



HAL
open science

A method to reduce false positives in a patent query

Johannes van Der Pol, Jean-Paul Rameshkoumar

► **To cite this version:**

Johannes van Der Pol, Jean-Paul Rameshkoumar. A method to reduce false positives in a patent query. 2021. hal-03287970

HAL Id: hal-03287970

<https://hal.science/hal-03287970>

Preprint submitted on 16 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A method to reduce false positives in a patent query

Johannes van der Pol^{a*}, Jean-Paul Rameshkoumar^a

^aVIA Inno, GREThA UMR-CNRS 5113, Université de Bordeaux

*Corresponding author: Johannes van der Pol; johannes.van-der-pol@u-bordeaux.fr

Abstract

The aim of this paper is to present a method that allows researchers and analysts to reduce the number of false positives in a patent query. Patents are not only used for prior art searches but increasingly for competitive analyses and the analysis of the evolution of technology. When these cases focus on specific technological domains, non-experts will aim to identify patents related to their focus-technology. In certain cases this can require complex queries to contain thousands of patents. It then becomes difficult to identify false positives. We present a method that allows researchers and analysts to refine their query on large datasets.

Keywords: Patent Query ; Patents ; Competitive Intelligence ; Technology Mapping ;
JEL: C80 ; O31 ; O34

1 Context and motivation

Patents are increasingly used as a data source for analyses that go beyond prior art searches. In Economics alone, patents have been used since the 1980's for the measuring of R&D output (Grabowski and Mueller (1972); Jaffe, Trajtenberg, and Henderson (1992)). Since then, patents have been used for a variety of uses related to the understanding of the evolution of technologies (Saint-Jean, Arfaoui, Brouillat, and Virapin (2020); van der Pol and Rameshkoumar (2018)), for competitive technological intelligence (Coates et al. (2001); Flamand (2016)), for measuring science-industry interactions (Han and Magee (2018); Tijssen, Yegros-Yegros, and Winnink (2016)) and more largely, for the analysis of technological innovation systems Frigant and Talbot (2005)). These analyses use patent data to identify citations between technological domains and firms, collaborations, emergence of technological concepts, inventor collaboration, patent transfers and so on (Ernst (2003); Trippe (2003)). The quality of the result of these analyses is highly dependant upon the quality of the patent dataset that is used. The analysis of a technological domain by non-experts implies the building of query to find relevant patents without specific knowledge of the technology. Any false positive can resulting in incoherent results in terms of citations, collaborations, and textual elements, which we want to avoid since they can lead to false interpretations end hence result in bad decisions. This makes the query a vital piece of the work even though there is no patent query that can ensure all relevant patents will be retrieved (Trajtenberg (1987)). For these reasons, and the ever increasing number of patents, it is important to be able to quickly and efficiently identify false positives in a patent query. This issue is different from the patent retrieval issue that has been largely documented (Khode and Jambhorkar (2017); Shalaby and Zadrozny (2019)). We are not concerned with the identification of patents close to a given patent, but rather with patents relevant to a technological domain.

In some case the construction of a query is a simple task, for instance if the domain one wishes to analyse is defined by one specific patent classification. Often however, this is not the case. 5G, structural composite materials for aeronautics, lithium ion batteries for cars, 3D skin printing and green tyres are some striking examples. The technologies do not have one specific classification and therefore require combinations of inclusions/exclusions of both different classifications and keywords. In such a case there is a high risk for false positives due to homonyms, acronyms, bilingual homophones, paraphrases and synonyms. It would be presumptuous for anyone to affirm that they know all about a technology and even experts are often surprised by the applications of a given technology. The difficult part of cleaning a patent query is to identify which patents are really out of scope and which are the beginning of a new application or trajectory. One could suggest to simply look at the classifications and exclude anything that does not make sense. However, excluding a classification can result in excluding relevant patents. Bamboo and tires do not appear to have much in common and yet bamboo fibres can be used in tires. Excluding bamboo might remove certain false positives but will also result in the removal of true positives.

Whether it is for the purpose of understanding the economics of innovation behind these technologies, the industrial dynamics of the strategic behaviour of a firm, analysts and researchers need to be able to identify relevant patents. Wether one has a technical understanding or not of the technology, queries can bring in false positives for a variety of reasons which we will discuss further in this paper.

