Networks with Hierarchical Structure: Applications to the Patent Domain

Nikolai Nefedov

Abstract In this paper we introduced a graph-based metric to measure a similarity between weighted sets of classifications codes defined as nodes on hierarchical taxonomy trees. We applied this metric to build relationship networks among companies and to find company peers (communities) in IPR (intellectual-property rights) domain based on patent portfolios.
To characterize evolution of patent portfolios for companies we used weighted sets of international patent classification codes (IPC), where each IPC weight corresponds to a number of IPC codes in a company patent portfolio aggregated to a given hierarchy level over a given period of time.
We used the suggested graph-based similarity at different hierarchical IPC levels to build corresponding networks and detected communities over different time periods. To track communities evolution in time we developed a cluster-matching algorithm to align community labels over time. Then we study evolution of communities in time to identify changes in a company strategy and its peers at the given time.
The suggested methodology may be applied to other domains that include hierarchical classification sets such as trademarks, legal documents, scientific papers, lawsuits etc.

1 Introduction

1.1 Patent networks

Patent network analysis is widely used to identify technology trends and formulate a technology strategy of a company, e.g., [1]. Typically patent networks are built using relationships among individual patents based on patent citations [2] or text analysis of patent abstracts, specifications, claims etc [3]. Recently patent text analytics is

Nikolai Nefedov (e-mail: nikolai.nefedov@thomsonreuters.com)✉
Thomson Reuters Labs, Switzerland and Swiss Federal Institute of Technology, Zurich (ETHZ), Switzerland

H. Cherifi et al. (eds.), *Complex Networks & Their Applications V*,
Studies in Computational Intelligence 693,
DOI 10.1007/978-3-319-50901-3_60

extended by using weighted keyword-based patent networks [4]. These methods usually are based on pairwise comparison of single patents complimented with total amount of patents in different technology sectors that allows to identify technology trends. On the other hand, in order to formulate a company strategy it is also important to know about activities of competing companies (peers) in relevant technology domains.

Finding company peers implies a comparison of profiles of companies and several attempts have been made to create company profiles or "fingerprints" reflective of assets and endeavors of the company. This may be done in several dimensions, e.g., fingerprint dimensions may include patent portfolio, trademarks, as well as products, fundamentals, geography, market associations, etc. At such fingerprints different taxonomy schemes (e.g., sets of classification codes) are widely used to describe dimensions. In this paper we address only patent portfolio domain.

Comparison of companies in IPR domain requires comparison of patent portfolios which include different amount of patents (patent weights) in different (and not necessarily overlapping) IPC categories. Besides, companies may have large patent portfolio volumes that makes difficult to differentiate and identify changes of topics using patents citations or patents text analytics. In this paper we used hierarchical International Patent Codes (IPC)[5] that are assigned by patent examiners and cover content of patents in more than 100 countries. Currently hierarchical IPC codes contain 8 sections (one letter), 130 classes (2-digit number), 639 subclasses (one letter), 7434 groups with 65152 subgroups (one-to-three digit number). In the following we refer these hierarchy levels h_k by number of symbols they contain, i.e., IPC1, IPC3, IPC4, IPC7. To compare patent portfolios we need to define a similarity between weighted sets of hierarchical objects.

1.2 Similarity measures

Similarity is widely used concept and many similarity measures have been suggested [6]. For example, a semantic measure in an IS-A taxonomy based on a shared information content of the shortest common distance between two words/concepts in a lexical taxonomy is proposed in [7, 8]. As its generalization, an universal definition of similarity from information theory point of view was developed in [9]. However, these concepts mainly address a similarity between single objects, while to compare patent portfolios we need to define similarity between sets of weighted hierarchical elements. On the other hand, methods to calculate similarity between sets of objects typically do not take hierarchy into account (e.g., cosine similarity).

In this paper we propose a similarity measure to compare weighted sets of hierarchical objects and applied it for patent portfolios comparison. The proposed similarity measure allowed us to present relations among objects, e.g. companies, as a *connected* graph; it is hardly possible with other types of similarity such as cosine similarity. Then we applied network analysis to find peers and analyze peers evolution in time. Also, the proposed method allows us to map activities of companies on a connected technology map to provide a view on a broader technology evolution.

