highest average Coreness ranking is out of 12. To summarize, the IPC section grouping does not well reflect the actual grouping by invention families or citations. An endogenous community formed from real classification assignments may stride over multiple IPC sections, and an IPC section could include subclasses from more than one families. Actually such references to other subclasses within or cross sections can also be found in the IPC definitions, including limiting references and non-limiting references, as explained in the IPC Guide (WIP17b).

We use a set of keys to capture endogenous communities while representing distinctive IPC sections as much as possible. In the following results, A61K will be used as an example as one of the keys from the family-cohort network. Information of the complete keys of both family-cohort and citation networks is available upon request.

2.6.3 Endogenous Communities Characteristics

A more revealing way to present the results is 2D plotting, using the rows to represent the 639 4-digit IPC codes and the columns to represent the 36 years from 1978 to 2013. The rows are arranged according to the IPC index, grouped by sections A to H as labeled along the Y axis on the left. A non-white square at Row i and Column j means the i_{th} subclass is a persistent member in the j_{th} year. The color coding is based on the cross-N average Coreness ranking R_{iN}, where red indicates the highest ranking, i.e. largest cross-N Coreness, and white the lowest.

Figure 1 shows the distribution of the consistent subclasses in the endogenous community centering A61K in the family-cohort network. It can be observed that of all times, the most consistent community members regardless of network clusters number are in Section A and Section C, among which only a few are indicated in the IPC definitions as "references," as labeled out along the Y axis. This shows our results provide richer information than the IPC and can help avoid potential misses in IPC searching.

When looking into the potential trends over time, we found that not all the subclasses meeting the "trend" criteria defined in Section 4.3 imply a disappearing or emerging pattern. Some are just weaker consistent candidates. We argue that it is better to perform the same analysis with inputs from the text similarity network in Figure 2. It should be recognized, though, that thorough interpretation and validation of such "trends" requires domain knowledge and experts' inputs.

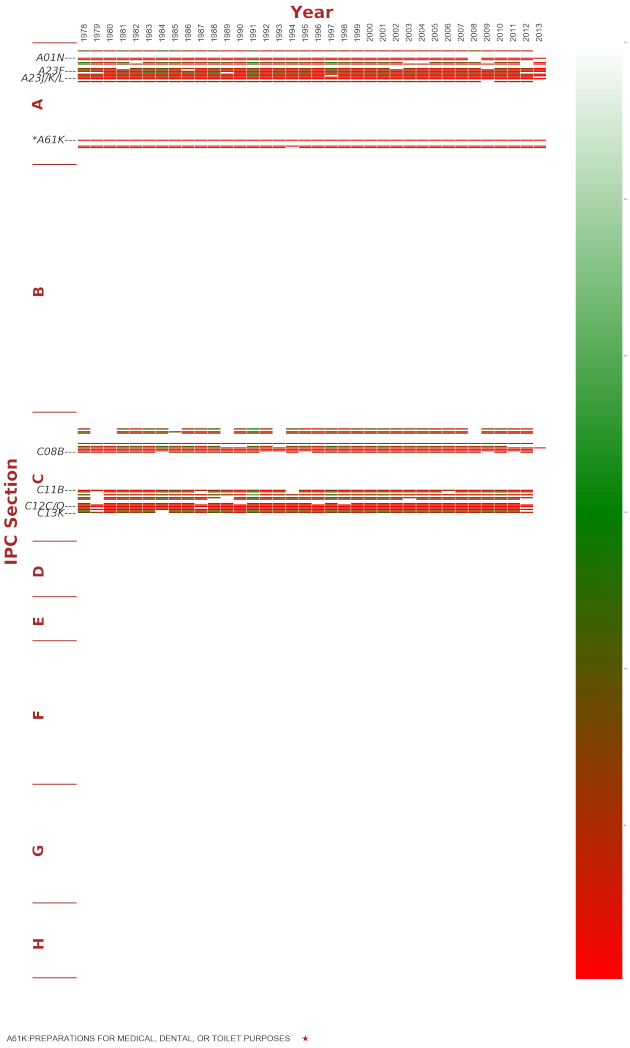


Figure 1: Consistency of the Endogenous Community of A61K Based on Family-Cohort Network

2.7 Verification and Discussion

2.7.1 Verification of Markov Chain Network Clustering

To verify the validity of the Markov Chain network clustering method used in this study, we compare it with the Louvain method for community detection based on greedy modularity optimization (BGLL08). We use the same temporal family-cohort network matrices as described in Section 3.3. In both methods, the nodes with no connections have been removed. To improve stability with the Louvain method, we check the network structure using different resolution levels: 0.5, 0.8, 1, 1.2 and 1.5. As explained by Lambiotte et al (LDB08a), in network partitioning, time can be seen as an intrinsic resolution parameter that affects the scale of network structure.

The result shows that the number of partitions decreases from resolution 0.5 to 1.2, and then increases with some temporal networks at resolution 1.5. Lambiotte et al show a stability framework where the stability of a partition in a continuous-time random walk process is a non-increasing, convex function that goes toward zero when time approaches ∞. It can therefore be inferred that among the 5 tested levels, resolution 1.2 when the average number of partitions is 4.89 may be the closest to the limit where natural clustering cannot be identified going beyond. The result from our method shows some largely overlapping community structures. Moreover, the persistence probability of the entire family-cohort network maintains at 0.32664 for community number N from 4 to 6, and then drops slightly for N = 7 and 8. Therefore the Louvain method would result in similar numbers of partitions as ours.

Using the Markov Chain community identification method we are able to control the results by either specifying Persistence Probability or the community number, a highly desired flexibility in practice. It easily allows obtaining a certain number of communities as needed and helps realize community tracking.

2.7.2 Verification of Consistency Results

To further verify the consistency results is to compare it with the direct counts of member subclasses in the given key's community, a naive measure of how many times a subclass is connected with the key. To avoid being affected by the total application volume variation from year to year, we calculate the average ranking of the connection quantities over the years. Although a simpler and more intuitive measure, the naive counts ranking suffers from a shortcoming due to lack of global normalization. For example, a certain subclass has more connections than another other subclasses with key A, but it might actually have more connections with key B.

As expected, the results are largely similar to the results of our network method, but with differences due to the reason stated above. The naive count results also have just a small portion overlapped with the IPC-defined references, confirming the validity of additional information brought by the network approach.

2.7.3 Verification with Patent Similarity Based on Text Matching

In a most recent study, Arts et al used text matching to measure the technological similarity between individual patents (ACG18). The study applies a text-mining approach to the titles and abstracts of patents to calculate the similarity between any two utility patents. They found that the text-matched patents are significantly more likely to be in the same family and to cite each other.

The "closest match" data shared in their paper contains the closest text-matched patent filed in the same year for each patent. We looked up these patent numbers in the OECD Triadic database and retrieved the IPC subclass codes. Any two subclasses from the two matching patents are assigned with the similarity index. The subclass-level similarities are aggregated for each application year, and averaged by the number of times they are in a pair of closest match of patents. The result is a matrix of 639 x 639 for each year from 1978 to 2004, in which each element is

the average similarity between two IPC subclasses. Thus, similarities between patents are converted to similarities between IPC subclasses. We then apply the Markov chain clustering method to the networks, divide the networks to 8 communities and find out the subclasses which are in the same community with the keys defined in Section 5.2. The results are in Figure 2.

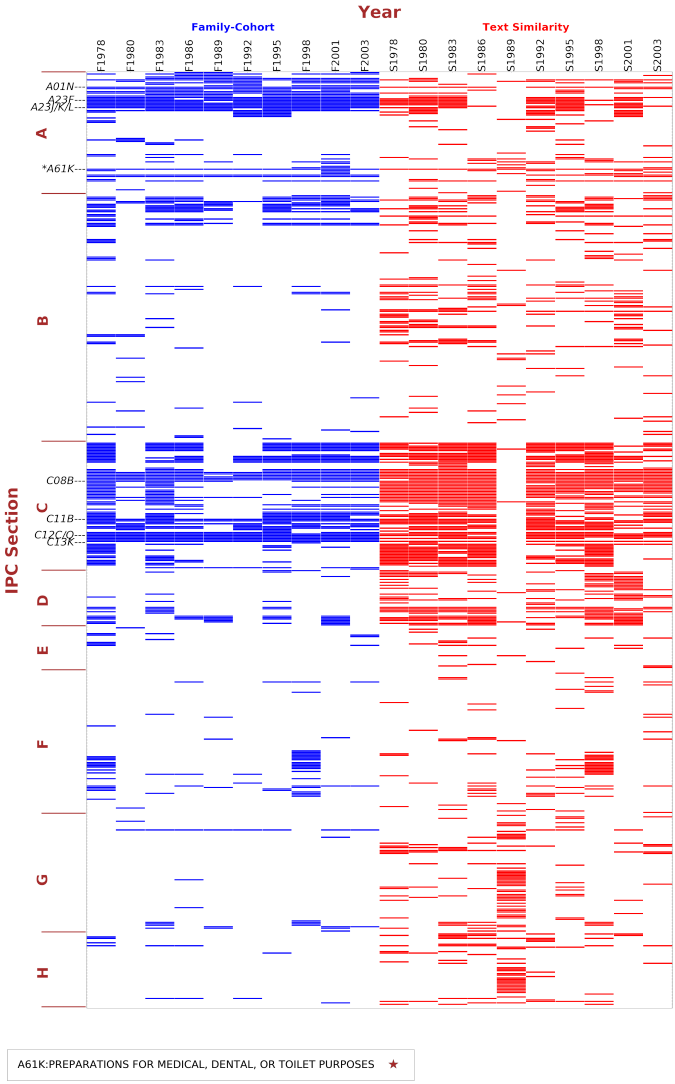


Figure 2: Comparison of Family-Cohort Network and Text-Match Similarity Network for A61K

Comparison with the family-cohort network shows that the most temporally consistent community members are shared between the two networks, while the text-matched network has a sparser distribution. This is likely because the wide coverage of key words makes it more sensitive to changes. But as pointed out by the authors as limitations of their work, patents with little discriminatory power, spelling variants and synonyms, and missing context could all lead to false matching and noises. Classifications assigned by the patent authorities are more consistent. Nevertheless, the text-matching method will be a good reference in technological changes detection.

2.8 Conclusion

To summarize, the systematic method based on a network approach provides an overview of how actual patent classifications are clustered and distributed in terms of family-cohort and citations. Using Coreness as a measurement, the most significant communities and their cores can be identified and tracked over time, enabling the possibility to study the characteristics of these communities, or actually, of the community containing any given subclass of interest. Further filtered communities include only the most persistent members despite community breakdown. The most consistent members with $\geq 80\%$ occurrence of all the available years are partially overlapped with the IPC-defined references, yet more informative. The method can improve the IPC search platform for patent applicants and inventors, analysts interested in technology development, national and regional patent authorities, and the international classification system development.

While the consistency result has been verified using several other approaches, the technological changes result has a great potential in technological innovation trends detection, but needs further work. A refined method is required to highlight the real signals while reducing noises. In addition to the text similarity network, the method proposed by Miranda et al to capture the temporal activities in the form of "pulses" (MDL+17) could be considered for continued efforts in the future.

Chapter 3

Community Evolution in Patent Networks: Technological Change and Network Dynamics

Abstract

When studying patent data as a way to understand innovation and technological change, the conventional indicators might fall short, and categorizing technologies based on the existing classification systems used by patent authorities could cause inaccuracy and misclassification, as shown in literature. Gao et al. (GZR17a) have established a method to analyze patent classes of similar technologies as network communities. In this paper, we adopt the stabilized Louvain method for network community detection to improve consistency and stability. Incorporating the overlapping community mapping algorithm, we also develop a new method to identify the central nodes based on the temporal evolution of the network structure and track the changes of communities over time. A case study of Germany's patent data is used to demonstrate and verify the application of the method and the results. Compared to

the non-network metrics and conventional network measures, we offer a heuristic approach with a dynamic view and more stable results.

3.1 Introduction

Patent data has attracted the interest of researchers as a way to measure and understand innovation and technological change, especially with the increased availability of online electronic database and the efforts made by worldwide patent authorities to consolidate and harmonize patent data at international level (MDW+08b; OEC09).

Gao et al. (GZR17a) have introduced an approach to construct networks based on the OECD Triadic Patent Family database (DK04), to identify communities and the community cores. The comparison against the International Patent Classification (IPC) system (WIP17a; WIP17b) shows that the endogenous communities can provide a more accurate and complete list of potentially associated IPC classes for any given patent class. This association is indicated by being the most consistent nodes in the community containing the given node, as measured by an indicator named coreness. However, that approach was unable to effectively capture the temporal evolution of a community over time due to the difficulty in community tracking.

This paper continues to address this unsolved problem. For community identification, we use an improved Louvain modularity optimization algorithm. To define community cores, we have developed a heuristic approach to detect the central groups of nodes based on the intrinsic characteristics of the temporal networks. As for community tracking, we use a method to find the "best match" based on majority nodes mapping to the reference community. Verification and robustness checks show that our findings are sound and reliable. We also present a case study to demonstrate the real-world implications of our results.

3.2 Background

Since the Sixties patent data has been used by many researchers to measure patent quality, their economic value and possible impact on technological developments and economy (GS63; CS69; Gri98; SDC13b; HJ14). Most of the well-recognized conventional indicators are straightforward measures, such as the number of patent applications and publications, time needed from filing to grant (grant lag), number of different technology classification codes involved (patent scope), forward and backward citation counts, etc. Such indicators may be used to track technological changes and innovation, but when considered alone, will fall short due to their simplicity and lack of context, resulting in bias and sometimes contradicting conclusions (BW08; DM15; HJT01; H$^+$05; HSV03).

In the light of this, we carried out the previous research (GZR17a) to study patent data from a network perspective (AAK16), which lays the foundation for the motivation of this paper. More specifically, two types of networks are constructed based on how individual patents grouped into the same family, and how patents in different families cite each other. In both networks, the nodes are the 4-digit subclass level IPC codes following WIPO's IPC scheme of 2016 (WIP16a). This paper focuses on the former type, the family cohort network, in which any two of the total of 639 nodes are connected when they are both found in patents of the same patent family. The more times two subclasses nodes are found to share the same family, the more intense they are linked in the network. Based on this construction mechanism, a community of closely connected nodes indicates that the represented technological fields are more likely to be found in the same inventions. For example, pharmaceutical products in IPC class A61 and enzymology or microbiology in class C12 frequently co-occur in patent families and they are found to be in the same network community.

Application inventions usually involve more than one technology field. A car, for example, consists of many parts serving different functions. Innovations in molecular material science could stimulate the birth of a new type of tire, or a more efficient type of fuel, which then brings a

new design of engines involving mechanical and electronic innovations. Along the technological trajectories there are many cases like this. To find out how an established community of technological changes over time, splitting up and merging with other technologies, is not only interesting in the retrospective observation of technological development trends, but also helps in understanding the interactions between science and technology and policy making, market drives and other socio-economic factors.

3.3 Data

The dataset used for the analysis is retrieved from the February, 2016 edition OECD patent database (OEC17b). In addition, ISO country codes from the OECD REGPAT database (MDW+08b) are used to sort out patent families by country.

In this paper, a patent family from a country is defined as a family containing at least one patent of which at least one applicant is from that country. The applicant's country is used instead of the inventor's country because the applicants designate the owners or party in control of the invention, mostly firms (OEC09). Therefore it reflects the innovative performance of the given country's firms, while the inventor's country is usually the inventor's professional address.

The REGPAT database is most reliable for OECD and EU countries since it is based on two sources: patent applications to the European Patent Office (EPO) and filed under the Patent Co-operation Treaty (PCT) from 1977 to 2013. We chose to focus on Germany for our case study and we use the data from year 1980 to 2013 for more consistent data quality. Germany has the largest number of patent applications among all the EU countries, and ranks third for patent production among the OECD countries.

3.4 Methodology

The analysis mainly consists of three parts, to be described in the following paragraphs: community identification, central nodes identification and community tracking over time.

3.4.1 Community Identification

In our previous study (GZR17a), we used the Lumped Markov Chain method proposed by Carlo Piccardi (Pic11) to detect clusters in networks. This method produces satisfying results for a single static network with sufficiently strong clustering structure. However, for our purpose to analyze the temporal evolution of a network, essentially a network in multiple time slices, this method would treat each time slice as a separate network without connection to each other, which is not appropriate for the continuous technological development issue of interest. Also, the marginal results observed show that the detected community structure is very sensitive to the input network. In other words, although the network is not supposed to have dramatic change from one snapshot in time to the next, a small change could cause significant transformation in the resulting communities.

To better capture the network's temporal properties and overcome the instability, we use a modification of the Louvain modularity optimization method for community detection. This modification, namely the Stabilized Louvain Method, proposed by Aynaud and Guillaume (AG10b), has been proved to achieve more stable results in tracing communities over time. The Louvain method finds the community structure with maximum modularity by looking for modularity gain through iterations (BGLL08). The modification, essentially, is to change the initial partition of the network at time *t* to the detected partition at time *t-1*, thus the initial partition is constrained to take into account the communities found at the previous time steps, making it possible to identify the real trends.

The algorithm implementation is based on the Python module using NetworkX for community detection (Ayn09). We split up the database

by the earliest priority year of patent family, and execute the algorithm for each year, using the detected network partition as the initial partition for the next year.