We will show in this paper how we use classification networks to assess the coherence of a query, identify what is removed when we exclude a classification and how we identify quickly the classifications to verify. The aim of this paper is to provide analysts and researchers with a method to reduce false positives even without knowledge of the technology.

This paper is organised as follows, we will start by identifying how false positives emerge. Using this information we will explain how we use classification networks to identify problems in a query. We will show an example of the method before concluding the paper.

2 How false positives emerge

When building a query for complex technologies it is common to combine classifications with keywords. The use of keywords is common especially when classifications are too broad or inconsistent. There are no classifications for 5G technologies. To identify 5G patents, we would combine different keywords describing the underlying technologies with classifications on telecommunication. If we search for patents on silice reinforced rubber we would combine the classification for rubber with different keywords for silica. However powerful, keywords can bring in a lot of false positives. Suppose we would like to create a dataset containing all patents related to "carbon". We would search for the keyword "Carbon" in the text of the patent. This would bring in many relevant patents but consider the following patent (US20050150283A1) from which the description reads :

"FIG. 11 shows an embodiment in which lines 140 form a diamond-shaped network 141, which is connected at node 142 to transponder 136. [0085] Fiber-like lines 140 are advantageously made of steel cable, carbon, electrically conducting plastic, and other electrical conductors known from aeronautics, for example, and combined with other materials or fibers, e.g. carbon, aramide, steel cable plastic, electrically conducting plastic ceramic fiber, etc."

Terms such as "carbon", "aramide" can be keywords used for certain queries but it is clear from this text that the patent itself is not related directly to these technologies. This problem occurs often when building patent queries using keywords and is unavoidable. Other reasons for false positives/negatives are related to the terms themselves :

- Synonyms : A query should include all synonyms of a term. If not included some patents might be missed (false negatives/silence). *e.g* tire / tyre
- Homonyms : Results in capturing information that is irrelevant. *e.g* if aim to find patents related to trains (the transport vehicle) using the term "train" can capture patents containing variations of the verb "to train" resulting in false positives/noise.
- Bilingual homophones : searching in different languages is a problem and can bring in irrelevant patent documents (soy (the bean) and soy (verb "to be" in Spanish), tire (for a car) and tire ("to pull" in French)) resulting in false positives / noise.
- Paraphrase : This problem is especially present in patents. Patent authors will aim to be as vague as possible in their patents. The latter can result in authors not using certain terms (camera = a tool for taking pictures). This can result in missing relevant information (false negatives / silence).
- Acronyms : search engines do not always take caps into account therefore an acronym can be confused with a word. For instance, positron emission tomography (PET) will be confused with the word pet.

It is often complicated to find a way to remove the false positives without knowing if one does not remove any true positives (through the exclusion of a classification for instance). Even for an expert who knows the right keywords it might be complicated to know whether something that is being excluded could not be a true positive. In the method proposed here, we allow for false positives in the first stage. We then analyse the dataset to identify what can be excluded without risking removing true positives.

3 Classification Networks for query assessment

For the validation of the coherence of a patent query we use a classification network. False positives that come from keywords still contain classifications. By analysing how classifications are related it will be easier to assess when a patent is completely out of scope, closely related to the core of the query or valid for the query.

For instance if a patent contains a classification that is not otherwise connected to the core IPC classifications of the query we would consider it to be out of scope. If there are some connections, it might be a specific application or new trajectory, if the connection is strong we would consider the classification to be in the core of the technology. Let’s clarify these points.

3.1 Connecting classifications

A classification network is build from the classifications present on the patents of the dataset. Whenever two or more classifications are present on a patent, we connect these classifications. In figure 1 we show how the classifications (in this case IPC) are used to create a network. This process is repeated for all patents of the dataset as shown in figure 2 creating a large network. Whenever there is a classification in common between patents they will connect.