The paper is organized as follows: Section 2 outlines a graph-based metric to compare weighted hierarchical sets. In Section 3 we built patent portfolio evolution for a number of companies at different hierarchical IPC levels. Next we used the suggested metric to calculate pairwise similarities between companies in IPR domain at different hierarchical levels followed by construction of corresponding networks and their evolution in time. To find peers (communities) we applied community detection methods [10, 11, 12] at different IPC hierarchical levels h_k for different years (2008-2014). To track communities evolution in time we developed a cluster-matching algorithm to align community labels over time based on [13]. Finally, we analyzed evolution of communities in time to identify changes in a company strategy and its peers at a given time.

2 Comparison of weighted hierarchical sets

2.1 Preliminaries

Let's consider a set C of objects c_i, where $|C| = N_c$ is a total number of objects. Relations between objects $\{c_i, c_j\}$ may be presented as a weighted undirected graph $G(C,E,\mathbf{S})$, where $E = \{e_{il}\}$ is a set of edges $e_{ii} \in \{0,1\}$ and $\mathbf{S}$ is a similarity matrix, $s_{i,j} = s(c_i,c_j) \in \mathbf{S}$, , $i,j = 1,\ldots,N_c$, is similarity between c_i and c_j. Hierarchical attributes for a given object c_i may be presented as a tree $T_i(\mathbf{a}(h_k))$, where c_i is the the root and attributes $\mathbf{a}(h_k)$ are nodes of c_i on the tree at a hierarchical level h_k. As an example, let's consider objects c_1 and c_2 with attributes taken from a set $\mathbf{a} = \{A,B,C,D,E,F,G,H\}$ corresponding to IPC1 as shown at Fig. 1. Similarity between objects c_i and c_j (shown by dashed lines) usually is defined as a function of intersection of corresponding subsets $\mathbf{a}(c_i)$ and $\mathbf{a}(c_j)$, e.g., $s(c_i,c_j) = f|\cap(\mathbf{a}(c_i),\mathbf{a}(c_j))|$ (cf. Fig. 2).

In the following we will call relations graph $G(C,E,\mathbf{S})$ as a network to avoid confusion with graphs presenting taxonomy trees T_i.

2.2 Weighted taxonomy trees

Figure 2 illustrates the suggested approach to define relationships between objects c_1 and c_2 with weighted hierarchical attributes at levels IPC1, IPC3 and IPC4. In case of patent portfolios, weights $w_n(h_k)$ may present a number of IPC codes aggregated to level h_k within considered IPC class (B02F, B02,B at Fig.2). Let's assume that objects c_1 and c_2 have, among others, patents in IPC code B02F, Fig. 2. Then this IPC category contributes to similarity $s(c_1,c_2)$ at three hierarchical levels {B, B02, B02F} (see dashed lines between c_1 and c_2) such that the deeper we go down on the tree, the higher similarity is: $s(c_1,c_2,h_1) < s(c_1,c_2,h_2) < s(c_1,c_2,h_3)$. For example, if we compare IPC classes B02G and B02F, then for these codes only 2 layers {B02,B} contribute to similarity; note no similarity between B0G2 and F04.

Generalization to weighed hierarchical sets and its applications is briefly outlined below. In particularly, a patent portfolio for a company c_j may be presented as a set of tuples $P_j(h_k) = \{a_i(h_k), w(a_i(h_k))\}$, where $a_i(h_k) = IPC_i(h_k)$ is the i-th IPC code in patent portfolio at the k-th hierarchy level, $w(a_i(h_k))$ is its weight, $i \in N_j(h_k)$ is a number of different IPCs in $P_j(h_k)$. In our case $w(a_i(h_k))$ is a number of IPCs aggregated from all patents containing $IPC_i(h_k)$ code. Note that since there may be multiple IPCs characterizing a single patent, this definition applies both to patent portfolios and to single patents. In the following we call tuples $P_j(h_k)$ as aggregated IPCs at the level h_k. For example, patent portfolios aggregated to h_k = 3 level and sorted by weight for companies c_1 ='Samsung Electronics' and c_2 ='Panasonic' are presented as $P_1(3) = \{\{G06F, 10251\}, \{H04N, 7800\}, \{H01L, 6634\}, \ldots\}$. and $P_2(3) = \{\{H04N, 5920\}, \{G06F, 4989\}, \{H01M, 2616\}, \ldots\}$, respectively.