3.4.2 Central Nodes Identification

There are many different ways to define centrality within a community and/or a network, from the classic definitions by degree, betweenness, closeness, Eigenvector, PageRank, etc, to many customized concepts in empirical and theoretical researches (Fre78; WF94; VCLC08). For example, in our previous work (GZR17a) "Coreness" has been defined as a measure of weighted centrality, based on the probability to be present in the community and the intra-community centrality of each node. However, similar to community detection, centrality measures are not designed for temporal evolving networks and the adoption of one metric out of the others is usually an ad-hoc choice.

A more heuristic concept of cores, as defined by Seifi and colleagues, is certain sets of nodes that different community detection algorithms or multiple execution of a non-deterministic algorithm would agree on (SJR+13). They summarized that for a static network, there are two types of algorithms to identify such sets of nodes: by adding perturbations to the network, and by changing the initial configuration. In the first type, small perturbations such as removing a fraction of links and putting them back on random pairs of nodes, are used to create slightly different networks from the original and produce different partitioning results for comparison and finding of the consensus communities. However, for a network that changes over time, such perturbations naturally exist in each time slice. In fact, they are the temporal changes to be discovered. Therefore, the latter type is more appropriate. Wang and Fleury experimented with the overlapping community technique in a series of works (WF10b; Wan12; WF13). Our method is similar to the concept of Wang and Fleury's fuzzy detection method to identify modular overlaps, which are groups of nodes or sub-communities shared by several communities (Wan12), with a different implementation.

We describe the overlapping community mapping algorithm and the central nodes identification methods in a 4-step procedure:

I. Given the network partition P in the reference time slice t, identify a community C ($C \in P$). C is the target community of interest to be mapped to in the following time slices.

P is obtained using the Stabilized Louvain Method described in the previous subsection. For a network with the total set of nodes E, P = $\{C_1, C_2, ...C_k\}$, where:

$$\bigcup_i C_i = E, i \neq j \Rightarrow C_i \cap C_j = 0$$

II. In the network with partition P' of a following time slice t', find the community with the most nodes in C, and that is the mapped community C' of C. The change of C from t to t' is considered the change between C and C'. This step can be illustrated by the pseudo codes in Algorithm 1:

Algorithm 1 Pseudo-code of Community Mapping

```
C ← Nodes of the community to be mapped in time slice t
P' ← Partition of the network in time slice t'
C'_i ← The i_th community in P'
N_i ← C ∩ C'_i
C' ← N_1
loop:
if N_i ≤ N_{i+1} then
    i ← i + 1.
    C' ← N_{i+1}.
goto loop.
close;
```

III. Based on the communities detected in the previous step, take any node k, find the community C_0 it belongs to in the initial year T_0 in a certain time window of n years, and use the mapping algorithm to track C_0 in the following years within the time window.

VI. The more significant this node k is, the more likely it is to be found in the mapped communities. Each node will have a number W_k ($W_k \leq n$) of how many times it is included in the mapped communities throughout the time window. The group of nodes with the largest W_k will become the central sets in this time window. Step III and VI can be illustrated by the pseudo codes in Algorithm 2:

Algorithm 2 Pseudo-code of Central Nodes Identification

```
N ← length of the time window started with the initial year T_0
m ← total number of nodes in the network
k ← 1
loop1:
if k ≤ m then
    C_k0 ← the community containing the k_th node in the initial year T_0
    W_k ← 1
    loop2:
    j ← 1
    if j < N then
        C_kj ← the community mapped to C_k0 in the following year T_j
        if k ∈ C_kj then
            W_k ← W_k+1
    goto loop2.
goto loop1.
close;
```

This method uses the intrinsic temporal dynamics of the network to find the central nodes. It is intuitive and heuristic, independent of arbitrary ad-hoc choices of measures. The configuration of the initial year and the length of time window could significantly affect the results. Therefore, robustness checks using different lengths of rolling time windows are necessary to verify stability.

3.4.3 Community Tracking Over Time

After sets of central nodes are identified, it is then possible to track the community containing them through the years. The tracking method is

the same as the mapping algorithm described above. Visualization helps to show that the central nodes are the persistent "cores" of the community under tracking whereas the "peripheral" nodes reflect the changes over time.

3.5 Case Study

Using the 3-step method described above, we perform a case study using data of patent families with Germany as the applicant's country. As the largest economy of the EU, Germany also ranks top among all the EU countries in terms of IP filings, including patent applications. Data from the World Intellectual Property Organization (WIPO) (sd17) shows that 176,693 patents have been filed to Germany's patent office in 2016 from residents and abroad, more than twice of 71,276 from France, the second place in EU. WIPO's statistics also reports that the top 5 fields of technology associated with patent applications are transport; electrical machinery, apparatus; mechanical elements; engines, pumps, turbines; and measurement.

Analysis Configuration

In our method, there are several adjustable parameters:

- *Community Detection Resolution*. In the first step, the Louvain method allows for different resolution settings, an implementation of the idea raised by Lambiotte and colleagues that time plays the role of an intrinsic parameter to uncover community structures at different resolutions (LDB08b). To test the influence of resolution, we run community detection using different resolutions ranging from 0.5 to 2.

- *Overlapping Community Reference*. In the second step, there are two ways to choose the reference year: For any year T_t of the non-initial years in the time window, always refer to the initial year T_0, or refer to the previous year T_{t-1}. The latter would mediate the de-

pendency of the initial year. We have applied both types of time referencing and compared the results.

- *Time Window Setting*. As mentioned in the previous section, the initial year's network partition is used as the reference for the following years' community mapping. The time window length is important for two reasons: first, depending on the pace of technology development and potential events driving the changes, the period of time that the initial year would remain valid as the reference varies; and second, longer time windows would require a node to be more "central" to appear at all time or most of the time, and therefore would result in smaller sets of central nodes than shorter time windows. To address these concerns, we used different rolling window settings, including 5 or 10-year time windows with the initial year rolling from year to year (for example, 1980-1989, 1981-1990, ...), and 5 or 10-year time windows with the initial year rolling 5 years apart (for example, 1980-1989, 1985-1994, ...).

Results

Community detection - quantities and sizes: For community detection, we apply the stabilized Louvain method on the entire time range from 1980 to 2013 because technological development is continuous through all the years.

We first check the number of communities detected at different resolution levels. As each node represents a subclass in the IPC scheme, not all of them would appear in every year's patent applications. In addition, some patent families contain just a single subclass. Such cases would result in "orphan communities", communities that have only one node without connection to any other nodes. There are also some very small communities with 2 or 3 nodes. Figure 10 shows the community structure of selected years with resolution set to 1.0, including all the small communities and orphans with nodes layout using Fruchterman-Reingold force-directed algorithm (HSSC08; FR91). Each sample year has an average of 151 orphan nodes plus 7 nodes in small communities

with no more than 5 nodes. So many isolated nodes and small communities will cause too much noise in the analysis. To focus on the meaningful clusters, we have excluded all the communities with 5 nodes or less from the detected partitions. Quantities of the remaining communities are shown in Figure 3.

Contrary to the common wisdom that higher resolutions correspond to finer, and therefore more partitions, the figure shows that after excluding the very small communities, the lowest resolution 0.5 has the most communities in all the years, and resolutions 1.8 and 2.0 have the fewest. Figure 3 also shows that the community numbers generally have a decreasing trend over the years. This is due to the mechanism of the stabilized algorithm where each year's initial partition builds on the previous year. With the enhanced stability, it becomes easier to identify clusters with time. It is noteworthy that the decrease of number of non-tiny communities over time does not indicate the breakdown of weakly connected communities, but rather community merging, including the situation where a community splits into 2 or more smaller parts which merge into other large communities.

Likewise, one should be aware that the disappearance of a portion of nodes in a community does not mean such nodes abruptly disconnect from the central nodes of the community. They are most likely still connected, but have become more closely connected with another set of central nodes, or are replaced by other nodes that are closer to the original central nodes. The methodology of cluster identification involves such "competition" at all times.

We also check the community sizes. Figure 4 shows the average number of nodes in community for all the years at different resolutions. Overall, the community size increases with resolution, and from the earlier years to the more recent years.

The first-step results show that although the algorithm detects more, finer communities under higher resolutions, a lot of them are very small communities. As a result, at the higher resolutions the community size distribution tends to be more polarized, with fewer but more aggregated communities, and more tiny communities than at the lower resolutions.

Central nodes - occurring rate: Similarly, different resolutions would result in different sets of central nodes. We define an indicator named "occurring rate" as the number of occurrence of each node in the mapped communities, divided by the total number of years in the time window. For a year-to-year rolling time window setting, the average occurring rate of all the nodes over a certain time window is calculated as

$$O_r = \frac{\sum_{t=1}^{34-N+1}(\frac{\sum_{i=1}^{m}(\frac{n_{ir}}{N})}{m})}{34-N+1}, r \in \{0.2, 0.5, 1.0, 1.2, 1.5, 1.8, 2.0\}, m \leq 639 \tag{3.1}$$

where m is the total number of nodes in year t after excluding those very small communities with less than 5 nodes; n_{ir} is the occurrence of the i_{th} node at resolution r in each time window during the community mapping process including the initial year; and N is the length of the time window.

We use the configuration of 10-year windows rolling from year to year to demonstrate this result. When N is equal to 10 in Equation 1, the calculated mean values and standard deviations of the occurring rates at various resolutions are shown in Figure 5. Using both mapping algorithms, the lowest average occurring rates are at resolution 1.0.

While there is no benchmark for the absolutely ground truth to determine which resolution is the "best", for our analysis purpose there are some preferred qualities: lower average occurring rates are more desirable because such community structures can better reflect the changes over time: Figure 3 and 4 show that the higher resolutions generate fewer and larger communities, which indicates that the community sizes tend to polarize at higher resolutions, with fewer large communities and more tiny communities, or even disconnected single-node communities.

At lower resolutions, the number of communities larger than five increases, which might also bring more instability (the number of distinct communities decreases from 13 to 8 at resolution 0.5). Therefore, we choose resolution 1.0 as the setting for the next step, to identify communities and track them over time.

The statistical behavior shown above under different resolutions is

related to the problem known as "resolution limit" (FB07), that the modularity optimization method may fail to identify communities smaller than a certain scale. Lambiotte and colleagues have also verified in their framework that partitions beyond a certain resolution limit are obtained at small time where the optimal partition is the finest (LDB08b).

Central nodes at resolution 1.0: Using the overlapping algorithm, at resolution 1.0, we select different time window configurations to identify the central nodes, each with the two referencing methods described above. Figure 6 shows the central nodes plotting under the 10-year time window setting, rolling from year to year. The threshold of the central nodes is set to be the length of the time windows (34 for the all-year setting and 10 for the rolling windows). That is, only the most persistent nodes with an occurring rate of one within the time window are colored in the figure. So under the all-year setting there are fewer central nodes. If a node is central using both referencing methods (colored green), it is more likely that the initial community has not gone through significant reshuffling. If it is only central when referring to the initial year (colored red), then in at least one of the following years in the time window, the initial community has probably experienced some changes that are not in a consistent direction. For example, when merging and then splitting, by referring to the previous year a node might be left out in the minority part of the merged community. If a node is central only when referring to the previous year, it is likely that it has just drift away from the initial community during accumulated changes. For some nodes, they would become red first, and then turn to green. This means the changes have stabilized.

Figure 6 shows several noteworthy trends, highlighted as framed areas 1-4. However, at this moment it is too soon to relate these signals with real-world facts since it is not yet clear how central nodes are grouped into different communities. At this stage, the visualization provides a guidance for the potential trends to take a closer look at. Overall, it also shows the most persistent central nodes, such as IPC Class C07-C08 (organic chemistry and organic macromolecular compounds), and H03-H04

(electric circuitry and electric communication technique).

Community tracking: At this step, any chosen community in the initial year can be tracked to analyze its changes over time. We use two examples to illustrate our approach. Since the endogenous communities do not have meaningful names, we refer to them by one of the representative central nodes they contain in year 1980's partition: B01D, defined in IPC as "separation in physical or chemical processes"; and B60R, "vehicles, vehicle fittings, or vehicle parts" not provided for in other categories under class B60, "vehicles in general".

Figure 7 and 8 show the results of tracking the two communities above, respectively. Both communities cover multiple IPC sections, as discussed by Gao et al. (GZR17a). The two figures show that the consistently overlapping parts of the two communities are different most of the time. B01D's community mainly consists of various physical or chemical processes treating materials and tooling (classes B01-B06), artificial materials from glass to cement and ceramics (C02-C03), petrol and gas industries (C10), and metallurgy and metal surface treatment (C21-C23). Such a composition suggests the application of physical or chemical processing techniques in the inventions of certain industries. For B60R, its community covers the majority of Section B, E and classes F01-F17, a combination of machinery, mechanical engineering, vehicles and transportation, building and construction. Relating to Figure 6, the central nodes of these two communities contribute to a majority of the central nodes, including the framed areas 1 and 2. This is consistent with the WIPO statistics about Germany's top technology fields of patent applications (for the complete IPC definitions, please refer to WIPO's IPC Scheme (WIP16a)).

Next, we focus on the major differences between the two figures. From 1990 to 1999, B21-B30 "moves" from Figure 8 to Figure 7. Those subclasses focus on technologies related to metal working, machine tools, and hand tools, which are likely to be applied in both communities. The temporary "move" turns back after 1999. This is an example of marginal clustering. Another similar case is the "move" of classes F22-F25 from

Figure 7 to Figure 8 from 2000 to 2009. This part represents technologies related to combustion process, heating and refrigeration. After robustness check, the "moves" still exist. This indicates that instead of an artifact due to time window configuration, the "moving" technologies are closely connected to both communities and the network clustering algorithm captures the changes in the relative connectivity. These two "moving" parts also provide an explanation to the framed areas 1 and 3 in Figure 6: the temporary community switches may result in the rise and fall of a set of central nodes in the following or preceding rolling time windows.

In Figure 8, we should also notice the spread to Section G and H starting from the 1990s. This is a consistent trend, getting stronger in the last 4 years. Compared to Figure 7 which also covers a part of Section G, the community containing B60R incorporates more technologies in digital computer (class G06), electric devices and power supply and distribution (class H01 and H02). This observation is in line with WIPO's report of electrical machinery as the second top technology field of patent applications (sd17).

Discussion

Technological change in Germany's automotive industry: To make sense of data analysis findings based on real-world technological trends is always difficult. In most empirical analysis, reliable methods and domain knowledge in the industry are both essential.

In Figure 8, the technology community containing B60R takes up more than half of Germany's patent filing activities, with the most persistent parts being IPC classes B60-B67, Section E, and F1-F16. These technologies can be considered as the mainstream of this community: vehicles and transportation, building and construction, machine and engines (for the details of these IPC schemes, please refer to Table 1 in Appendix).

Clear changing trends can also be observed. Aside from the marginal "moves" like B21-B23 as discussed above, we focus on the more consistent trends, such as the increasing involvement of Section G and H, specifically, classes G01, G05, G06, H01 and H02, shown in the bottom

framed area in Figure 8. These are the technologies related to measuring and testing, controlling and regulating, computing, electric elements, and electric power. This trend started from 2000, and became significantly stronger since 2010 (for the relevant IPC schemes of the classes and the subordinate central subclasses, please refer to Table 2 in Appendix).

Germany's dominating industrial sectors include automotive, machinery and equipment, electrical and electronic, and chemical engineering. These sectors not only contribute to the national GDP, but also are the focal points of innovation of this country. Among the top ten German organizations filing the most PCT patents, at least 6 have automotive as its major or one of the major operations, including vehicle manufacturers like Continental Automotive GMBH and Audi AG, automotive components and assembly suppliers like Robert Bosch Corporation and Schaeffler Technologies AG & Co. KG, and research institutes like the Fraunhofer Society (sd17). Germany Trade & Invest (GTAI), the economic development agency of the Federal Republic of Germany reported that internal combustion engine energy efficiency, alternative drive technologies (including electric, hybrid, and fuel cell cars), and adapting lightweight materials and electronics are the current major market trends (GTA17). From electronic technologies, software solutions to metallurgy, chemical engineering, automation and drive technologies, innovation in the automotive industry drives and benefits from a number of other sectors.

In fact, these trends in the automotive sector are not limited to Germany, but Germany's case is more noticeable and representative given its outstanding concentration of R&D, design, supply, manufacturing and assembly facilities. The automotive industry does not just source from other sectors for innovative technological support. When Enkel and Gassmann examined 25 cases of cross-industry innovation, automotive is observed as both the result and source of the original idea (EG10). The interactive sectors range from the ones with a closer cognitive distance like aviation and steel industry to the more distant ones like sports, medical care and games. These cases all occurred between 2005 to 2009, and indeed, the cross-industry technological interactions have become more

dynamic starting from 2000, as the shuffles observed in Figure 6-8 of our analysis. In 2009, Germany's Federal Ministry for Environment, Nature Conservation, Building and Nuclear Safety issued German Federal Governments National Electromobility Development Plan (Bun09) specified a serial action plan to promote electromobility in Germany, which defines 2009 to 2011 for market preparation, 2011 to 2016 as market escalation and 2017 to 2020 as mass market. The first stage focuses on research and development. The Plan also identifies batteries as the weakness of Germany's automotive sector on the path to the leading position in electromobility. The increased activities in Section G and H starting from 2000 might be a reflection of this policy. However, this is up to validation when more data covering the following years will become available.

Robustness check: For community tracking, we have performed the analysis under 10-year and 5-year time window settings, and found the results to be very close. The results presented in Figure 7 and 8 are based on 5-year time windows. In addition, we have done robustness checks using the other community mapping method and with time window shifts, shown in Figure 11 and 12 respectively using the example of B60R's community. In Figure 11, when referring to the initial year, the colored blocks layout is the same as Figure 8 except for the colors used, which is merely due to the difference in the definition in the mapping methods. We find similar results in Figure 12: There is no difference from Figure 8 except for the 1-year shift. We have performed such robustness checks for other communities and obtained the similar results. This indicates that the community mapping method is stable and consistent in identifying central nodes and tracking communities.

