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

Yuan Gao, Zhen Zhu, Massimo Riccaboni

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 measuring 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.

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.

Yuan Gao
IMT School for Advanced Studies Lucca, Lucca, Italy, e-mail: yuan.gao@imtlucca.it

Zhen Zhu
IMT School for Advanced Studies Lucca, Lucca, Italy e-mail: zhen.zhu@imtlucca.it

Massimo Riccaboni
IMT School for Advanced Studies Lucca, Lucca, Italy; Department of Managerial Economics, Strategy and Innovation, Katholieke Universiteit Leuven, Leuven,Belgium e-mail: massimo.riccaboni@imtlucca.it

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[4, 3]. 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 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 [23] 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[18].

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 [9]. Lai and Wu proposed a new classification system based on co-citations to replace IPC or USPC and improve categorization accuracy [10]. 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 [22] 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 [6, 13, 20, 25, 26]. 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.

2.1 Data

The data used in this analysis is mainly retrieved from the February, 2016 edition OECD patent database [16], 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 [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 [21].

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 [15, 5]. 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.

3 Network Construction

3.1 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.

3.2 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 [7]. 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.

3.3 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 [19, 8, 7, 18]. 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 [18].

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.

4 Network Analysis Methods

4.1 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 [17]. 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

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} \tag{1}$$

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

$$U_{cc} = \frac{\sum_{i,j \in C_c} \pi_i p_{ij}}{\sum_{i \in C_c} \pi_i} \tag{2}$$

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.

4.2 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

$$C_{ic} = \frac{1}{\frac{\bar{d}_{ij}n}{n-1}} U_{cc} = \frac{n-1}{\bar{d}_{ij}n} U_{cc}, \tag{3}$$

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.

4.3 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.

5 Results

5.1 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.

5.2 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 [24].

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.

5.3 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.

6 Verification and Discussion

6.1 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 [2]. 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 [12], 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.

6.2 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.

6.3 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 [1]. 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. 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.

7 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" [14] could be considered for continued efforts in the future.

References

1. Arts, S., Cassiman, B., Gomez, J.C., Cassiman, B., Gomez, J.C.: Text matching to measure patent similarity (2017)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10,008 (2008)
3. Boyack, K., Börner, K., Klavans, R.: Mapping the structure and evolution of chemistry research. Scientometrics **79**(1), 45–60 (2008)
4. Boyack, K.W., Klavans, R., Börner, K.: Mapping the backbone of science. Scientometrics **64**(3), 351–374 (2005)
5. Dernis, H., Khan, M.: Triadic Patent Families Methodology (2004). DOI 10.1787/443844125004. URL http://dx.doi.org/10.1787/443844125004
6. Foglia, P.: Patentability search strategies and the reformed IPC: A patent office perspective. World Patent Information **29**(1), 33–53 (2007)
7. Hall, B.H., Jaffe, A., Trajtenberg, M.: Market value and patent citations. RAND Journal of economics pp. 16–38 (2005)
8. Harhoff, D., Scherer, F.M., Vopel, K.: Citations, family size, opposition and the value of patent rights. Research policy **32**(8), 1343–1363 (2003)
9. Harris, C.G., Arens, R., Srinivasan, P.: Comparison of IPC and USPC classification systems in patent prior art searches. In: Proceedings of the 3rd international workshop on Patent Information Retrieval, pp. 27–32. ACM (2010)
10. Lai, K.K., Wu, S.J.: Using the patent co-citation approach to establish a new patent classification system. Information processing & management **41**(2), 313–330 (2005)
11. Lai, R., D' Amour, A., Yu, A., Sun, Y., Torvik, V., Fleming, L.: Disambiguation and Co-authorship Networks of the U . S . Patent Inventor Database pp. 1–38 (2011)
12. Lambiotte, R., Delvenne, J.C., Barahona, M.: Laplacian dynamics and multiscale modular structure in networks. arXiv preprint arXiv:0812.1770 (2008)
13. Marie-Julie, J.M.: 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)
14. Miranda, F., Doraiswamy, H., Lage, M., Zhao, K., Gonçalves, B., Wilson, L., Hsieh, M., Silva, C.T.: Urban Pulse: Capturing the Rhythm of Cities. IEEE transactions on visualization and computer graphics **23**(1), 791–800 (2017)
15. OECD: OECD Patent Statistics Manual, 1 edn. OECD PUBLICATIONS, France (2009). DOI 10.1787/9789264056442-en. URL http://www.oecd-ilibrary.org/science-and-technology/oecd-patent-statistics-manual_9789264056442-en
16. OECD: OECD patent databases - OECD (2017). URL http://www.oecd.org/sti/inno/oecdpatentdatabases.htm
17. Piccardi, C.: Finding and testing network communities by lumped Markov chains. PLoS ONE **6**(11) (2011). DOI 10.1371/journal.pone.0027028
18. Squicciarini, M., Dernis, H., Criscuolo, C.: Measuring Patent Quality: Indicators of Technological and Economic Value. OECD Science, Technology and Industry Working Papers (03), 70 (2013). DOI 10.1787/5k4522wkw1r8-en. URL http://www.oecd-ilibrary.org/science-and-technology/measuring-patent-quality_5k4522wkw1r8-en
19. Trajtenberg, M., Henderson, R., Jaffe, A.: University versus corporate patents: A window on the basicness of invention. Economics of Innovation and new technology **5**(1), 19–50 (1997)
20. Veefkind, V., Hurtado-Albir, J., Angelucci, S., Karachalios, K., Thumm, N.: A new EPO classification scheme for climate change mitigation technologies. World Patent Information **34**(2), 106–111 (2012)
21. WIPO: IPC 2016.01 (2016). URL http://www.wipo.int/classifications/ipc/en/ITsupport/Version20160101/
22. WIPO: IPC Definitions_20160101 (2016). URL http://www.wipo.int/ipc/itos4ipc/ITSupport_and_download_area/
23. WIPO: About the International Patent Classification (2017). URL http://www.wipo.int/classifications/ipc/en/preface.html
24. WIPO: Guide to the International Patent Classification (2017). URL http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf
25. Yoo, H., Ramanathan, C., Barcelon-Yang, C.: Intellectual property management of biosequence information from a patent searching perspective. World Patent Information **27**(3), 203–211 (2005)
26. Yoon, B., Park, Y.: A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research **15**(1), 37–50 (2004)