2.3 Similarity between weighted hierarchical sets

Typically methods to calculate similarity (e.g., cosine similarity) do not take hierarchy into account. For example, cosine similarity between patents having rather similar IPC codes A01B11 and A01B12 is zero. Similar to patent portfolios comparison, the problem exists in patent to patent comparison since even a single parent may be categorized by a set of IPC codes. Furthermore, it is not clear how to take into account weights at different hierarchical levels and define a normalization to compare *weighted sets* of hierarchical classification codes, such as patent portfolios with multiple IPCs. In this section we briefly outline the proposed method to compare weighted sets of hierarchical objects where sets have the same cardinality. More detailed generic description of the proposed method is rather involved and to appear elsewhere.

Let's define $p(a_l, c_i) = [a_l, \ldots, c_i] = p(a_l^i)$ as a sequence of nodes on T_i forming the shortest path from node a_l to root c_i. Then we may define a similarity s between nodes a_l and a_m as a number of common nodes between paths $p(a_l^i)$ and $p(a_m^j)$:

$$s(a_l, a_m) = s(p(a_l^i), p(a_m^j)) = \left| \bigcap (p(a_l^i), p(a_m^j)) \right| . \tag{1}$$

Clearly, $s(a_l, a_l) = |p(a_l)|$ corresponds to a number of hierarchical levels on the path from a_l to the root on T_i. Similarly, $s(a_l, a_m)$ may be seen as a number d of shared hierarchical levels or a distance $d(a_l, a_m)$ on T. In this settings s is a linear function of d. On the other hand, for irregular trees such as IPCs taxonomy, contributions to similarity may not necessary depend linearly on h_k. To take this property into account we included function $f(h_k)$ into the normalization below. Recall that the longer a classification code, the more information it provides, i.e., $s(h_k)$ is a monotonically increasing function of h_k.

Let **a** and **b** be portfolios for companies c_1 and c_2. Then a normalized similarity s_n between two codes from **a** and **b** on the same taxonomy tree may be written as

$$s_n(a_l(h_k), b_m(h_k)) = \frac{s(a_l(h_k), b_m(h_k))}{f(h_k)} \,. \tag{2}$$

It may be shown that a normalized similarity between unweighted hierarchical sets **a** and **b** at level h_k may be presented as below

$$s_n(\mathbf{a}, \mathbf{b}, h_k) = \frac{1}{C_{max}} \sum_{l}^{N} \sum_{m}^{N} s_n(a_l(h_k), b_m(h_k), f(h_k)) \,, \tag{3}$$

where

$$C_{max} = 1 + (N-1) f(h_{max} - 1) \Big/ f(h_{max}) \,. \tag{4}$$

A normalized similarity between weighted hierarchical sets **a** and **b** (patent portfolios) aggregated to h_k level may be written as

$$s_n^{(w)}(\mathbf{a}, \mathbf{b}, h_k) = \frac{1}{C_{max}^{(w)}(f, N, h_{max})} \sum_{l}^{N} \sum_{m}^{N} \Phi\left(w_l^{(a)} w_m^{(b)}, W^{(a)}, W^{(b)}\right) s_n(a_l, b_m, f(h_k)) \,. \tag{5}$$

Note that there may be different ways to define function $\Phi()$. For example, by applying the same methodology as in (1) for weights we may derive a weighting symmetric function as below

$$\Phi\left(w_l^{(a)}, w_m^{(b)}, W^{(a)}, W^{(b)}\right) = \min\left(\frac{w_l^{(a)}}{W^{(a)}}, \frac{w_m^{(b)}}{W^{(b)}}\right) \tag{6}$$

$$W^{(i)}(h_k) = \sum_{m} w_m^{(i)}(h_k), \quad i = a, b \tag{7}$$

The max similarity in (Eq.5) is reached when all IPC codes in both portfolios are located in the same IPC class at the lowest hierarchy level.

This methodology may be extended to comparison of two ontologies with a difference that instead of a single underlying tree as in the case above, there may be several (or a forest of) underlying trees. It implies that mapping of ontology objects and similarity calculations should be aggregated over relevant subsets of underlying trees.

3 Patent portfolios comparison

3.1 Evolution of patent portfolios

Companies change direction and enter new areas of technology and may cease operating in long-involved areas of technology. In this section we analyzed evolution of company patent portfolios at different hierarchical levels to detect changes in a company activities. As a data source we used Derwent Patents Database [14]

available via *Thomson Innovation*[15] and built patent portfolios for 10^5 companies covering totaly about 3×10^6 patent families registered in the USA during period 2008-2014.