Central nodes identification methods comparison: Alternative to the community mapping and central nodes identification method, we try to rank nodes by their betweenness centrality. Betweenness centrality is one of the most widely used measures of vertex centrality in a network (Bav48; Bea65; Fre77). Compared to other centrality measures using degree or closeness, betweenness represents the connectivity of a node

as a bridge connecting two other nodes along a shortest path. We use it as an example to demonstrate the similarity and difference between our method and the conventional network centrality measures. The betweenness centrality of each node is calculated to find the nodes with the highest centrality values. In order to avoid outstanding impact from a single year, we use aggregated data from 3 consecutive years to form a network, based on which the centrality is calculated for the first of the 3 years. We present here the results comparison for the same years from 1980 to 2004. As a major difference between the two algorithms, ours provides a set of central nodes all with the same occurring rate of 1, but the betweenness centrality value of most nodes are different, ranking them from high to low. So when using the betweenness centrality method, we take the M nodes ranking highest by centrality values, with M being the size of central nodes set in the same time period using our method. For example, the central nodes set in the time windows starting with 1980 has 131 nodes, and the top 131 nodes with highest betweenness centrality rankings in the aggregated period of 1980-1982 are used for comparison. The matching rates are shown in Table 3 in Appendix, averaging at 32.45%. Figure 9 shows the distribution over IPC scheme using both methods. The central nodes based on our algorithm are the ones shared by both referring methods.

The two algorithms are different by definition, and offer different information as the comparison shows. It is difficult to verify the results against ground truth, but we argue that our method has two important advantages. First, there is no arbitrary control of the number of central nodes. To study the interaction of technologies in cohesive families, to have a set of central nodes rather than a given number of top centrality nodes is intuitively closer to the real-world situation. Second, our method identifies the set of central nodes based on tracked communities over a time window, while the betweenness centrality calculated is for a single time period (3 years in the demonstrated example) - additional efforts are needed to track communities over time in order to calculate the centrality values for continuous time periods. It would only be more inaccurate to simply aggregate data in a time window of 10 years

and calculate the centrality. These issues stand true for all other centrality measure. Figure 9 also shows the central nodes identified using our proposed method are more consistent and concentrated, while the top betweenness centrality nodes are more spread out over the whole IPC scheme.

Similarities between the two results also confirm the persistent and changing trends shown in Figure 6: bio-technology in agriculture and food (A01, A21, A23), chemical technology in medical science and pharmaceutics (A61), material separation and other processing (B01), machine tools (B23), Vehicles and transport (B60-B65), organic chemistry (C7-C9), biochemistry (C12), engine technology (F1-F16), physics measuring, testing, computing and controlling (G01, G05 and G06) and electronic technology (H01-H04) are more persistent. And increasing centrality is found with B21-B23, B60-B61,C12-C13, H03-H04.

Comparison with multislice community detection and tracking method: To study networks that evolve over time, another methodology is to treat the changing network as slices at different points in time based on quality functions. Mucha et al. (MRM$^+$10) proposed a method to generalize the problem of network community structure detection using interslice coupling adjacency matrices consisting of coupling parameters between nodes in different slices. The generalized algorithm offers flexible configurations for both the resolution parameters as we used in the Louvain modularity clustering algorithm, and the interslice coupling parameter indicating connection among slices under Laplacian dynamics. This solution is applicable to the multiplex community detection task we have. To compare the results, we applied the algorithm proposed by Mucha et al. in the same 5-year time windows as shown in Figure 8 and 9, with the same resolution set to 1.0 and the coupling parameter as 1.0. and then find the communities containing subclasses B01D and B60R, respectively. The algorithm also obtains clusters based on modularity optimization, and generates a considerable amount of very small communities. Same for the orphan nodes. Therefore, communities with 5 nodes or less are also excluded in the results for comparison, as shown in Figure 13 and

14.

Comparing Figure 13 with Figure 7, and Figure 14 with Figure 8, obvious similarities can be observed. The "move" of B21-B30 from 1990 to 1999 is not shown for B01D. But from 1995 to 1999, most nodes in this section drop out for B06R, although they did not "move" to the community containing B01D. It verifies the marginality of this section, that they tend to have close connection to several different communities.

As mentioned before, it's hard to determine the result of which algorithm is closer to the truth. Each method has its unique properties. The algorithm by Mucha et al. has the advantage of providing an overall picture of all the communities and their changes over time, but we have found that as the continuous time period increases, the number of clusters detected will decrease, which reduces the sensitivity to changes. When applied on shorter time period, the 2 methods have 2 steps in common: communities identification and tracking. For the first step, we argue that our method has higher stability and consistency given the Stabilized Louvain Method. Additionally, our method is capable to find the central nodes of a community, which is meaningful in the situation of this study.

Comparison with conventional patent metrics: Compared to the simpler, more straightforward metric used in conventional patent data analysis, the network approach is more complicated and costs more computational resources. However, we propose the network method for its advantage in studying the structure of an inter-connected system. In (GZR17a), the authors showed that ranking nodes by their connections with a given "key" subclass produced different results than the network clustering method, although largely similar. In the network perspective, nodes are clustered based on their relative proximity instead of the absolute counts or frequencies. Consider the situation where a node k is connected to nodes in 2 clusters A and B, where A has more nodes than B and therefore gives N more occurrence/connections. A simple measure will put k as a key node in A, but the network algorithm might attribute k to B if there are other nodes in A with even stronger connections to each

other.

Secondly, some structural changes may be anticipated by economic historians and policy makers as results from known actions or decisions, but they don't usually roll out as expected, with likely differences in timing or extent. As compared to traditional methods, our approach is better suited to detect structural change and paradigmatic shifts in the technological landscape.

3.6 Conclusion

Through the three-step procedure, we demonstrated a way to improve community detection for temporal evolving networks, and more importantly, to track the community changes over time. Using Germany as a case study, we have verified this procedure by combining industry literature and robustness checks. Methodologically, our method contributes to the literature of temporal networks analysis with a new approach. Comparisons with conventional methods have helped to prove its validity and advantages. In terms of application, it is the first of such in patent data analysis. Although the subject of interest here is technological evolution, we expect the proposed approach to become a powerful tool for studying similar systems.

3.7 Limitations and future Work

We focus our analysis on selected technological fields. Neither Figure 6 nor Figure 9 distinguishes the central nodes by communities. It is because the communities are not exogenously defined, and to track all the communities requires selection of a node in each community in the initial year. In fact, none of the methods discussed can show how all the communities change over time in one picture with satisfying accuracy, sensitivity and stability. Our method is more efficient in showing which nodes are the most central and investigating the evolution of the community containing certain technologies of interest.

Given that the method utilizes the information embedded in the network changes, it can be generalized for other temporal networks studies. However, the result verification still requires more work due to the reasons mentioned in the Discussion section. A next step in our research is the application in other countries or regions to expose the method to a more comprehensive check. This will also provide an opportunity to study how various factors, including policy decisions, market trends, economic growths, national or regional resources, human resources, government and business investment, would interact with technological exploration.

3.8 Appendix

Table 1: IPC Schemes of the Persistent Technologies in Community Containing B60R

Section	Class	Scheme
B	B60	VEHICLES IN GENERAL
	B61	RAILWAYS
	B62	LAND VEHICLES FOR TRAVELLING OTHERWISE THAN ON RAILS
	B63	SHIPS OR OTHER WATERBORNE VESSELS; RELATED EQUIPMENT
	B64	AIRCRAFT; AVIATION; COSMONAUTICS
	B65	CONVEYING; PACKING; STORING; HANDLING THIN OR FILAMENTARY MATERIAL
	B66	HOISTING; LIFTING; HAULING
	B67	OPENING OR CLOSING BOTTLES, JARS OR SIMILAR CONTAINERS; LIQUID HANDLING
E	E01	CONSTRUCTION OF ROADS, RAILWAYS, OR BRIDGES
	E02	HYDRAULIC ENGINEERING; FOUNDATIONS; SOIL-SHIFTING
	E03	WATER SUPPLY; SEWERAGE
	E04	BUILDING
	E05	LOCKS; KEYS; WINDOW OR DOOR FITTINGS; SAFES
	E06	DOORS, WINDOWS, SHUTTERS, OR ROLLER BLINDS, IN GENERAL; LADDERS
	E21	EARTH OR ROCK DRILLING; MINING
	E99	SUBJECT MATTER NOT OTHERWISE PROVIDED FOR IN THIS SECTION
F	F01	MACHINES OR ENGINES IN GENERAL; ENGINE PLANTS IN GENERAL; STEAM ENGINES
	F02	COMBUSTION ENGINES; HOT-GAS OR COMBUSTION-PRODUCT ENGINE PLANTS
	F03	MACHINES OR ENGINES FOR LIQUIDS; WIND, SPRING, OR WEIGHT MOTORS; PRODUCING MECHANICAL POWER OR A REACTIVE PROPULSIVE THRUST, NOT OTHERWISE PROVIDED FOR
	F04	POSITIVE-DISPLACEMENT MACHINES FOR LIQUIDS; PUMPS FOR LIQUIDS OR ELASTIC FLUIDS
	F15	FLUID-PRESSURE ACTUATORS; HYDRAULICS OR PNEUMATICS IN GENERAL
	F16	ENGINEERING ELEMENTS OR UNITS; GENERAL MEASURES FOR PRODUCING AND MAINTAINING EFFECTIVE FUNCTIONING OF MACHINES OR INSTALLATIONS; THERMAL INSULATION IN GENERAL
	F17	STORING OR DISTRIBUTING GASES OR LIQUIDS

Table 2: IPC Schemes of the Central Nodes in Section G and H in Community Containing B60R

Code	Scheme
G01	***Measuring; Testing***
G01F	MEASURING VOLUME, VOLUME FLOW, MASS FLOW, OR LIQUID LEVEL; METERING BY VOLUME
G01G	WEIGHING
G01H	MEASUREMENT OF MECHANICAL VIBRATIONS OR ULTRASONIC, SONIC OR INFRASONIC WAVES
G01L	MEASURING FORCE, STRESS, TORQUE, WORK, MECHANICAL POWER, MECHANICAL EFFICIENCY, OR FLUID PRESSURE
G01M	TESTING STATIC OR DYNAMIC BALANCE OF MACHINES OR STRUCTURES; TESTING OF STRUCTURES OR APPARATUS, NOT OTHERWISE PROVIDED FOR
G01P	MEASURING LINEAR OR ANGULAR SPEED, ACCELERATION, DECELERATION OR SHOCK; INDICATING PRESENCE OR ABSENCE OF MOVEMENT; INDICATING DIRECTION OF MOVEMENT
G01W	METEOROLOGY
G05	***Controlling; Regulating***
G05B	CONTROL OR REGULATING SYSTEMS IN GENERAL; FUNCTIONAL ELEMENTS OF SUCH SYSTEMS; MONITORING OR TESTING ARRANGEMENTS FOR SUCH SYSTEMS OR ELEMENTS
G05D	SYSTEMS FOR CONTROLLING OR REGULATING NON-ELECTRIC VARIABLES
G05G	CONTROL DEVICES OR SYSTEMS INSOFAR AS CHARACTERISED BY MECHANICAL FEATURES ONLY
G06	***Computing; Calculating; Couting***
G06M	COUNTING MECHANISMS; COUNTING OF OBJECTS NOT OTHERWISE PROVIDED FOR
G08	***Signalling***
G08G	TRAFFIC CONTROL SYSTEMS
G10	***Musical instruments; Acoustics***
G10K	SOUND-PRODUCING DEVICES; METHODS OR DEVICES FOR PROTECTING AGAINST, OR FOR DAMPING, NOISE OR OTHER ACOUSTIC WAVES IN GENERAL
H01	***Basic electric elements***
H01H	ELECTRIC SWITCHES; RELAYS; SELECTORS; EMERGENCY PROTECTIVE DEVICES
H02	***Generation, conversion, or distribution of electric power***
H02G	INSTALLATION OF ELECTRIC CABLES OR LINES, OR OF COMBINED OPTICAL AND ELECTRIC CABLES OR LINES
H02H	EMERGENCY PROTECTIVE CIRCUIT ARRANGEMENTS
H02K	DYNAMO-ELECTRIC MACHINES
H02P	CONTROL OR REGULATION OF ELECTRIC MOTORS, ELECTRIC GENERATORS OR DYNAMO-ELECTRIC CONVERTERS; CONTROLLING TRANSFORMERS, REACTORS OR CHOKE COILS

Table 3: Matching Rates of Central Nodes by the Community Mapping Method and the Betweenness Centrality Method

Year	common	different	match (%)
1980	53	78	40.46
1981	38	56	40.43
1982	33	60	35.48
1983	33	60	35.48
1984	32	60	34.78
1985	30	65	31.58
1986	42	54	43.75
1987	39	56	41.05
1988	36	63	36.36
1989	45	63	41.67
1990	48	79	37.80
1991	34	58	36.96
1992	24	49	32.88
1993	21	52	28.77
1994	32	67	32.32
1995	31	63	32.98
1996	22	69	24.18
1997	27	63	30.00
1998	36	80	31.03
1999	39	87	30.95
2000	53	96	35.57
2001	22	74	22.92
2002	18	59	23.38
2003	5	34	12.82
2004	6	28	17.65

Figure 3: Number of communities at different resolutions

The x-axis indicates years from 1980 to 2013, and the y-axis indicates number of communities.

Figure 4: Average community size at different resolutions

The x-axis indicates years from 1980 to 2013, and the y-axis indicates number of nodes in the community.

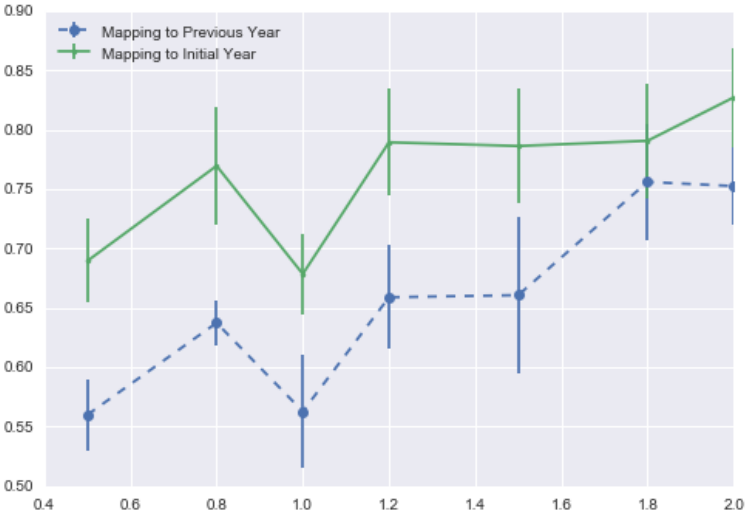


Figure 5: Occurring rate statistics at different resolutions

The mean value and standard deviation of the occurring rate over all the nodes, using the 10-year time window rolling from year to year, including 25 time windows with initial years from 1980 to 2004. The x-axis indicates various resolution values, and the y-axis is the scale of mean values. Results from two mapping methods are shown in this figure.

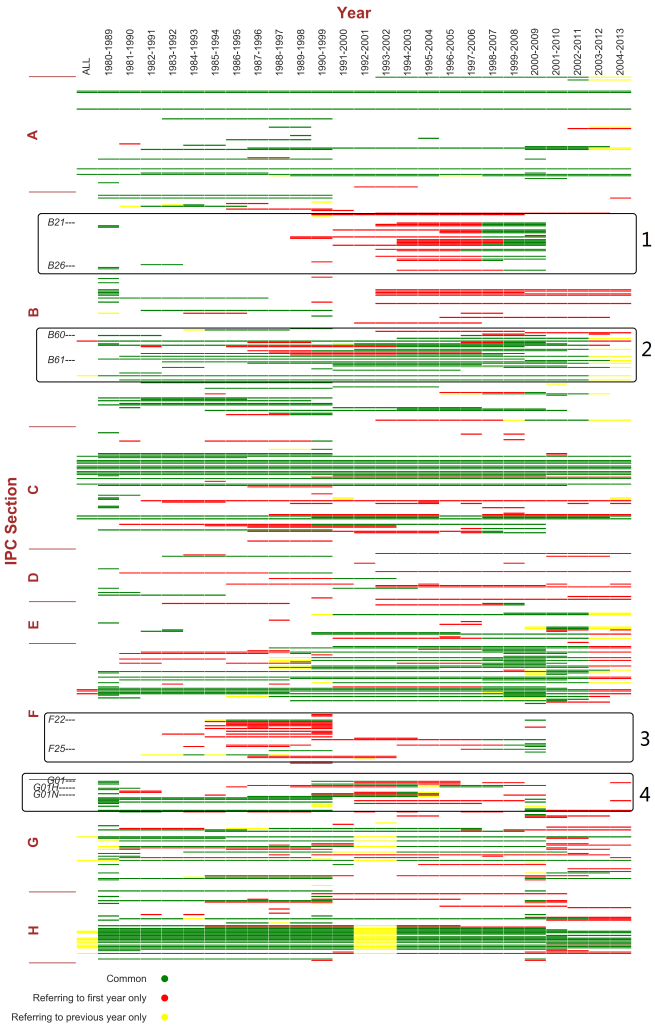


Figure 6: Central nodes of 10-year time window rolling from year to year

X-axis indicates time windows from 1980-1989 to 2004-2013, except for the first column labeled "ALL", which is the all-year condition. Y-axis indicates the IPC subclasses ordered in IPC index. Colored blocks indicate central nodes in a time window using at least one referencing method: Green represents central nodes both methods have in common, red for those central only by referring to the initial year, and yellow for those central only by referring to the previous year.