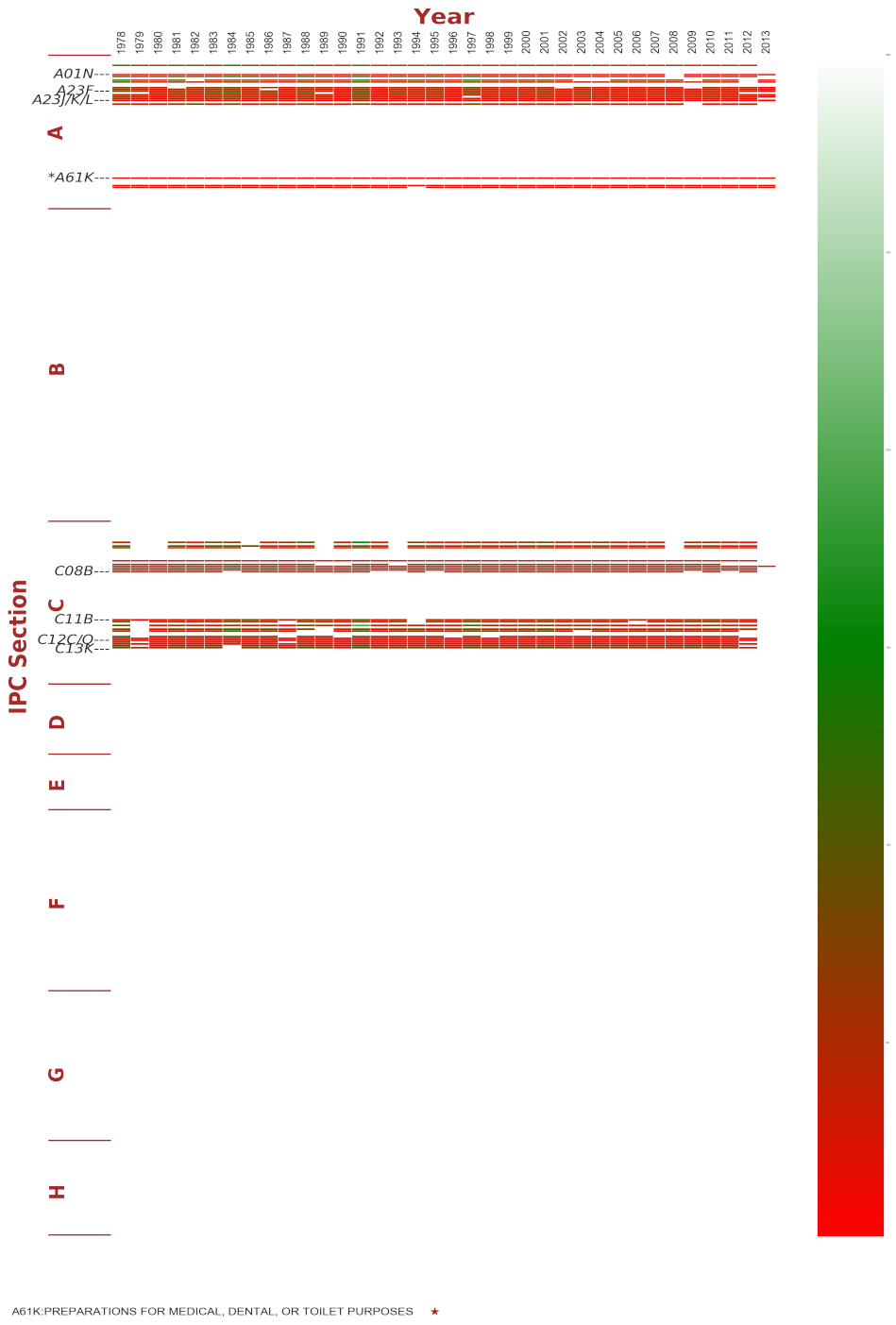

Fig. 1: Consistency of the Endogenous Community of A61K Based on Family-Cohort Network