As an example, Fig. 3 shows IBM patent portfolio evolution at different hierarchical IPC levels over time. Here colors correspond to different IPC codes for patents within IBM patent portfolio, labels on side color-bars indicate patents mapping to the highest hierarchical level IPC1. The absence of patents in a particular IPC category is denoted by blue color (black color in paper version). For example, one can notice a blue color line during 2010-2014 at Fig. 3a (y=19 corresponds to IPC3 = G07) and Fig. 3b (y=38,39 correspond to IPC4 = G07C,G07F). It indicates that IBM stopped patent activity in measurement equipment for registering tokens. On the other hand, from 2010 there is growing activity in IPC3=B81 (y=8 at Fig. 3a) corresponding to nano-technology, in particular, in field of manufacturing of devices and systems on substrate IPC4=B81C (Fig. 3b).

Note that new trends may not easily be observed at a very course or a very granular hierarchy levels, so we used cross-level analysis to detect changes and then digging for more details.

3.2 Networks evolution

Networks are dynamic and changing over time with some companies becoming peers and other peer companies losing the association as a peer company due to a number of reasons. Over time companies enter the competitive landscape and fall out of the landscape. Dynamic network analysis and models to describe evolution of communities are under intensive studies, in particular, in social networks (SN) domain, e.g., [16, 17, 18]. In this paper we do not consider models for SN communities evolution, but primary addressing a discovery mode to look for disruptive changes that modifies competition profile in IPR domain.

In particular, given sets of classification codes (e.g., IPC-based patent portfolio) defined on the same classification tree we analyzed peers (communities) evolution using the following steps:

(a) define graph-based similarity metric as a function of distance between nodes on the underlying classification tree;
(b) calculate pair-wise similarity between nodes by mapping nodes (IPCs) from different portfolios to the underlying classification tree (see Eq.2);
(c) calculate similarity metric between sets of weighted classification codes (e.g., general case Eq.5, examples Eq.6, Eq.7) and build network snapshots for different time periods;
(d) apply community detection algorithms to network snapshots to find stable communities based on random walk [12] within each time snapshot;
(e) build a reference network by aggregating all network snapshots over time and applied community detection algorithms to find communities within;

(f) use aggregated community labels as a reference and matched community labels from different network snapshots to the reference community labels;
(g) steps above allow us to analyze communities evolution over time, detect company peers at given time and predict new trends.

Fig. 4 shows a network example built using 10 IPC codes with largest weights in each patent portfolio for the top 300 companies with largest patent portfolio volumes. We found that the suggested method results in a connected network, but for visualization purposes Fig. 4 shows only 5% of largest similarity values. As one can see, even under this simplification, the suggested method results in several connected clusters which allows to find mapping to technology areas and its relations. Also it easy to detect companies which are active in several technological areas, such as 'Siemens', 'Samsung', 'Hitachi Chemical' and 'Funai Electric'.

Fig. 5 presents an example of evolution of peer communities in time before (on the left) and after (on the right) community labels matching for the top 100 companies with the highest patent portfolios volumes. The first column on the left on both figures shows references for communities matching. All nodes (company IDs) in time snapshots are grouped according to the reference layer grouping.

As one can see from Fig. 5b, the largest part of competitive landscape stays mainly stable (shown by yellow in online version), while some companies are moving or exploring other technology domains. At the same time one group of companies (green in online version) keeps investing in another technology domain (orange in online version) in 2009 and 2013, while staying in its main domain the other time.

4 Conclusion

In this paper we propose a similarity measure to compare weighted sets of hierarchical objects. As an example, we consider company patent portfolios characterized by hierarchical IPC codes. Using the suggested similarity measure we build network snapshots for different time periods and applied network analysis to find company peers in IPR domain. It allows us to study peers evolution at different hierarchical levels and find changes in competitive landscape. The suggested methodology may be applied to other domains that include hierarchical classifications.

Acknowledgements This work was supported by Thomson Reuters Global Resources. The author would like to thank anonymous reviewers for comments and pointing to missing references.

References

[1] Valverde S. et al, Topology and Evolution of Technology Innovation Networks. Phys. Rev. E 76, 056118 (2007).
[2] Verspagen B., Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. Advances in Complex Systems, 10(1), 93–115 (2007).
[3] Yoon B. et al, A systematic approach for identifying technology opportunities: keyword-based morphology analysis. Technol. Forecast. 72, 145–160, Elsevier (2005).