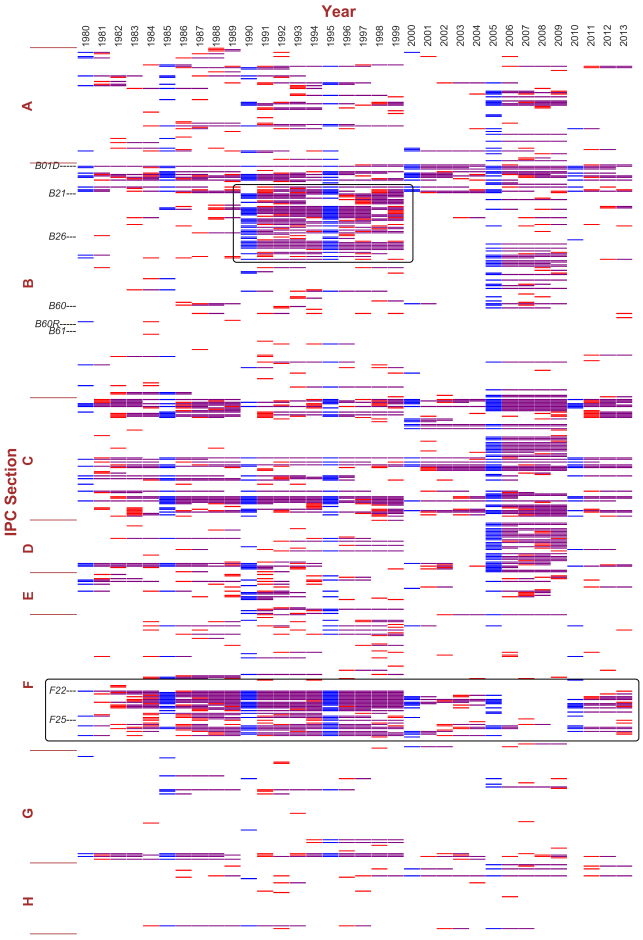


Figure 7: Tracking community B01D in consecutive 5-year time windows, mapping to the previous year

X-axis indicates years from 1980 to 2013, y-axis indicates the IPC subclasses ordered in IPC index. The community mapping is based on consecutive 5-year windows. In each time window, the initial year's community containing the central node set represented by B01D is shown in blue. In the 4 years following, colored nodes represent the mapping communities: red - the node does not exist in the reference community (community of the previous year), purple - the overlapping part.

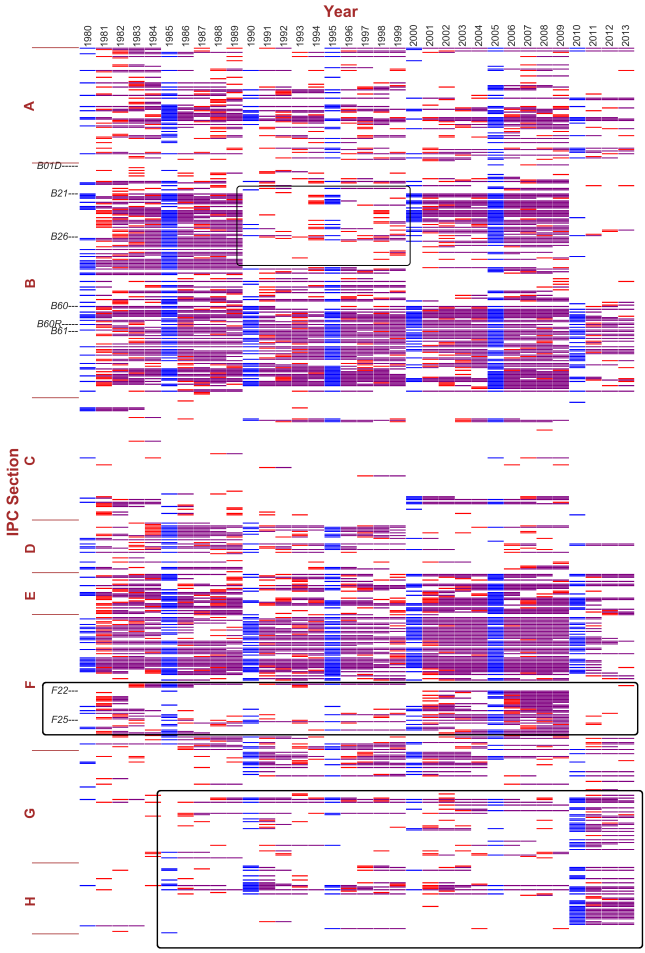


Figure 8: Tracking community B60R in consecutive 5-year time windows, mapping to the previous year

X-axis indicates years from 1980 to 2013, y-axis indicates the IPC subclasses ordered in IPC index. The community mapping is based on consecutive 5-year windows. In each time window, the initial year's community containing the central node set represented by B60R is shown in blue. In the 4 years following, colored nodes represent the mapping communities: red - the node does not exist in the reference community (community of the previous year), purple - the overlapping part.

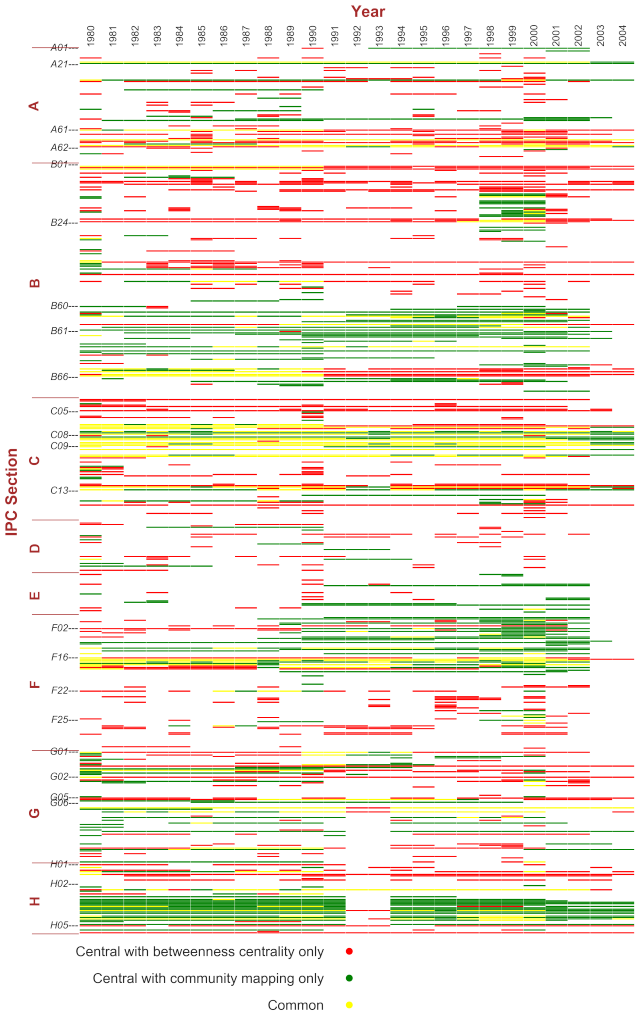


Figure 9: Central Nodes Distribution by Comparing the Community Mapping Method with the Betweenness Centrality Method

X-axis indicates years from 1980 to 2004, y-axis indicates the IPC subclasses ordered in IPC index. The community mapping is based on consecutive 10-year windows. Betweenness centrality values are calculated on 3-year aggregation period started with the same labeled year as the other method. Colored blocks indicate the central nodes in common and different between the two method.

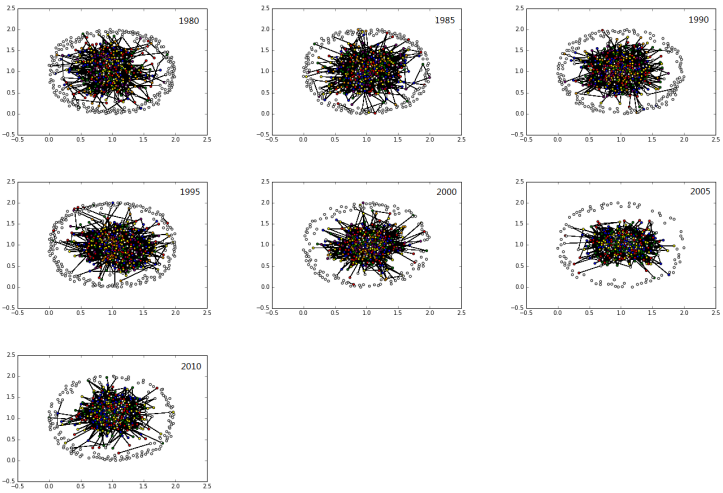


Figure 10: Community structures of the individual sample years based on Louvain modularity optimization algorithm

Resolution is 1.0. Major communities with more than 5 nodes are in the center, with different colors indicating each unique community, surrounded by small communities with 5 nodes or less in white color.

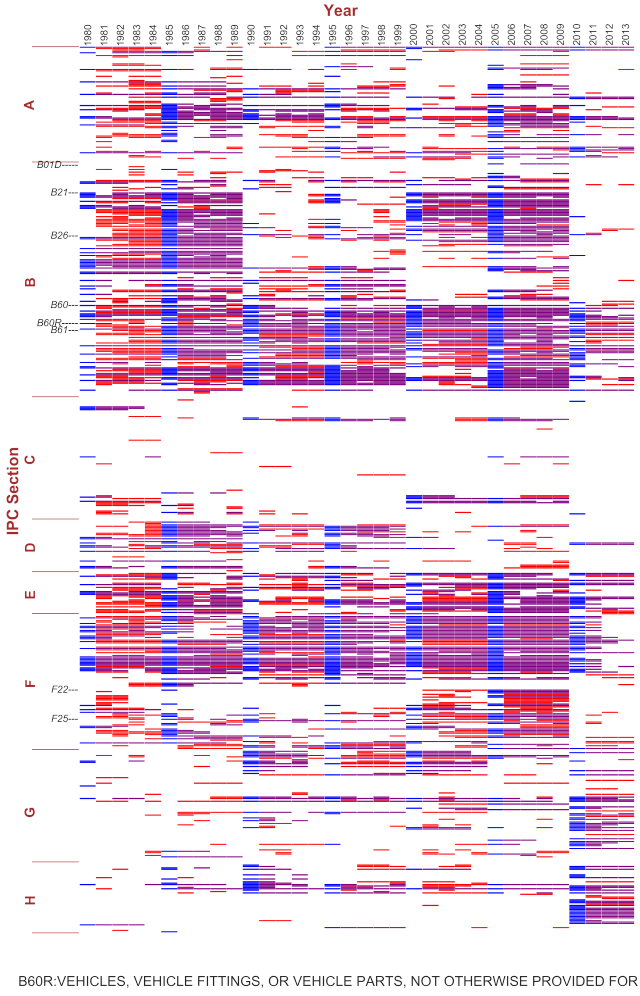


Figure 11: Tracking community B60R in consecutive 5-year time windows, mapping to the initial year

This figure differs from Figure 6 that the overlapping community mapping reference is the initial year of each time window, using the same color coding definitions as Figure 6.

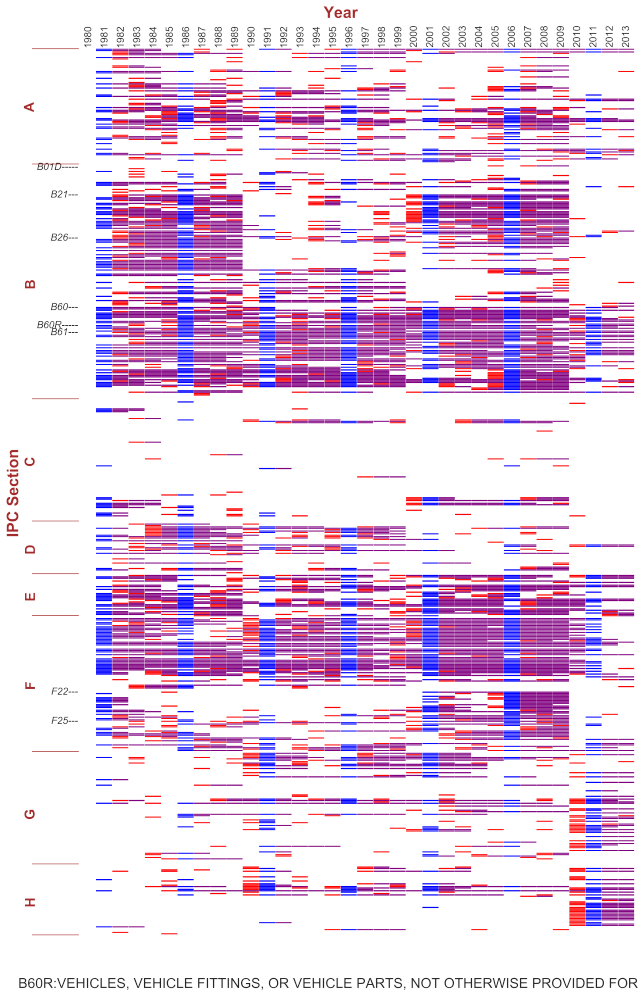


Figure 12: Tracking community B60R in consecutive 5-year time windows, starting from 1981, mapping to the previous year

This figure differs from Figure 6 that the all the time windows are shifted 1 year forward, using the same color coding definitions as Figure 6.

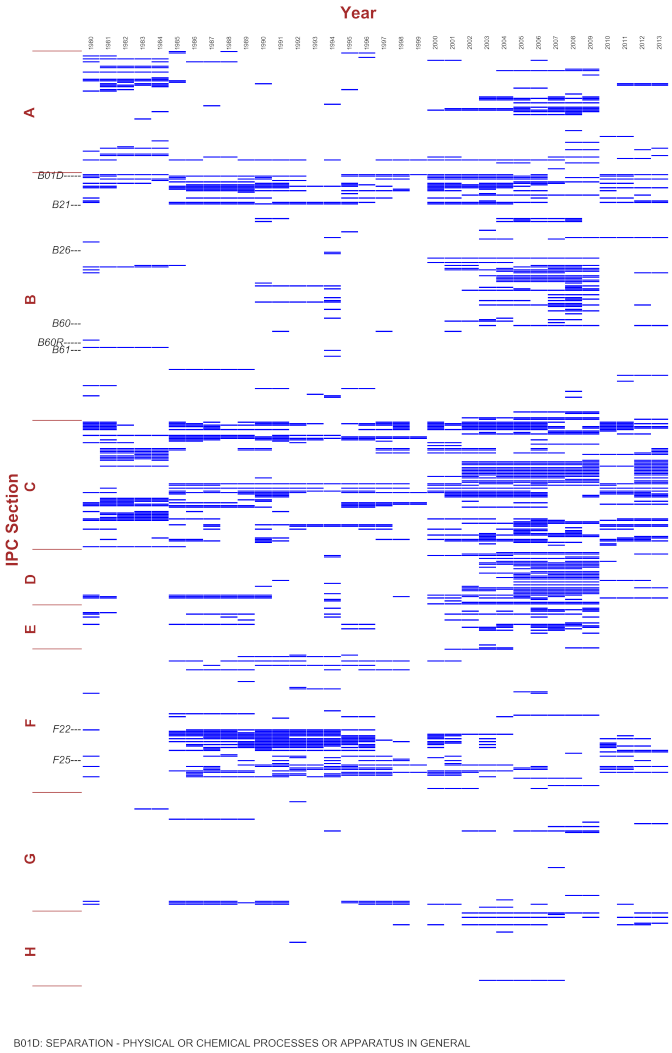


Figure 13: Communities containing B01D in consecutive 5-year time windows based on the multislice community detection method

Starting from 1980-1984. Nodes in blue color are in the same community with B01D in each year.

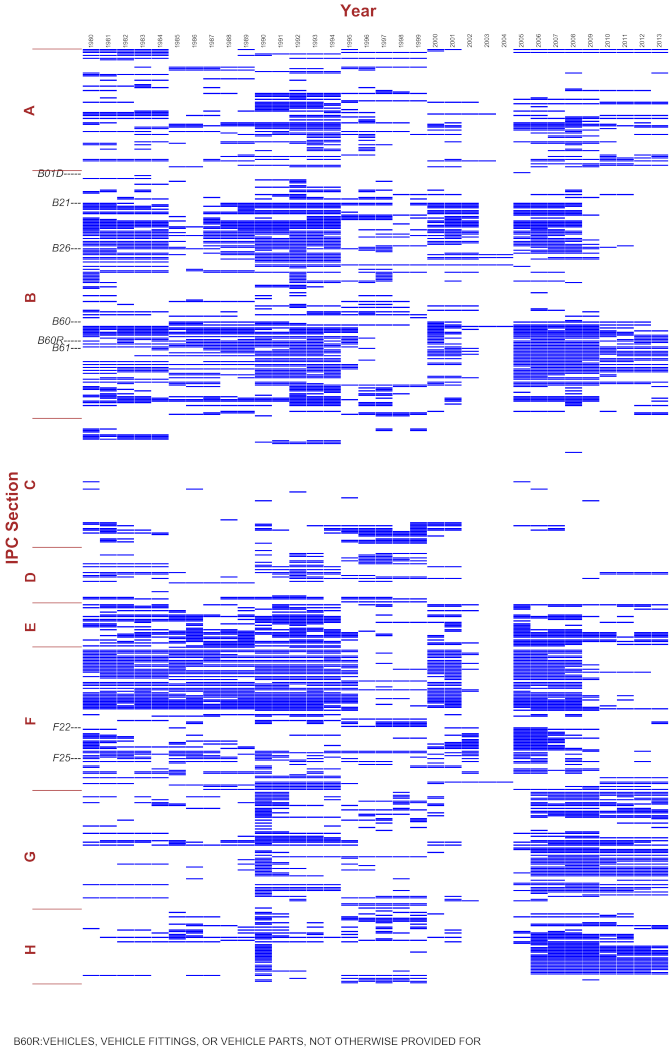


Figure 14: Communities containing B60R in consecutive 5-year time windows based on the multislice community detection method

Starting from 1980-1984. Nodes in blue color are in the same community with B60R in each year.