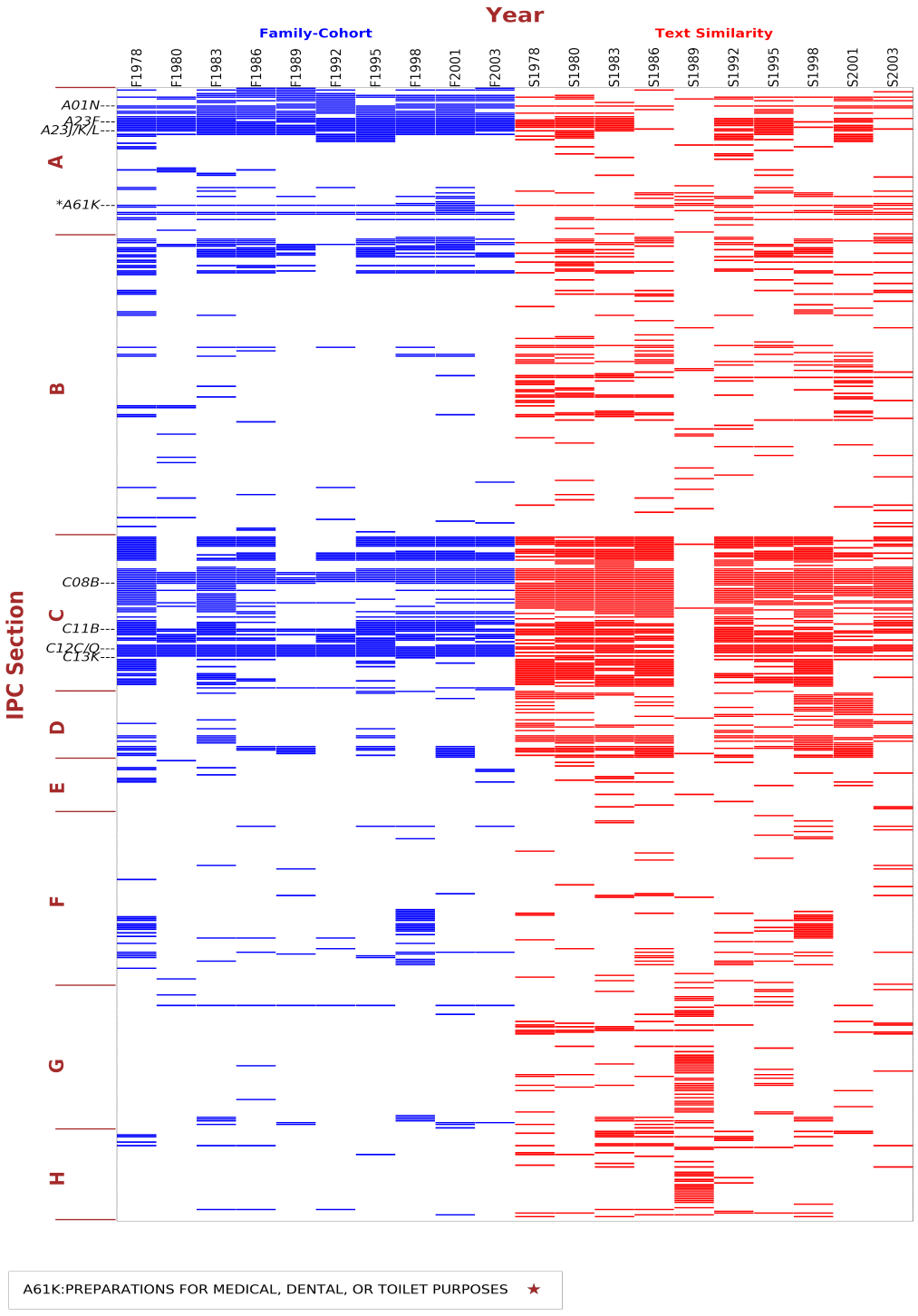

Fig. 2: Comparison of Family-Cohort Network and Text-Match Similarity Network for A61K