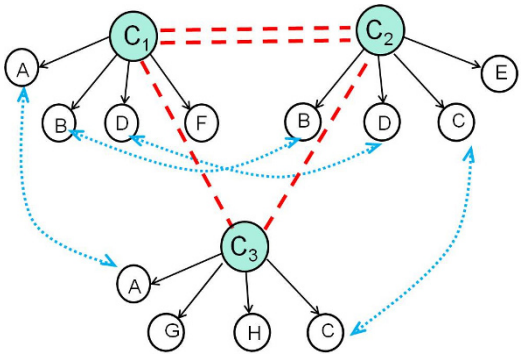

Fig. 1: Nodes with attributes.

[4] Lee S. et al, An approach to discovering new technology opportunities: Keyword-based patent map approach. Technovation 29(6), 481–497, Elsevier (2009).
[5] International Patent Classification, http://www.wipo.int/classifications/ipc/en/
[6] Cha S–H., Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. J. Math. Models and Methods in Applied Sci. 1(4), 300–307 (2007).
[7] Resnik P., Using information content to evaluate semantic similarity in a taxonomy. Proceedings IJCAI, 448-453 (1995).
[8] Resnik P., Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. J. Artif. Intel. Res. 11, 95–130 (1999).
[9] Lin D., An Information-Theoretic Definition of Similarity. Proc. Int. Conf. on Machine Learning, 296–304 (1998).
[10] Newman MEJ, Fast algorithm for detecting community structure in networks. Physical Review E 69, 066133 (2004).
[11] Blondel V. et al, Fast unfolding of communities in large networks. J. Stat. Mech.: Theory and Experiment, 1742-5468 (10), P10008+12 (2008).
[12] Lambiotte R. et al, Laplacian Dynamics and Multiscale Modular Structure in Networks. ArXiv:0812.1770v3.
[13] Nefedov N., Analysis of Communities Evolution in Dynamic Social Networks. Studies in Computational Intelligence: Complex Networks IV, 476, 39–46, Springer (2013).
[14] DWPI: http://ipscience.thomsonreuters.com/product/derwent-world-patents-index-dwpi
[15] http://ipscience.thomsonreuters.com/product/thomson-innovation
[16] Palla G. et al, Quantifying social group evolution. Nature 446, April, 664–667 (2007).
[17] Lin Y-R et al, Analyzing Communities and Their Evolutions in Dynamic Social Networks, ACM Trans on Knowledge Discovery from Data. 3(2), 8 (2009).
[18] Brodka P. at al, GED: the Method for Group Evolution Discovery in Social Networks, Soc. Netw. Anal. Min. 3(1), 1-14 (2013).

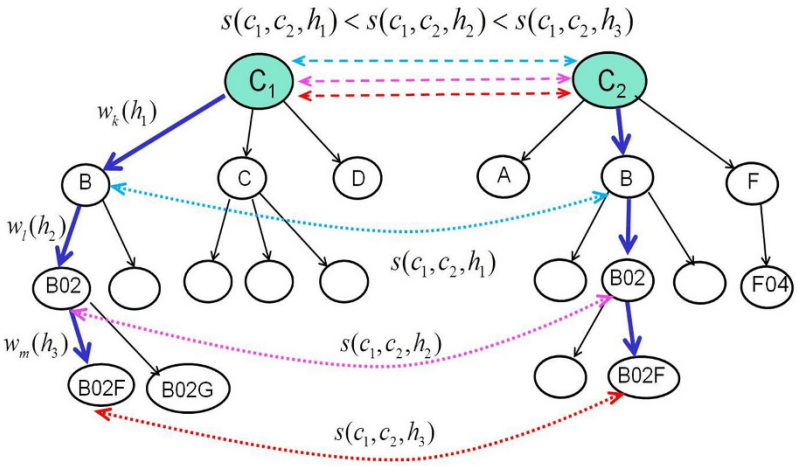

Fig. 2: IPCs as taxonomy trees.