Chapter 4

More Central in Export, More Central in R&D: A Network Approach to the Pharmaceutical Industry

Abstract

Literature suggests that cross-national knowledge diffusion can boost overall globalization of innovation. International R&D collaboration is one of the ways of such diffusion. Researchers have found innovation as a positive factor for firm and national exporting. This paper analyzes whether the positive relationship exists the other way around: from export to innovation? I perform an analysis in the particularly R&D-driven pharmaceuticals sector using a network approach to test whether being an exporting center in the international pharmaceuticals market could make the country become more central in international R&D collaboration. I construct a cross-national inventor collaboration network based on pharmaceutical patent co-inventions, and use the World Input-Output Database as the international trade network. Through community detection and centrality calculation for both networks, I obtain a

panel data set involving each country's "Coreness" in the two networks to perform regression analysis with other relevant variables. The results show being a trading hub with a favored balance have positive effects for the country to co-invent more patents with more foreign countries in the following 5-6 years, but the effects of drug prices remain inconclusive. This analysis presents an opportunity to help decision makers optimize trade strategies which may enhance domestic R&D impact and contribute to globalized innovation.

4.1 Introduction

The globalization of innovation has been well recognized like the other globalization notions in information, culture, merchandise and finance. Indeed it is not a stand-alone movement, but in close conjunction with the other globalization trends. Due to the increasing importance of knowledge and information in the modern economic processes, governments, firms and researchers have been interested to understand the interactive relationships between innovation and the other factors in order to achieve optimum resource planning strategies and high-efficiency research and development (R&D) activities.

Compared to more result-oriented measures of innovation, such as numbers of patent applications or publications, numbers of new commercialized inventions launched to market, the position of a country in the globalized innovation network has more implications in the long term. The extent of cross-national R&D collaboration not only represents the country's domestic innovative capability, but also indicates its global influences, exposure to knowledge diffusion, opportunities to exploit new applications, and access to international markets and production resources. Archibugi and Iammarino consider three components of technological innovation in their empirical study: the international exploitation of nationally generated innovations through bilateral trade and overseas production of innovative goods, international R&D activities and acquisitions of or joint venture with existing foreign R&D entities by multinational firms; and global techno-scientific collaborations mainly through

joint scientific projects among research and academic institutes (AI02).

Throughout the literature, the relevant factors of innovation have been discussed repeatedly. Among these an important one is international trade. The recognized role of trade in facilitating technology flows and knowledge spillover (Sag02) indicates that trade should promote innovation and narrow down global technology gaps. But it may be a double-bladed sword depending on a country's position in the global market and technological development. As Grossman and Helpman explained, countries, especially the less developed ones, can benefit from the spillovers of their trade partner countries' accumulated knowledge, but may also lose from the insufficient exploitation of their own research investments (GH90). Agion et al. found evidence that the relationship between product market competition and innovation follows an inverted-U curve where competition encourages competitors in a close race but discourages the laggard (ABB$^+$05). The positive effect of innovation and R&D investment in export performance has been widely proved from country level to firm level (Gre90; GTW94; Wak98a; Wak98b; And99; RL02) Meanwhile, empirical studies have found that import from low-wage countries like China have stimulating effects on the importing countries' innovation (BIS07; BSZ08; BDVR16). However, the impact of export on innovation is less analyzed in empirical studies. Is it possible to motivate innovation or expand the global impact of domestic innovation through trade policies and guiding strategies? This chapter will try to answer this question.

In this paper I focuses on the pharmaceutical sector. As DiMasi and colleagues found in a series of studies, the R&D process for new drugs is very lengthy, often taking over a decade, and very costly (DHGL91). As science progresses, the complexity of pharmaceutical R&D increases. Since the mid-1990s, rising risks in successful development, decrease of pharmaceutical R&D productivity and the increase of times and expenditures needed for new drugs development have been reported in many studies (DiM00; DiM01; RP07; DTZ09; Mun09; PMR11; DHG03; DGH16; AB06). This provides further motivation for this analysis to explore whether and how to improve global pharmaceutical innovation

through trade strategies.

Classic and neo-classic studies on trade and innovation generally use R&D investment and patenting activity to measure innovation and trade volume, prices and revenue for trade indicators. In fact, today both trade flows and technology diffusion work in the form of networks. More recent studies began to adopt network approaches to study economic and innovative relations and the dynamics in their transfer and propagation(SP07; ZXT10; ACOTS12; FRZ16), but none has yet studied the relationship between the two networks. In fact, compared to companies in other technological fields, top R&D companies in pharmaceuticals and chemicals have the largest inventor teams and the most countries involved in generating new technologies (DDD$^+$17). Therefore it is more meaningful to study pharmaceuticals using network measures. This chapter takes an initiative to measure how well countries are centrally connected in the innovation network and the trade network. This approach also positions this analysis in the context of global innovation.

4.2 Network Construction

This analysis uses two main datasets for network construction. The innovation collaboration networks or co-invention networks are developed from the OECD patent database[1] (OEC17a), and the trade networks are from the World Input-Output Database (WIOD)[2] (TDL$^+$15).

4.2.1 Co-invention Network (CIN)

I use patents from the OECD Triadic database (DK04) to reduce home bias due to geographic location and to focus on patents deemed to be more valuable. According to the IPC concordance table[3] developed by the World Intellectual Property Organization (WIPO) (Sch08) which links IPC codes to thirty-five technology fields, patents with at least one IPC

[1]Updated version of March, 2018. Data is available upon request at http://www.oecd.org/sti/inno/intellectual-property-statistics-and-analysis.htm

[2]2016 release, available at http://wiod.org/database/wiots16

[3]Updated version of February, 2016

code in the pharmaceuticals field are selected to form a pharmaceutical patent dataset. The OECD REGPAT database (MDW$^+$08a) is used to cross-match with the dataset and provide inventors' locations. From 1980 to 2014, 43 different inventor-hosting countries[4] are found and used as network nodes. Between any of the two countries, each unique co-invented patent is defined as a connecting edge, undirected. Patents are divided by their first-priority years to construct time-slice networks.

For the purpose of this research, inventor country is used instead of applicant country. Consider the case where a patent is co-invented by employees working at multiple overseas sites of the same organization, while the applicant country is only where the headquarter location.

4.2.2 World Input-Output Network (WIOD)

Using input-output data to study cross-national trade is a well-recognized practice in economic research. Literature shows that it goes back to at least 1950s (Leo53) and remains active among recent studies (JN12; KWW14; PRTZ17). The WIOD provides bilateral inter-country and inter-sector input-output tables that reflect the domestic and international supply and use of products in different sectors. The 2016 Release covers 56 sectors classified following the International Standard Industrial Classification revision 4 (ISIC Rev. 4[5]) in 44 countries (including 28 EU countries and 15 other major countries, and a residual Rest-of-World) over the period from 2000 to 2014 (TLSdV16). For each year, a World Input-output Table (WIOT) is readily prepared as a network matrix in which each sector of a country is a node and the input-output values between them are the edges. For this research I only use the pharmaceutical products sector[6] of each country's supply and use data. The result is a 44×44 matrix for each year, in which an element at the crossing of row i and column j

[4]Countries in the OECD patent database are provided as ISO 2 country codes

[5]The classification is published and maintained by the United Nation Statistics Department (UNSD) and available at https://unstats.un.org/unsd/publications/catalogue?selectID=396

[6]ISIC Rev. 4 code C21: Manufacture of basic pharmaceutical products and pharmaceutical preparations

represents the output of pharmaceutical products from country i to country j.

Between the 43 countries from the CIN and the 43 countries (ROW excluded) from the WION, 32 countries are in common and will be used for the relationship analysis afterwards. Table 7 in Appendix lists out the 32 common countries and those only in one of the 2 networks.

4.3 Network Analysis

In this section I introduce the methods used for network analysis. First I will describe the two networks in 2D matrix, and then conduct community detection and calculate the nodes' centrality.

4.3.1 Network Matrix Description

Considering the lengthy times needed for pharmaceutical R&D and to compensate for statistical instability due to co-invented patent quantity variances in some countries from year to year[7], for the CIN, data is aggregated in 5-year windows. Figure 15 shows an example of the CIN in the time window 1995-1999. The 43 countries are arranged by rows and columns. The CIN is undirected, and the cells are normalized by column considering the gaps in scale between developed countries with major pharmaceutical multinational companies (MNC) and the less developed countries. So a cell of row i and column j represents the importance of invention collaboration with country i for country j. A cell with deeper color indicates the collaboration with i plays a bigger role for the overall pharmaceutical patenting activities of j. Note that Qatar and Uganda have no inventors contributing to any pharmaceutical patents in this time period.

Unsurprisingly, for most of the countries, the majority of pharmaceutical inventions happen within the border. But there are a few exceptions. For example, Indonesia's collaboration with Japan and the Netherlands

[7] In particular, less developed countries that mainly generate patents from subsidiaries of large overseas companies

generates more patents than its domestic collaboration. The Philippines and Thailand are extreme examples of such case. Both countries barely have any domestic innovation activities compared to overseas collaboration: The Philippines with the U.S., the UK and Belgium, and Thailand with Germany, the U.S. and the UK. It is noteworthy that both countries have developed deeper-colored diagnosis cells in the following time period. This intuitively supports the theory that R&D collboration facilitates technology transfer and thus helps the less developed countries develop their own innovation resources.

Meanwhile, some countries are "popular" innovation partners universally, such as the United States, Germany and the UK. While overseas collaboration only counts for a small portion of their pharmaceutical inventions, joint innovation partnership with them contributes significantly to many other countries, especially the less developed ones like China, Greece, the Phillippines, Thailand, Turkey and Ukraine.

Figure 19 in Appendix shows such rendering of CINs of the other time periods. Obviously, with time more countries are participating in the global pharmaceutical innovation and producing patents. The "popular" partners remain popular over time, but overall the dominance of domestic invention has reduced, with the diagnosis becoming lighter-colored.

Figure 16 shows an example of the WION in the aggregated time window 2000-2004. The trade volumes are also normalized by column. Similarly, the namely Germany, the U.S. and the UK. In addition, the WION shows other "popular" countries like Belgium, Switzerland and Ireland. Note that the input and output links in the WION are directed, but the trade volume values in the matrix does not differentiate between the two directions. For example, Switzerland is largely importing and the United States is mainly exporting. It takes further analysis to relate information of the two networks.

Figure 20 in Appendix shows the WION rendering for the other time periods. WIOD does not have pharmaceutical trade data for Russia and Turkey. Overall, during the 15 years the WION does not show obvious changes. It is noteworthy that the dominating significance of domestic

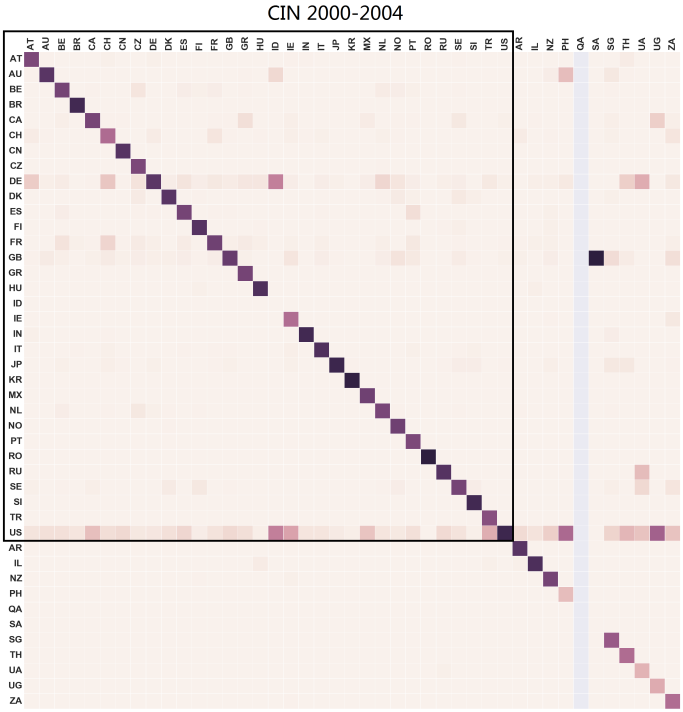


Figure 15: 2D Rendering of Co-invention Network Aggregated from 2000 to 2004

trade is less universal compared to the CIN.

4.3.2 Community Identification

In the first two chapters (GZR17b; GZKRng), I have introduced two methods to identify naturally formed network clusters: the Lumped Markov Chain method (Pic11) and the Stablized Louvain method (AG10a). The

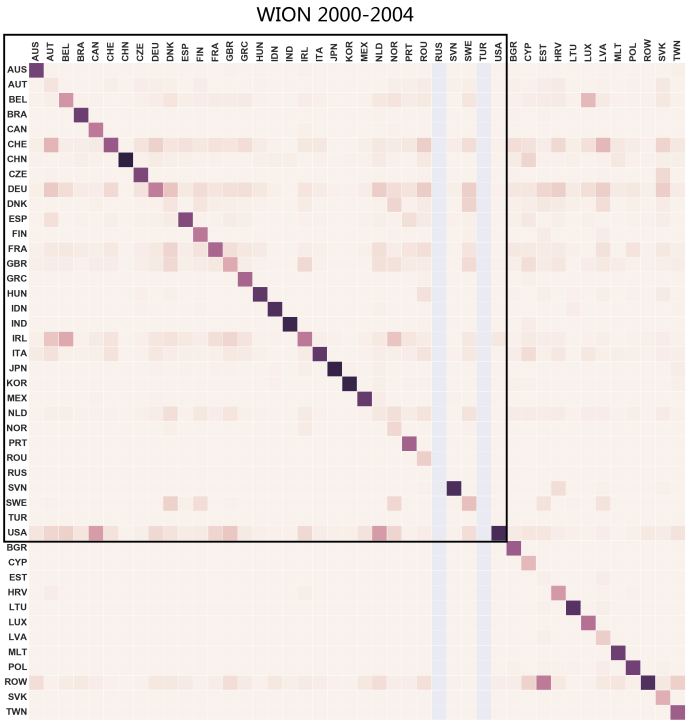


Figure 16: 2D Rendering of WIOD Network Aggregated from 2000 to 2004

former has been applied by Piccardi et al. in an analysis of the WIOD network (PRTZ17). Indeed it is a preferred method in this analysis because the input-output system can be modeled as a regular Markov chain. In this system, the output, or supply side analysis can be viewed as a "salesman" randomly walking from one industry to the next to sell products (MB09), while the input, or demand side can be modeled as a "purchasing agent" visiting one industry to the next to buy products (Leo53).

Since I'm using the WIOT with only one sector, and for the research interest in this analysis, I do not customize the WION with a "salesman" or "purchasing agent" perspective as Piccardi et al. did, but instead use a country's input-output ratio to tell whether the country has a favorable or adverse balance of pharmaceutical trade.

To observe the community changes over time, I conduct community detection for time slices they each covers. For the CIN, 5-year aggregate rolling time windows are used and for the WION each time window is simply a single year. This is because the CIN potentially has more changes, and smaller numbers of patents for some countries in a single year might lead to fluctuant noises. Rolling time windows of more years' aggregated data can enhance stability. Figure 17 and 18 show the over time network partitioning of the two networks using the Lumped Markov chain method. They share a common period of 2000 to 2014. For comparison purpose, the persistence probability threshold is configured so that it would be eight community in each time window for either network. In each figure, the rows represent countries in the network and are labeled with ISO 2 or 3 codes, and the columns are time windows. Each color represents a unique community. A time window is treated as a independent static time slice for network clustering, therefore the colors are not consistent across time windows. Only countries of the same color in the same time window belong to the same community. But for demonstration, I have adjusted the community IDs to enhance visual consistency. Countries are indexed according to the region they are located, according to the M49 standard published by the United Nations[8]. Table 8 in Appendix provides details of the region codes and the grouping of countries. The regional grouping is used as an exogenous reference to the endogenous communities. In both networks, countries that can't be clustered to any community due to lack of data or very weak connections are in black.

Comparing Figure 17 and Figure 18, it is obvious that the CIN shows more dynamics over time, whereas the WION is more stable. Both net-

[8]The assignment of countries or areas to specific groupings, available at https://unstats.un.org/unsd/methodology/m49/

works have a "main" community containing the majority of the nodes, colored as cyan, but the CIN's "main" community is smaller. In the WION, other smaller communities besides the "main" community are very small and have more single-node communities than the CIN. This remains true with varying persistence probability thresholds. Therefore, the WION tends to be more balanced and more persistently integrated.