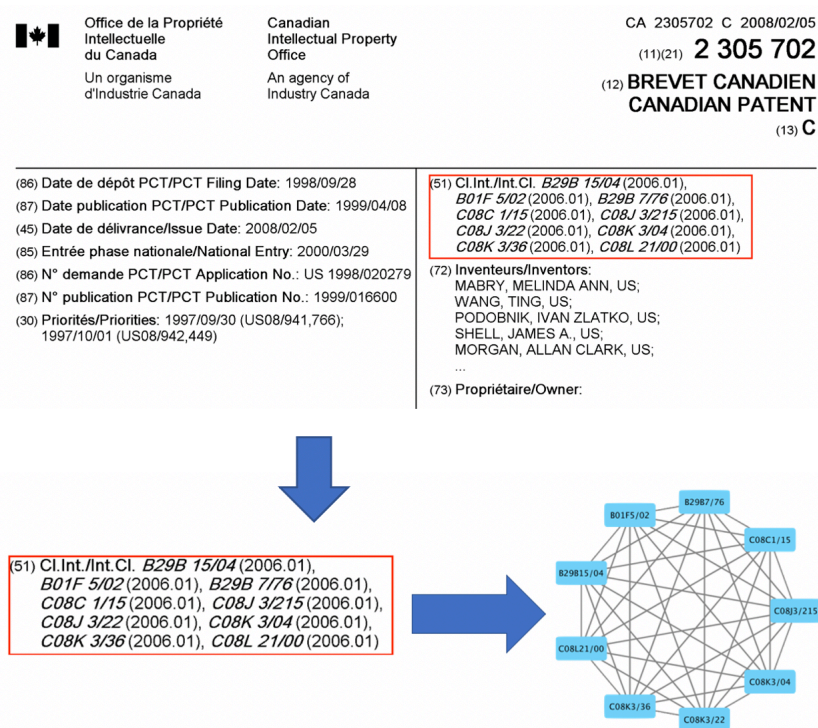


Figure 1: The IPC network is build by connecting IPC classes present on a same patent. This is done for all patents in a given portfolio.

Figure 3 shows a classification network for a technological domain. In this network of classifications we can see that there is a heterogeneity between the classifications in terms of the number of connections (as there should be). Some classification are more central to the network, some are more at the periphery. Others are not connected at all (components on the top right of the figure).

The structure of the networks provides us with insight on the technology we are analysing. The

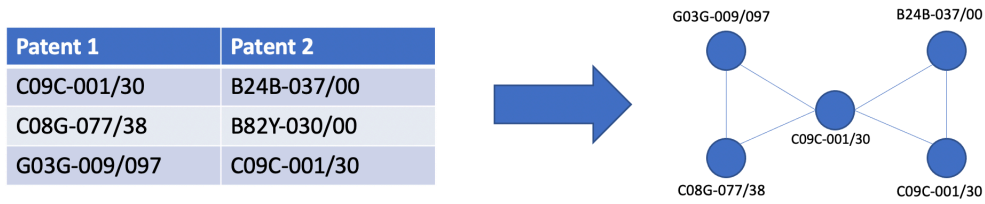


Figure 2: How the classifications of different patents are combined to create a network.

theory of innovation economics states that innovation is achieved by the combination of existing knowledge (Schumpeter (1942), Nelson and Winter (1982)). The means that patents related to a technological domain are somehow connected since they use similar underlying knowledge. We use classifications as proxies for these pieces of knowledge. A technological is build from inter-connected pieces of knowledge, and it stands to reason that all these pieces should be connected and form a whole. The network represents the interconnection of this knowledge. Knowledge that is not connected (the components on the top right) appear to have no relation and might be out of scope. There are classifications that make up the core (C08K and C08L for instance) and are clearly at the heart of the technology. Some classification are related to this core but they are further away. In terms of knowledge this would imply that they are either applications of the core technology or false positives that are related to the core but out of scope nonetheless.

van der Pol and Rameshkoumar (2018) have shown through a dynamic analysis how this type of network forms over time. They showed that an IPC network emerges with a technological core. Applications arrive later connecting to the core and changing its structure. In order to identify false positives from such a network we will use a typical three stage analysis of this network: analysis of the components, the communities and finally the nodes.