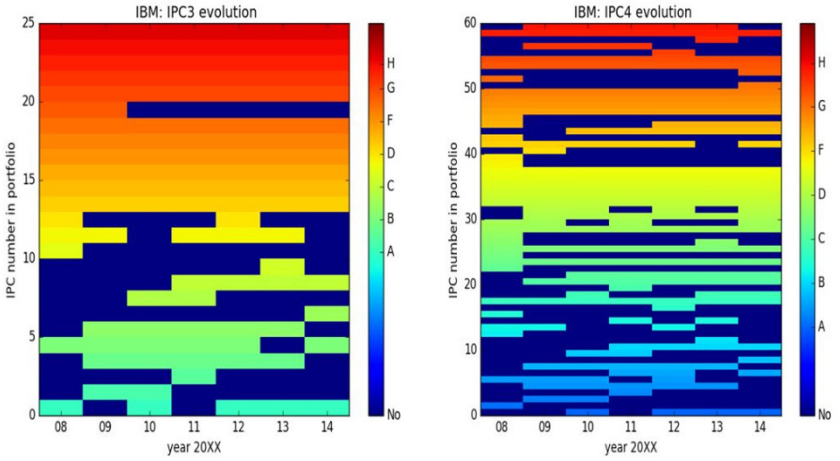

Fig. 3: Example of patent portfolios evolution in time at difference hierarchy levels. Company: IBM; hierarchical levels IPC3 (left, *a*) and IPC4 (right,*b*). Colored bars indicate mapping to the highest hierarchical level IPC1 (colored figures online).

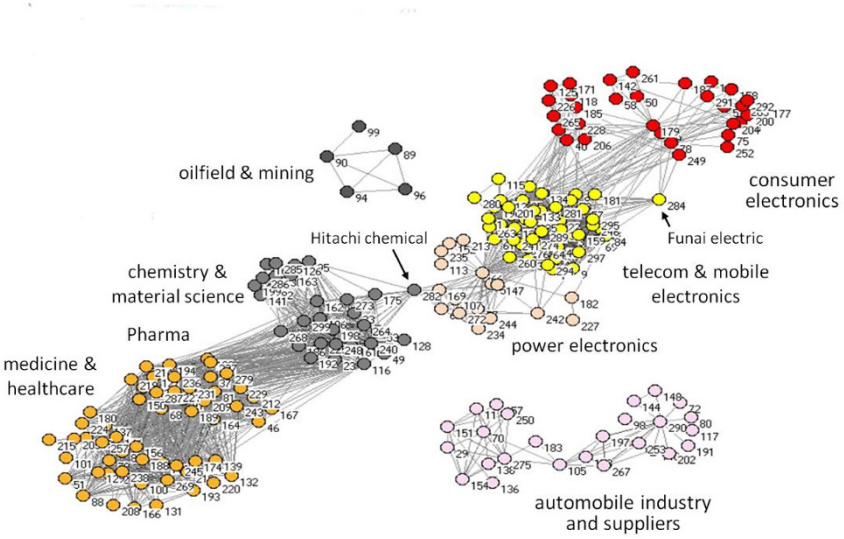

Fig. 4: Mapping patent portfolios of top 300 companies on technology categories: network with 5 % of strongest similarities to highlight technology categories; hierarchical level IPC4; 10 IPCs in each portfolio with the largest weight (colored figure online).

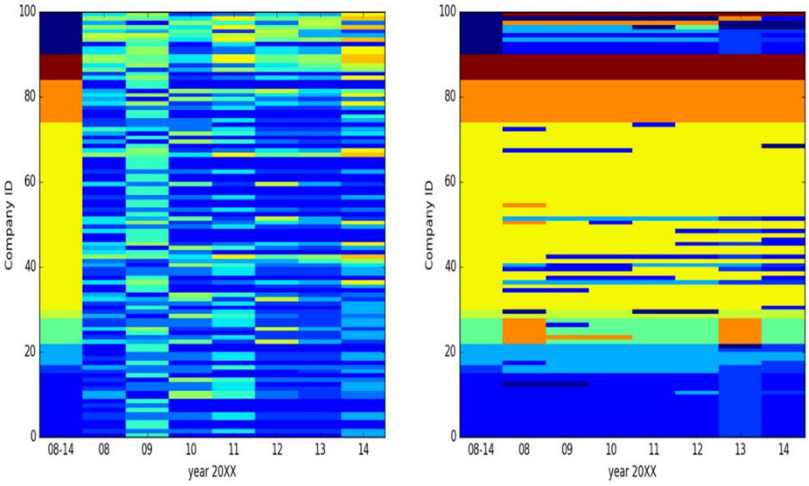

Fig. 5: Example of evolution of peer communities (shown by colors) in time before (left, *a*) and after (right, *b*) community labels matching for the top 100 companies with the highest patent portfolios volumes. The first column on the left on both figures is used as a reference for communities matching. This reference corresponds to communities detected in an aggregated network built over time period 2008-2014. All nodes (company IDs) in time snapshots are grouped according to the reference layer grouping.