Although the two networks have only partial countries in common, one community can still be observed in both figures: Japan and South Korea (and Taiwan in the WION) at the end of the first decade of the current century. More communities are identified in the CIN: Countries in America plus China and Japan formed a community from mid-1980s to early 1990s, from which the American countries broke away after 1994. A possible relevant fact is the launch of North American Free Trade Agreement (NAFTA) in 1994. The NAFTA might have strengthened the innovative cooperation among the U.S., Canada and Mexico with favored trade terms, although the three countries are absorbed into the "main" community for at least the next six years or more, before Canada forms a small community with Brazil for a short period. A larger and more persistent community among the North European countries is also captured, including mainly Denmark, Finland, Norway and Sweden, and briefly Belgium. In the time window of 1955-1999 it even expands to include three South European countries, but most often it is Spain, Italy and Portugal that are more closely connected. Besides, a small community between India and Thailand can also be observed in the early 2000s.

Network clustering is inevitably sensitive to the threshold setting of persistence probability, but the two examples provide proof that the random walking method is capable to capture communities that coincide with geographic grouping and potential trade cooperative relationship which is often related to geographic proximity as well.

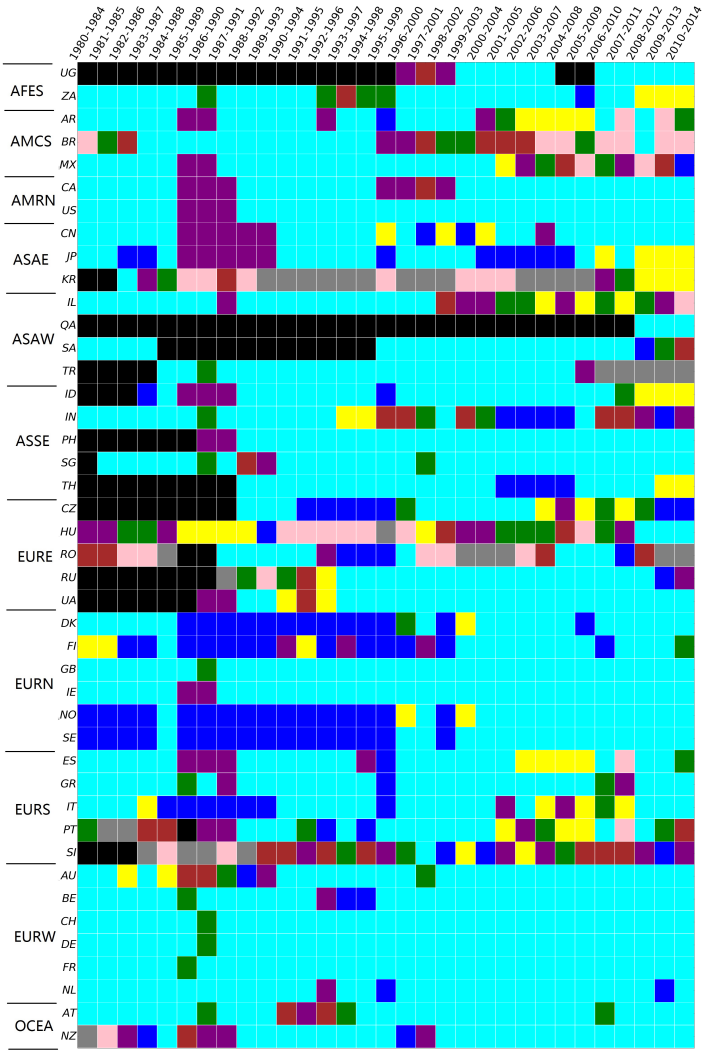


Figure 17: 8-Community Structure of the Co-invention Networks Using Lumped Markov Chain Method

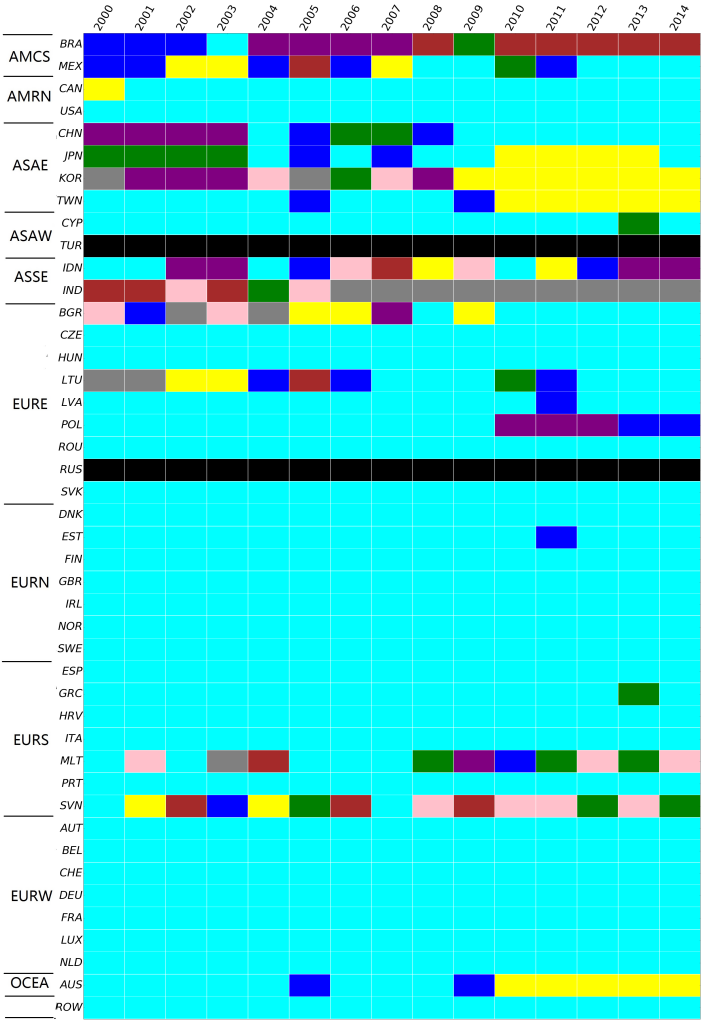


Figure 18: 8-Community Structure of the WIOD Networks Using Lumped Markov Chain Method

I have also performed community identification using the Stabilized Louvain method. The results are a much larger number of small communities. For the CIN, at resolution 1.0, the network of the first 5-year time window is partitioned into 37 clusters out of 43 nodes. Even using the stabilizing algorithm to build the network partition of each time slice considering the result partition from the previous time slice (AG10a; GZKRng), the last time window still has 19 communities with an average community size of 2.26. For the WION, at the same resolution of 1.0, the initial time window has 22 communities and the last has 16. Different from the CIN where the maximum community size is 4, the largest community in the WION has 21 nodes, which confirms the finding using the Lumped Markov chain method. However, the rest of the communities mostly have only a single node. Changing the resolution setting can reduce the number of partitions, but also increases the gap between community sizes. This method is not ideal for this analysis because it's not sensitive enough to the formation of smaller communities. Such a polarized partitioning also leads to inaccurate calculation of centralness using the community overlapping method (WF10a; GZKRng) because single-node communities tend to be dropped and most likely only the largest community can be used for mapping. Therefore, the following analysis will be based on the Lumped Markov chain method.

4.3.3 "Coreness Calculation

I use the same method as in Chapter 1 (GZR17b) to calculate "Coreness": the within-community closeness of each node weighted by its community's persistence probability. To avoid influence due to different "Coreness" scales of the two networks, "Coreness" ranking is used instead of the value. Considering sensitivity to threshold setting, an average of ranking at different thresholds is used, as the number of communities increases from 4 to 12. Take a further average over time, the top 5 "core" countries in the CIN are: the U.S., Germany, the UK, Indonesia and France, and for WION: Switzerland, Ireland, Germany, the UK and Denmark. Table 9 in Appendix shows the complete ranking of the 32

countries shared by the two networks.

4.4 Regression Analysis

I use linear regression models to examine the effects of trade on innovation. Since I'm interested in how well a country is centrally connected in global R&D collaboration, I define the dependent variable as "Coreness" of each node in the CIN, more specifically, "Coreness" ranking order of each node in the CIN of each 5-year time window. I also construct an additional analysis where the dependent variable is the forward citation number, which is well recognized as a patent quality measure, especially for patent technology importantce and potential economic values (SDC13a). It is therefore used as an alternate dependent variable to reflect the impact of pharmaceutical innovation. When using trade as an independing variable, I first check the more conventional and simpler measures, trade volumes. National or firm-level input and/or output volumes have been widely used by researchers to study the impact of R&D on trade performance (GH90; And99; RL02; BDVR16), together with other recognized indicators of patenting and R&D, including the number of patent applications and investment in innovation. Then I use the WION "Coreness" ranking in the similar way as CIN "Coreness" ranking to check the network approach. In addition, since I'm particularly interested in the impact of export, a trade balance variable is included in the model.

For the 29 common countries between the two networks[9], they are re-ranked for consistency while the relative original ranking orders in each network are preserved. For example, nodes A, B and C are ranked 1, 2 and 3 in the CIN network[10], and B is not in the WION nodes. Then A and C will be re-ranked as 1 and 2 for their CIN "Coreness" in the common nodes set. By doing this the "Coreness" ranking of the two networks are on the same scale.

[9]Out of the 32 shared countries, Russia and Turkey are removed due to missing WION data and Indonesia is removed due to lack of patent citation data.

[10]Smaller numbers represent larger "Coreness" values

For any node *i* in the CIN of the 5-year aggregate time window starting from year*t*, its invert "Coreness" ranking order as CIN_{it}:

$$CIN_{it} = 29 - CR_{it}, \tag{4.1}$$

where CR_{it} represents node *i*'s CIN "Coreness" ranking in year *t*. With the simple subtraction, larger CIN_{it} indicates a more persistently central node in the patent co-invention network.

For WION, the invert "Coreness" ranking is defined similarly as $WION_{it}$, and $WION_{it}^{'}$ is further weighted by the export trade balance:

$$WION_{it} = 29 - WR_{it}, \tag{4.2}$$

$$WION_{it}^{'} = WION_{it}\ E_{it}, \tag{4.3}$$

$$E_{it} = \frac{\sum_j O_{jit}}{\sum_j I_{ijt} + \sum_j O_{jit}}, \tag{4.4}$$

where WR_{it} represents node *i*'s WION "Coreness" ranking in year *t*, E_{it} is the share of export in total input and output volumes, I_{jit} is the input from country *j* to country *i*, and O_{ijt} is the output from country *i* to country *j*. E_{it} is a value between 0 and 1 that represents the extent of a country being favorable balanced in its overall bi-lateral pharmaceutical trade. Larger $WION_{it}^{'}$ value indicates a country being more persistently central in the international pharmaceutical trade market with a favorable balance.

Other factors that could influence national innovation capability are taken into consideration, including[11]:

1. Business Enterprise Expenditure on R&D (BERD) performed in the pharmaceutical industry: in current PPP Millions $
2. Percentage of BERD performed in the pharmaceutical industry: in percentage units

[11] 1 and 2 are obtained from OECD's Main Science and Technology Indicators (MSTI) dataset, release of 1 March 2018, available at https://www.oecd-ilibrary.org/science-and-technology/data/oecd-science-technology-and-r-d-statistics_strd-data-en. 3-6 are obtained from the OECD patent database, and 7 is based on the composite ranking in a 2016 ITIF report

3. Number of pharmaceutical patents filed, grouped according to the first-priority year and patent applicant country[12]

4. Number of individual inventors listed in the pharmaceutical patents filed, grouped according to the first-priority year and inventor country

5. The average number of outgoing citations from pharmaceutical patents filed, grouped according to the citing patent's first-priority year and patent applicant country

6. The average number of incoming citations received by each pharmaceutical patent filed, in 5 years after the patent publication, grouped according to the cited patent's first-priority year and patent applicant country

7. National bio-pharmaceutical price control index (WE16), used for control: 1 for high extent of control, 2 for moderate and 3 for low.

Among these variables, No. 6 is the forward citation number to be used as an alternate dependent variable. No. 5, also known as backward citation number, is an indicator of patent novelty and patentability (SDC13a) and represents the extent of inventors reply on prior art. It is used as an independent variable.

4.4.1 Results

Considering fixed effects of time and country, for the 30 countries in the time range from 2000 to 2010[13], results of the linear regression models are shown in Table 4 to Table 6. Analysis 1 uses CIN "Coreness" as dependent variable and includes trade volumes as independent variabls. The other two analyses use WION "Coreness" ranking as regressor: Analysis 2 uses CIN "Coreness" as dependent variable, and Analysis 3 uses

[12]Pharmaceutical patents are those patents with at least one IPC classification as pharmaceutical sector according to the IPC concordance table

[13]Note that for CIN, the time window starting 2010 covers 2010-2014

average incoming citation quantities. Analysis 2 and 3 also include estimations using national drug price control as a categorial variable[14].

Table 4: Analysis 1: The Changes of CIN "Coreness" on Input and Output Volumes

Dependent: CIN coreness ranking		Base	Including BERD variables		
		1a	2a	3a	4a
1	log of total input volume	-0.1200*	-0.2728*	-0.2756*	-0.2435
		(0.0611)	(0.1456)	(0.1446)	(0.1572)
2	log of total output volume	0.1954*	0.3016**	0.3198**	0.2942*
		(0.1065)	(0.1353)	(0.1317)	(0.1505)
3	log of average outgoing citation quantity	0.0528*	0.0324	0.0370	0.0368
		(0.0309)	(0.0357)	(0.0368)	(0.0355)
4	log of patent quantity	0.1568	0.2107	0.2167	0.2263
		(0.1098)	(0.1342)	(0.1326)	(0.1334)
5	log of inventor quantity	0.0116	-0.1422	-0.1456	-0.0716
		(0.2006)	(0.1982)	(0.1960)	(0.2195)
6	log of BERD in pharmaceuticals		0.0049		-0.2270
			(0.0887)		(0.1947)
7	percentage of BERD performed in pharmaceuticals			0.0838	0.2894
				(0.1291)	(0.2509)
8	Dummy variables: Year		Included in all		
	Number of countries	29	20	20	20
	Number of observations	289	195	195	195

* p<0.10, ** p<0.05,*** p<0.01
Note: Dependent variable is the CIN "Coreness". All the models are estimated by panel fixed-effect method. Estimations report coefficients and robust standard errors. 1a is the base model in this analysis. Input and output volumes are in millions of dollars.

Analysis 1 shows negative coefficients of input volume and positive coefficients of output volumes with medium-low significance levels. The R-squared value of Model 4a is low (0.0357), smaller than that of Model 4b (0.1835). In Analysis 2, variable 9 is insignificant, but when weighted by export trade ratio, variable 10 becomes consistently significant[15]. In

[14]High control countries: Australia, Spain, France, the UK, Norway. Moderate control countries: Austria, Belgium, Canada, Czech Republic, Germany, Hungary, Italy, Japan, Korea, Portugal, Romania, Slovenia. Low control countries: Switzerland, Mexico, the U.S.

[15]The export trade ratio is used as a weight but not a regressor due to lack of theoretical

Table 5: Analysis 2: The Changes of CIN "Coreness" on WION "Coreness"

	Dependent: CIN coreness ranking	Base		Including BERD variables			Include price control level	
		1b0	1b	2b	3b	4b	5b	6b
9	log of WION coreness ranking	-0.0414						
		(0.1287)						
10	variable 9 weighted by the share of output in total trade volume		0.2501**	0.4909***	0.4973***	0.4679**		
			(0.1107)	(0.1653)	(0.1637)	(0.1699)		
11	variable 10 X high price control						0.2475	0.3515
							(0.1571)	(0.3779)
12	variable 10 X moderate price control						0.2679*	0.5456**
							(0.1446)	(0.2112)
13	variable 10 X low price control						-0.2804	-0.0516
							(0.1983)	(0.4396)
3	log of average outgoing citation quantity	0.0288	0.0278	0.0313	0.0343	0.0348	0.0256	0.0309
		(0.0307)	(0.0314)	(0.0459)	(0.0473)	(0.0452)	(0.0315)	(0.0472)
4	log of patent quantity	0.0852	0.1239	0.2526	0.2542	0.2692	0.1246	0.2751
		(0.1245)	(0.1359)	(0.1590)	(0.1570)	(0.1587)	(0.1395)	(0.1635)
5	log of inventor quantity	-0.0508	-0.0319	-0.0697	-0.0855	0.0096	-0.0275	-0.0012
		(0.1737)	(0.1740)	(0.1461)	(0.1536)	(0.1656)	(0.1721)	(0.1786)
6	log of BERD in pharmaceuticals			-0.0493		-0.2636		-0.2164
				(0.0892)		(0.1538)		(0.1963)
7	percentage of BERD performed in pharmaceuticals				0.0210	0.2653		0.2197
					(0.1296)	(0.2242)		(0.2736)
8	Dummy variables: Year	Included in all						
	Number of countries	29	29	20	20	20	20	20
	Number of observations	273	273	190	190	190	273	190

* p<0.10, ** p<0.05,*** p<0.01

Note: Dependent variable is the CIN "Coreness". All the models are estimated by panel fixed-effect method. Estimations report coefficients and robust standard errors. Model 1b is the base model in this analysis, using weighted WION "Coreness". Model 5b and 6b include drug price control extent as a categorical control with 3 values from high to low.