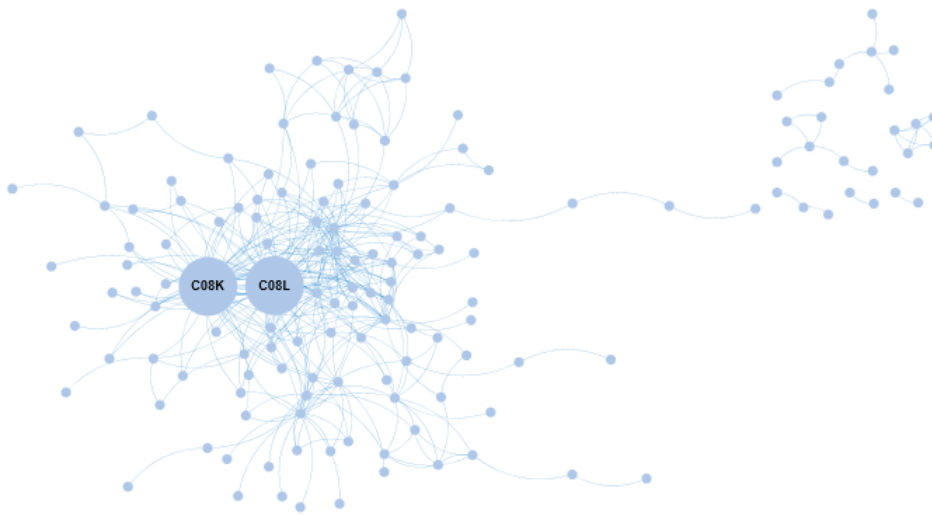


Figure 3: Example of network of IPC codes for a domain-level patent query. Source : Questel Orbit, network generated automatically with Intellixir.

4 Structural analysis for the identification of false positives

A normal network analysis works in three stages : an analysis of components, an analysis of clusters (or communities) and an analysis of the nodes ([van der Pol et al. \(2018\)](#)). The idea behind these steps is to explain how the network is build and identify the underlying rules at work. In our case we focus on a network that reflects innovation since it build by a combination of classifications. This means that we want to

4.1 Components

A component is a part of the network that is not connected to the rest of network ([Barabási \(2013\)](#)). In our case, figure 3 has eight components: the seven components on the top right and the large component next to it. These components contain knowledge that is not related to the core of the patent set. This does not automatically mean that the patents with these classifications should be removed. It is possible that these components contain knowledge that is relevant to the core but has not yet connected, but will in the future. The network allows us to quickly identify the classifications we need to check. Our experience shows that 95% of the time components reflect a mistake in the query.

If a mistake is identified one can simply remove the classifications from the query by excluding the classification. Since there is no link to other relevant classifications we now that exclusion will not result in the removal of true positives.

4.2 Communities

A second step in the analysis is the identification of communities inside the network. This step aims to segment the network into communities that represent a specific aspect of the technology (this could be a specific application of the technology, a subdomain, older technology etc.). We use this segmentation to identify groups of patents that combine different classification that are or are not in the core of the set we aim to build. Different techniques exist for network segmentation, *e.g.* K-means and modularity maximisation ([Blondel, Guillaume, Lambiotte, and Lefebvre \(2008\)](#)) are amongst the most popular ones.

In figure 4, nodes with the same colour are part of the same community¹. The results show nine different communities identified using modularity maximisation in Gephi ([Bastian, Heymann, and Jacomy \(2009\)](#)). The advantage of modularity maximisation is that the number of communities is defined by the algorithm. Modularity identifies nodes that are more densely connected between themselves than to the rest of the network. It maximises the number of links between nodes of the same community while minimising the number of links to the rest of the network. This means that the classifications in these communities are more densely connected to each-other than they are to the rest of the network. This implies that even though they are somehow related to the core of the technology but are combined with classifications that are a bit further from the core. An example is provided in figure 4 in which nine communities were identified, based on the structure these were already quite easily identifiable. By reading the classifications these communities contain we identified what these communities represent. This information is added next to the community. By reading these communities it becomes clear that these are applications of Silica but are not related to rubber. For instance the community on the upper left (urinals), is present because the patents contain the classification of silica, and a classification that related to non-organic compounds that was also present on some of the rubber patents. The network shows clearly that these patents are out of scope. However,