Table 6: Analysis 3: The Changes of Average Incoming Citation Quantity on WION "Coreness"

Dependent: Log of average incoming citation quantity			Base	Including BERD variables			Include price control level	
		1c0	1c	2c	3c	4c	5c	6c
9	log of WION coreness ranking	-0.0089						
		(0.0711)						
10	variable 9 weighted by the share of output in total trade volume		0.1014	0.3129	0.3401	0.3383		
			(0.1606)	(0.2188)	(0.2235)	(0.2185)		
11	variable 10 X high price control						0.5464*	0.2023
							(0.3061)	(0.5604)
12	variable 10 X moderate price control						-0.0478	0.44
							(0.1303)	(0.2601)
13	variable 10 X low price control						-0.103	-0.409
							-0.1712	-0.2908
3	log of average outgoing citation quantity	0.0414	0.0436	0.1877***	0.1839***	0.1842***	0.0375	0.1831***
		(0.0720)	(0.0726)	(0.0565)	(0.0571)	(0.0621)	(0.0754)	(0.0631)
4	log of patent quantity	-0.0369	-0.0191	0.1087	0.0935	0.0947	-0.0513	0.1139
		(0.1050)	(0.1151)	(0.0873)	(0.0809)	(0.0850)	(0.1228)	(0.0846)
5	log of inventor quantity	-0.0527	-0.0422	0.0349	-0.0369	-0.0302	-0.0935	-0.0289
		(0.1297)	(0.1313)	(0.2584)	(0.2335)	(0.2939)	(0.1343)	(0.3081)
6	log of BERD in pharmaceuticals			-0.1564		-0.0153		0.0022
				(0.1037)		(0.3096)		(0.3074)
7	percentage of BERD performed in pharmaceuticals				-0.1859*	-0.1722		-0.2104
					(0.1045)	(0.3147)		(0.3111)
8	Dummy variables: Year	Included in all						
	Number of countries	29	29	20	20	20	29	20
	Number of observations	250	250	172	172	172	250	172

* p<0.10, ** p<0.05,*** p<0.01

Note: Dependent variable is the average incoming citation quantity within 5 years after the patent's publication. All the models are estimated by panel fixed-effect method. Estimations report coefficients and robust standard errors. Model 1c is the base model in this analysis, using weighted WION "Coreness". Model 5c and 6c include drug price control extent as a categorical control with 3 values from high to low.

Analysis 3, variable 3 has a positive significant impact on the average number of incoming citations. The other factors are not statistically significant, including BERD investment, patent application quantity and inventor quantity.

4.5 Discussion

Although the statistical significance of weighted WION "Coreness" ranking in Analysis 3 is weak, its coefficients are consistently positive. Analysis 2 and 3 indicate that being an central exporting country in the global pharmaceutical market tends to promote patenting collaboration with foreign countries and produce patents with higher technological and economic values. Of course, as discussed earlier in literature review, the positive coefficient might be viewed as reverse causality: Being central in R&D collaboration promotes a country's position in export. In my analysis, the dependent variable CIN "Coreness" ranking is calculated over a 5-year window, while the independent variables are of the first year of the time window. I perform a further check with the independent variables leading one year ahead of the dependent variable, i.e. WION "Coreness" of year 2000 is used to model CIN "Coreness" of the aggregate time window from 2001 to 2005. The results are shown in Table 10 in Appendix. For Model 1b and 4b, adding one year lag still shows significant positive effects of variable 10.

This analysis is by no means a denial to the causality that innovation can bring up export in trade. It is to explore the causality the other way around. In the global market the causality in both directions can co-exist. Through this analysis and future work, I hope to provide evidence for countries and business to optimize trade strategy so that it could help improve their status in R&D collaboration, enhance their R&D impacts,

support that trade balance directly influences co-invention centrality. I also found weak correlation between E_{it} and log of $WION_{it}$ (Pearson correlation coefficient is 0.1036), and E_{it} and CIN_{it} (Pearson correlation coefficient is 0.0198). Further checks using a pooled linear regression model including E_{it}, $WION_{it}$ and their interaction term show that when holding E_{it} as a constant at different values, the positive marginal change of CIN_{it} caused by one unit change in $WION_{it}$ increases. The collinearity between E_{it} and $WION_{it}$ is low. Therefore I include E_{it} as a weight on $WION_{it}$.

increase opportunities to benefit from knowledge diffusion and international resources, and promote globalized innovation.

Literature hold that stricter drug price controls could systematically lower the prices of drugs sold in the country and inhibit market access, and the decrease in profits could harm R&D incentives for pharmaceutical firms (DC00; GSV05; WE16). Magaret also suggests that firms headquartered in countries with price controls reach fewer markets. But she also points out that consequently, firms could avoid price-controlled countries and limit market launches in additional countries (Kyl07). The existence of parallel trade in the European market and the relevant legalization and responding product launch strategies taken by pharmaceutical firms are also changing the game. Drug price is therefore a factor of interest in my analysis. However, due to difference in market access, new drug launch procedures, medicine packing and distribution channels, comparing cross-national drug prices is complicated. Let alone the relevant factors of national health care reimbursement policies. In this analysis I use the composite price control extent index from the ITIF report (WE16) as a categorical factor. The result of Model 5b, 6b and 5c, 6c also suffer from small sample sizes of each price control category and are difficult to interpret. The results tend to indicate that in countries with higher price controls, the positive effects of being central in export are more robust. A possible explanation is lower domestic market values could encourage firms to export to high-valued foreign markets, and consequently raise their chances of international R&D collaboration.

For the incoming citations received, the number of outgoing citations is consistently significant. This echoes the knowledge spillover literature: Stand on the shoulders of giants, and you can see further. It is also intuitively reasonable that when the number of cross-national co-invented patents increases, the total number of citations received will go up because co-inventors are motivated to cite their own prior inventions, and because the IP protection coverage has expanded to more countries.

4.6 Conclusion

This chapter attempts to answer the question "how does a country's position in the world trade network, export in particular, affect its position in global innovation in the pharmaceutical industry?" Through network community detection, I present descriptive analysis of how countries are grouped together in the trade network and the co-invention network. Further analysis using linear regression models shows that exporting more pharmaceutical products to more different foreign markets helps a country forge more co-invented pharmaceutical patents with other countries, and the resulting patents are more likely to be of higher quality and values. Compared to simple trade volume measure, the network approach is more accurate in estimating changes of CIN "Coreness" ranking.

The impact of drug prices and national price control policies is rather inconclusive and presents continuous research potential. Inclusion of reliable bi-lateral drug prices, reimbursement policies and market access regulations will be important in future work.

This study adds to the literature of relationships between innovation and trade. It also provides a new estimation approach using networks measures. In the trending discussions about enhancing innovation efficiency on national and firm levels theses days, the yet preliminary results of this analysis may be suggestive for decision makers in making national trade policies and for pharmaceutical business owners in considering foreign market accessing strategies.

4.7 Appendix

[16]Inventor-hosting country in ISO 2 codes
[17]WIOD countries in ISO 3 codes
[18]Countries appearing in co-invention networks only
[19]Countries appearing in WIOD only

Table 7: Co-invention countries and WIOD countries

Country	CIN[16]	WION[17]	Country	CIN[18]	Country	WION[19]
Austria	AU	AUT	Israel	IL	Croatia	HRV
Belgium	BE	BEL	New Zealand	NZ	Cyprus	CYP
Brazil	BR	BRA	Philippines	PH	Estonia	EST
Canada	CA	CAN	Qatar	QA	Latvia	LVA
China	CN	CHN	Saudi Arabia	SA	Lithuania	LTU
Czech Rep.	CZ	CZE	Singapore	SG	Luxembourg	LUX
Denmark	DK	DNK	South Africa	ZA	Malta	MLT
Finland	FI	FIN	Thailand	TH	Poland	POL
France	FR	FRA	Uganda	UG	Rest-of-World	ROW
Germany	DE	DEU	Ukraine	UA	Slovakia	SVK
Greece	GR	GRC			Taiwan	TWN
Hungary	HU	HUN				
India	IN	IND				
Indonesia	ID	IDN				
Ireland	IE	IRL				
Italy	IT	ITA				
Japan	JP	JPN				
Mexico	MX	MEX				
Netherlands	NL	NLD				
Norway	NO	NOR				
Portugal	PT	PRT				
Romania	RO	ROU				
Russia	RU	RUS				
Slovenia	SI	SVN				
South Korea	KR	KOR				
Spain	ES	ESP				
Sweden	SE	SWE				
Switzerland	CH	CHE				
Turkey	TR	TUR				
UK	GB	GBR				
USA	US	USA				

Table 8: Regions and Countries

Region	Name	WION	CIN
AFRE	Southeast Africa		UG
			ZA
AMCS	Central-south America		AR
		BRA	BR
		MEX	MX
AMRN	North America	CAN	CA
		USA	US
ASAE	East Asia	CHN	CN
		JPN	JP
		KOR	KR
		TWN	
ASAW	West Asia	CYP	
			IL
			QA
			SA
		TUR	TR
ASSE	Southeast Asia	IDN	ID
		IND	IN
			PH
			SG
			TH
EURE	East Europe	BGR	
		CZE	CZ
		HUN	HU
		LTU	
		LVA	
		POL	
		ROU	RO
		RUS	RU
		SVK	
			UA
EURN	Noth Europe	DNK	DK
		EST	
		FIN	FI
		GBR	GB
		IRL	IE
		NOR	NO
		SWE	SE
EURS	South Europe	ESP	ES
		GRC	GR
		HRV	
		ITA	IT
		MLT	
		PRT	PT
		SVN	SI
EURW	West Europe	AUT	AU
		BEL	BE
		CHE	CH
		DEU	DE
		FRA	FR
		LUX	
		NLD	NL
OCEA	Oceanic	AUS	AT
			NZ
		ROW	

Table 9: Average "Coreness Ranking Over Time and Persistence Probability Thresholds

CIN	CIN ranking	WION	WION ranking
US	1.242424	CHE	1.081481
DE	3.818182	IRL	2.548148
GB	5.813131	DEU	3.962963
ID	8.247475	GBR	4.177778
FR	8.560607	DNK	6.274074
CH	10.26768	AUT	8.874074
KR	13.02222	NLD	10.33333
CA	16.14646	SWE	10.80741
JP	18.1404	BEL	11.01481
AU	18.43939	FRA	11.4963
BE	18.78788	ITA	12.02222
IE	18.82323	NOR	13.77037
GR	18.90088	USA	14.77778
ES	19.95834	CZE	15.54815
NL	20.63952	ESP	16.45926
SE	21.99495	HUN	16.64444
MX	24.22778	FIN	17.32593
RO	24.30357	ROU	23.14074
CZ	25.01389	CAN	23.23704
DK	25.33334	GRC	24.85926
AT	26.34443	PRT	27.76944
CN	27.74564	AUS	31.4418
IT	27.81881	CHN	32.52523
BR	28.00131	JPN	32.78214
NO	28.93723	KOR	32.812
IN	29.00292	MEX	32.88056
PT	29.16162	IDN	34.30324
FI	32.62067	SVN	35.55486
HU	33.80053	BRA	36.62247
SI	35.1996	IND	39

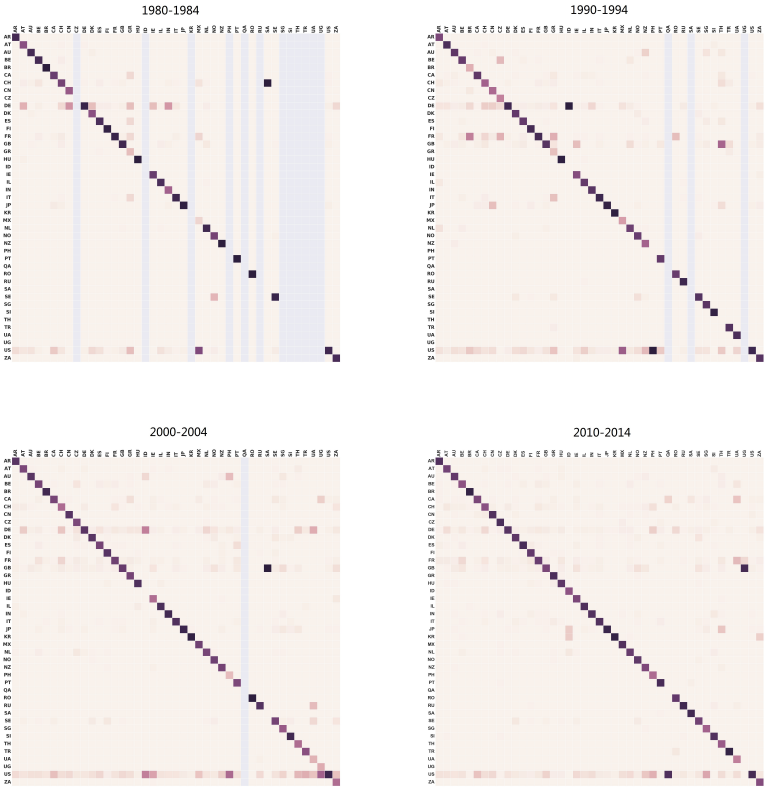
Table 10: Regression with Increased Lag of One Year

Dependent: CIN coreness ranking			1-year lag		1-year lag
		1b	1b+1	4b	4b+1
10	variable 9 weighted by the share of output in total trade volume	0.2501**	0.3675**	0.4679**	0.5658**
		(0.1107)	(0.1553)	(0.1699)	(0.2096)
3	Log of average outgoing citation quantity	0.0278	0.0019	0.0348	-0.0111
		(0.0314)	(0.0248)	(0.0452)	(0.0293)
4	Log of patent quantity	0.1239	0.2978**	0.2692	0.5459***
		(0.1359)	(0.1381)	(0.1587)	(0.1640)
5	Log of inventor quantity	-0.0319	-0.0000	0.0096	0.0988
		(0.1740)	(0.2106)	(0.1656)	(0.2712)
6	Log of BERD in pharmaceuticals			-0.2636	-0.4011
				(0.1538)	(0.2541)
7	Percentage of BERD performed in pharmaceuticals			0.2653	0.3040
				(0.2242)	(0.3048)
8	Dummy variables: Year	2000-2010	2000-2009	2000-2010	2000-2009
	Number of countries	29	29	20	20
	Number of observations	273.0000	248.0000	190.0000	172.0000
	R-squared	0.0783	0.1098	0.1557	0.2265

* p<0.10, ** p<0.05,*** p<0.01

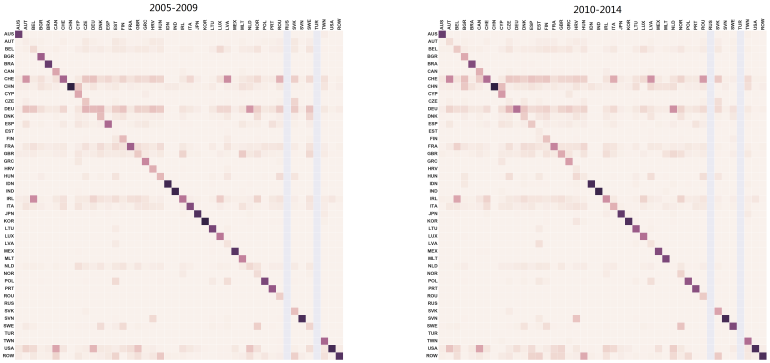
Note: Dependent variable is the CIN "Coreness". All the models are estimated by panel fixed-effect method. Estimations report coefficients and robust standard errors. By adding 1 year of lag, independent variables of year *t* are used to predict the dependent variable of the aggregate time window *t+1-t+5*.

Figure 19: Co-invention Networks Rendering of Multiple Time Periods



2D Rendering of Co-invention Networks in Four 5-year Time Windows

Figure 20: WIOD Networks Rendering of Multiple Time Periods



2D Rendering of WIOD Networks in Two 5-year Time Windows

Chapter 5

Conclusion

5.1 Conclusion

The author of this thesis is interested in a series of issues in technological innovation. Rather than the conventional indicators that measure only a single dimension, this thesis focuses on the information contained in connections, including the cohorting connections among different technologies, the citation connections among different technologies and countries, and the international collaboration connections among individual co-inventors. Therefore, the author takes a network perspective throughout the studies, develops and applies methodologies for appropriate research purposes.

This research is driven by the needs to rethink about today's technological innovation in a socio-economic context, in order to maintain the motivation for continuous innovation while optimizing the output of innovation investment. The first two chapters center around the issue of technology cohort as an indication of how emerging technologies reshape the global innovation landscape. Referenced with literature and results from conventional measures, the network approaches prove to be consistent with the empirical baselines and are more efficient to capture temporal changes, which is an interesting problem in networks studies. This is demonstrated in particular in Chapter 2, where the author de-

velops a systematic analysis based on some known good practices in literature. In addition to the methodological contribution, our results also suggest improvements to the authority-defined technology classification platform to provide more updated, accurate and thorough information for users with various purposes. The last chapter brings in more economic intuition. While more literature investigate whether innovation can empower production and increase profits, the author focuses on how innovation can be benefited when considering trading strategies in the pharmaceutical market. Although due to data limitation, the causality can't be solidly concluded, the analysis has found a consistent positive relationship between centralities in the trade network and the co-invention network.

Overall, this research reveals new ways to interpret patent data, to understand technological evolution, knowledge diffusion and the relationships between innovation and economics, and hopefully, to help decision makers optimize strategies in the globalization context.

This is a far-reaching goal and the relevant research field is vast. The studies presented in this thesis represent only a small part of this field. Further work remains abundant, including in-depth analysis of interesting technology sectors or regions, inclusion of improved and extended empirical data sets, and further studies of innovation in microeconomics. The thesis ends here, but the pursuit will continue.

References

[AAK16] Daron Acemoglu, Ufuk Akcigit, and William R Kerr. Innovation network. *Proceedings of the National Academy of Sciences*, 113(41):11483–11488, 2016. 25

[AB06] Christopher P Adams and Van V Brantner. Estimating the cost of new drug development: is it really $802 million? *Health affairs*, 25(2):420–428, 2006. 62

[ABB+05] Philippe Aghion, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt. Competition and innovation: An inverted-u relationship. *The Quarterly Journal of Economics*, 120(2):701–728, 2005. 62

[ACG18] Sam Arts, Bruno Cassiman, and Juan Carlos Gomez. Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84, 2018. 3, 19

[ACOTS12] Daron Acemoglu, Vasco M Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016, 2012. 63

[AG10a] Thomas Aynaud and Jean-Loup Guillaume. Static community detection algorithms for evolving networks. In *Modeling and optimization in mobile, ad hoc and wireless networks (WiOpt), 2010 proceedings of the 8th international symposium on*, pages 513–519. IEEE, 2010. 4, 67, 73

[AG10b] Thomas Aynaud and Jean-Loup Guillaume. Static community detection algorithms for evolving networks. In *Modeling and optimization in mobile, ad hoc and wireless networks (WiOpt), 2010 proceedings of the 8th international symposium on*, pages 513–519. IEEE, 2010. 27

[AI02] Daniele Archibugi and Simona Iammarino. The globalization of technological innovation: definition and evidence. *Review of International Political Economy*, 9(1):98–122, 2002. 62

[And99] Bob Anderton. Innovation, product quality, variety, and trade performance: an empirical analysis of germany and the uk. *Oxford economic papers*, 51(1):152–167, 1999. 5, 62, 74

[Ayn09] Thomas Aynaud. Community detection for NetworkX's documentation, 2009. 27

[Bav48] Alex Bavelas. A mathematical model for group structures. *Human organization*, 7(3):16–30, 1948. 39

[BBK08] Kevin Boyack, Katy Börner, and Richard Klavans. Mapping the structure and evolution of chemistry research. *Scientometrics*, 79(1):45–60, 2008. 7

[BDVR16] Nicholas Bloom, Mirko Draca, and John Van Reenen. Trade induced technical change? the impact of chinese imports on innovation, it and productivity. *The Review of Economic Studies*, 83(1):87–117, 2016. 62, 74

[Bea65] Murray A Beauchamp. An improved index of centrality. *Systems Research and Behavioral Science*, 10(2):161–163, 1965. 39

[BGLL08] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. 18, 27

[BIS07] Ann Bartel, Casey Ichniowski, and Kathryn Shaw. How does information technology affect productivity? plant-level comparisons of product innovation, process improvement, and worker skills. *The quarterly journal of Economics*, 122(4):1721–1758, 2007. 62

[BKB05] Kevin W Boyack, Richard Klavans, and Katy Börner. Mapping the backbone of science. *Scientometrics*, 64(3):351–374, 2005. 7

[BSZ08] Matteo Bugamelli, Fabiano Schivardi, and Roberta Zizza. The euro and firm restructuring. Technical report, National Bureau of Economic Research, 2008. 62

[Bun09] Die Bundesregierung. German Federal Government's National Electromobility Development Plan. *Development*, 2009. 39

[BW08] Mary Benner and Joel Waldfogel. Close to you? bias and precision in patent-based measures of technological proximity. *Research Policy*, 37(9):1556–1567, 2008. 25

[CS69] William S Comanor and Frederic M Scherer. Patent statistics as a measure of technical change. *Journal of political economy*, 77(3):392–398, 1969. 1, 25

[DC00] Patricia M Danzon and Li-Wei Chao. Cross-national price differences for pharmaceuticals: how large, and why? *Journal of health economics*, 19(2):159–195, 2000. 81

[DDD+17] Taro Daiko, Hélène Dernis, Mafini Dosso, Petros Gkotsis, Mariagrazia Squicciarini, Alexander Tuebke, Antonio Vezzani, et al. World top r&d investors: Industrial property strategies in the digital economy. Technical report, Joint Research Centre (Seville site), 2017. 63

[DGH16] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics*, 47:20–33, 2016. 62

[DHG03] Joseph A DiMasi, Ronald W Hansen, and Henry G Grabowski. The price of innovation: new estimates of drug development costs. *Journal of health economics*, 22(2):151–185, 2003. 62

[DHGL91] Joseph A DiMasi, Ronald W Hansen, Henry G Grabowski, and Louis Lasagna. Cost of innovation in the pharmaceutical industry. *J Health Econ*, 1991. 62

[DiM00] Joseph A DiMasi. New drug innovation and pharmaceutical industry structure: trends in the output of pharmaceutical firms. *Drug Information Journal*, 34(4):1169–1194, 2000. 62

[DiM01] Joseph A DiMasi. Risks in new drug development: approval success rates for investigational drugs. *Clinical Pharmacology & Therapeutics*, 69(5):297–307, 2001. 62

[DK04] Hélène Dernis and Mosahid Khan. Triadic patent families methodology. 2004. 10, 24, 63

[DM15] Jianwei Dang and Kazuyuki Motohashi. Patent statistics: A good indicator for innovation in China? Patent subsidy program impacts on patent quality. *China Economic Review*, 35:137–155, 2015. 25

[DTZ09] Eric David, Tony Tramontin, and Rodney Zemmel. Pharmaceutical r&d: the road to positive returns, 2009. 62

[EG10] Ellen Enkel and Oliver Gassmann. Creative imitation: exploring the case of crossindustry innovation. *R&d Management*, 40(3):256–270, 2010. 38

[FB07] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007. 35

[Fog07] Pasquale Foglia. Patentability search strategies and the reformed ipc: A patent office perspective. *World Patent Information*, 29(1):33–53, 2007. 9

[FR91] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991. 32

[Fre77] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977. 39

[Fre78] Linton C Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978. 28

[FRZ16] Gary D Ferrier, Javier Reyes, and Zhen Zhu. Technology diffusion on the international trade network. *Journal of Public Economic Theory*, 18(2):291–312, 2016. 63

[GH90] Gene M Grossman and Elhanan Helpman. Trade, innovation, and growth. *The American economic review*, 80(2):86–91, 1990. 62, 74

[Gre90] Christine Greenhalgh. Innovation and trade performance in the united kingdom. *the economic Journal*, 100(400):105–118, 1990. 5, 62

[Gri98] Zvi Griliches. Patent statistics as economic indicators: a survey. In *R&D and productivity: the econometric evidence*, pages 287–343. University of Chicago Press, 1998. 1, 25

[GS63] Zvi Griliches and Jacob Schmookler. Inventing and maximizing. *The American Economic Review*, 53(4):725–729, 1963. 25

[GSV05] Carmelo Giaccotto, Rexford E Santerre, and John A Vernon. Drug prices and research and development investment behavior in the pharmaceutical industry. *The Journal of Law and Economics*, 48(1):195–214, 2005. 81

[GTA17] GTAI. GTAI - Automotive Industry, 2017. 38

[GTW94] Christine Greenhalgh, Paul Taylor, and Rob Wilson. Innovation and export volumes and prices-a disaggregated study. *Oxford Economic Papers*, pages 102–135, 1994. 62

[GZKRng] Yuan Gao, Zhen Zhu, Raja Kali, and Massimo Riccaboni. Community evolution in patent networks: Technological change and network dynamics. forthcoming. 67, 73

[GZR17a] Yuan Gao, Zhen Zhu, and Massimo Riccaboni. Consistency and Trends of Technological Innovations: A Network Approach to the International Patent Classification Data. In *International Workshop on Complex Networks and their Applications*, pages 744–756. Springer, 2017. 23, 24, 25, 27, 28, 36, 42

[GZR17b] Yuan Gao, Zhen Zhu, and Massimo Riccaboni. Consistency and trends of technological innovations: A network approach to the international patent classification data. In *International Workshop on Complex Networks and their Applications*, pages 744–756. Springer, 2017. 67, 73

[H+05] Bronwyn H Hall et al. A note on the bias in herfindahl-type measures based on count data. *REVUE D ECONOMIE INDUSTRIELLE-PARIS-EDITIONS TECHNIQUES ET ECONOMIQUES-*, 110:149, 2005. 25

[HAS10] Christopher G Harris, Robert Arens, and Padmini Srinivasan. Comparison of ipc and uspc classification systems in patent prior art searches. In *Proceedings of the 3rd international workshop on Patent Information Retrieval*, pages 27–32. ACM, 2010. 8

[HJ14] Angela Hausman and Wesley J Johnston. The role of innovation in driving the economy: Lessons from the global financial crisis. *Journal of Business Research*, 67(1):2720–2726, 2014. 1, 25

[HJT01] Bronwyn H Hall, Adam B Jaffe, and Manuel Trajtenberg. The nber patent citation data file: Lessons, insights and methodological tools. Technical report, National Bureau of Economic Research, 2001. 25

[HJT05] Bronwyn H Hall, Adam Jaffe, and Manuel Trajtenberg. Market value and patent citations. *RAND Journal of economics*, pages 16–38, 2005. 11

[HSSC08] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008. 32

[HSV03] Dietmar Harhoff, Frederic M Scherer, and Katrin Vopel. Citations, family size, opposition and the value of patent rights. *Research policy*, 32(8):1343–1363, 2003. 11, 25

[JN12] Robert C Johnson and Guillermo Noguera. Accounting for intermediates: Production sharing and trade in value added. *Journal of international Economics*, 86(2):224–236, 2012. 64

[KWW14] Robert Koopman, Zhi Wang, and Shang-Jin Wei. Tracing value-added and double counting in gross exports. *American Economic Review*, 104(2):459–94, 2014. 64

[Kyl07] Margaret K Kyle. Pharmaceutical price controls and entry strategies. *The Review of Economics and Statistics*, 89(1):88–99, 2007. 81

[LDB08a] Renaud Lambiotte, J-C Delvenne, and Mauricio Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008. 18

[LDB08b] Renaud Lambiotte, J-C Delvenne, and Mauricio Barahona. Laplacian dynamics and multiscale modular structure in networks. *arXiv preprint arXiv:0812.1770*, 2008. 31, 35

[LDY+11] Ronald Lai, Alexander D' Amour, Amy Yu, Ye Sun, Vetle Torvik, and Lee Fleming. Disambiguation and Co-authorship Networks of the U . S . Patent Inventor Database. pages 1–38, 2011. 9

[Leo53] Wassily Leontief. Domestic production and foreign trade; the american capital position re-examined. *Proceedings of the American philosophical Society*, 97(4):332–349, 1953. 64, 68

[LW05] Kuei-Kuei Lai and Shiao-Jun Wu. Using the patent co-citation approach to establish a new patent classification system. *Information processing & management*, 41(2):313–330, 2005. 8

[MB09] Ronald E Miller and Peter D Blair. *Input-output analysis: foundations and extensions*. Cambridge University Press, 2009. 68

[MDL+17] Fabio Miranda, Harish Doraiswamy, Marcos Lage, Kai Zhao, Bruno Gonçalves, Luc Wilson, Mondrian Hsieh, and Cláudio T Silva. Urban Pulse: Capturing the Rhythm of Cities. *IEEE transactions on visualization and computer graphics*, 23(1):791–800, 2017. 22

[MDW+08a] Stéphane Maraut, Hélène Dernis, Colin Webb, Vincenzo Spiezia, and Dominique Guellec. The oecd regpat database. 2008. 64

[MDW+08b] Stéphane Maraut, Hélène Dernis, Colin Webb, Vincenzo Spiezia, and Dominique Guellec. The OECD REGPAT database: a presentation. *STI Working Paper*, 2008/2, 2008. 24, 26

[MJ05] J M Marie-Julie. Searching in flowchartsA PD toolsdoc pilot project at the borderline between text and image to access the most important features of an invention. *EPO internal publication*, 2005. 9

[MRM+10] Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *science*, 328(5980):876–878, 2010. 41

[Mun09] Bernard Munos. Lessons from 60 years of pharmaceutical innovation. *Nature reviews Drug discovery*, 8(12):959, 2009. 62

[OEC09] OECD. *OECD Patent Statistics Manual*. OECD PUBLICATIONS, France, 1 edition, 2009. 10, 24, 26

[OEC17a] OECD. Intellectual property (ip) statistics and analysis, 2017. 63

[OEC17b] OECD. OECD patent databases - OECD, 2017. 9, 26

[Pic11] Carlo Piccardi. Finding and testing network communities by lumped Markov chains. *PLoS ONE*, 6(11), 2011. 3, 12, 27, 67

[PMR11] Fabio Pammolli, Laura Magazzini, and Massimo Riccaboni. The productivity crisis in pharmaceutical r&d. *Nature reviews Drug discovery*, 10(6):428, 2011. 62

[PRTZ17] Carlo Piccardi, Massimo Riccaboni, Lucia Tajoli, and Zhen Zhu. Random walks on the world input–output network. *Journal of Complex Networks*, 6(2):187–205, 2017. 64, 68

[RL02] Stephen Roper and James H Love. Innovation and export performance: evidence from the uk and german manufacturing plants. *Research policy*, 31(7):1087–1102, 2002. 5, 62, 74

[RP07] Massimo Riccaboni and F Pammolli. *Innovation and industrial leadership: lessons from pharmaceuticals*. CENTER FOR TRANSATLANTIC RELATIONS, JOHNS HOPKINS, 2007. 62

[Sag02] Kamal Saggi. Trade, foreign direct investment, and international technology transfer: A survey. *The World Bank Research Observer*, 17(2):191–235, 2002. 62

[Sch08] Ulrich Schmoch. Concept of a technology classification for country comparisons. *Final Report to the World Intellectual Property Organization (WIPO), Fraunhofer Institute for Systems and Innovation Research, Karlsruhe*, 2008. 63

[sd17] WIPO statistics database. Statistical Country Profiles_Germany, 2017. 31, 37, 38

[SDC13a] Mariagrazia Squicciarini, Hélène Dernis, and Chiara Criscuolo. Measuring patent quality. 2013. 1, 74, 76

[SDC13b] Mariagrazia Squicciarini, Hélène Dernis, and Chiara Criscuolo. Measuring Patent Quality: Indicators of Technological and Economic Value. *OECD Science, Technology and Industry Working Papers*, (03):70, 2013. 8, 11, 25

[SJR+13] Massoud Seifi, Ivan Junier, Jean-Baptiste Rouquier, Svilen Iskrov, and Jean-Loup Guillaume. Stable community cores in complex networks. In *Complex Networks*, pages 87–98. Springer, 2013. 28

[SP07] Melissa A Schilling and Corey C Phelps. Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. *Management science*, 53(7):1113–1126, 2007. 63

[TDL+15] Marcel P Timmer, Erik Dietzenbacher, Bart Los, Robert Stehrer, and Gaaitzen J De Vries. An illustrated user guide to the world input–output database: the case of global automotive production. *Review of International Economics*, 23(3):575–605, 2015. 63

[THJ97] Manuel Trajtenberg, Rebecca Henderson, and Adam Jaffe. University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and new technology*, 5(1):19–50, 1997. 11

[TLSdV16] Marcel Timmer, Bart Los, Robert Stehrer, and Gaaitzen de Vries. An anatomy of the global trade slowdown based on the wiod 2016 release. Technical report, Groningen Growth and Development Centre, University of Groningen, 2016. 64

[VCLC08] Thomas W Valente, Kathryn Coronges, Cynthia Lakon, and Elizabeth Costenbader. How correlated are network centrality measures? *Connections (Toronto, Ont.)*, 28(1):16, 2008. 28

[VHAA+12] Victor Veefkind, J Hurtado-Albir, S Angelucci, K Karachalios, and N Thumm. A new EPO classification scheme for climate change mitigation technologies. *World Patent Information*, 34(2):106–111, 2012. 9

[Wak98a] Katharine Wakelin. Innovation and export behaviour at the firm level. *Research policy*, 26(7-8):829–841, 1998. 5, 62

[Wak98b] Katharine Wakelin. The role of innovation in bilateral oecd trade performance. *Applied Economics*, 30(10):1335–1346, 1998. 62

[Wan12] Qinna Wang. Overlapping community detection in dynamic networks, 2012. 28

[WE16] J John Wu and Stephen J Ezell. How national policies impact global biopharma innovation: A worldwide ranking. Technical report, Information Technology & Innovation Foundation, 2016. 76, 81

[WF94] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994. 28

[WF10a] Qinna Wang and Eric Fleury. Mining time-dependent communities. In *LAWDN-Latin-American Workshop on Dynamic Networks*, pages 4–p, 2010. 4, 73

[WF10b] Qinna Wang and Eric Fleury. Mining time-dependent communities. In *LAWDN-Latin-American Workshop on Dynamic Networks*, pages 4–p, 2010. 28

[WF13] Qinna Wang and Eric Fleury. Overlapping community structure and modular overlaps in complex networks. In *Mining Social Networks and Security Informatics*, pages 15–40. Springer, 2013. 28

[WIP16a] WIPO. IPC 2016.01, 2016. 9, 25, 36

[WIP16b] WIPO. IPC Definitions_20160101, 2016. 9

[WIP17a] WIPO. About the International Patent Classification, 2017. 8, 24

[WIP17b] WIPO. Guide to the International Patent Classification, 2017. 15, 24

[YP04] Byungun Yoon and Yongtae Park. A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, 15(1):37–50, 2004. 9

[YRBY05] Heahyun Yoo, Chandra Ramanathan, and Cynthia Barcelon-Yang. Intellectual property management of biosequence information from a patent searching perspective. *World Patent Information*, 27(3):203–211, 2005. 9

[ZXT10] S Xie Zeng, Xue M Xie, and Chi Ming Tam. Relationship between cooperation networks and innovation performance of smes. *Technovation*, 30(3):181–194, 2010. 63