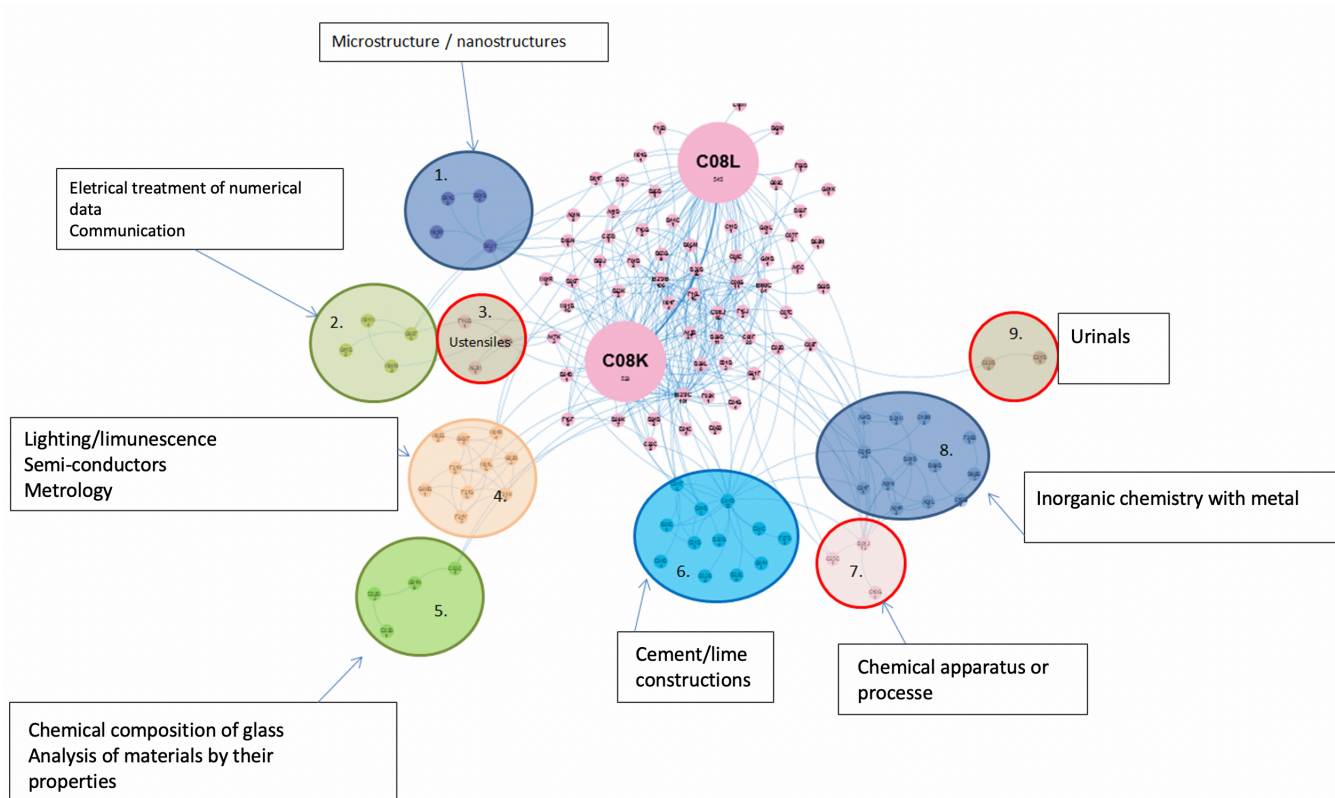


Figure 4: Example of an IPC network in which multiple communities were identified

contrary to a component, we cannot directly exclude these classifications from the query. For instance, if we would exclude all codes from community 1. we risk excluding a classification that would result in other relevant patents to be excluded as well. One therefore need to be careful with the exclusion method used. In the present case we would select the patents related to the core community (in the center) and keep only the patents related to the combination of those IPC codes. This means that we do not modify the query, this is done as modification directly to the dataset.

4.3 Nodes

The third and final step of the analysis focuses on the classifications themselves. We check here for two elements. 1. How the classifications of the query are positioned. 2. We check for gatekeepers.

The network informs us on how the nodes are connected. This means we know that if a certain classification is excluded from the query, which are classifications might be affected. This step also highlights the core of our query. If a query contains classifications (as is the case for the on in our example) one should expect these classifications to be central in the network. Some classifications can play a role of gatekeeper, in other words they connect different communities meaning that if the community is not a false positive, this code defines the application of the technology which is of interest since it allows for a segmentation of the query itself so one can perform an analysis on the application and the core of the technology separately.

5 Conclusion

In this paper we show how a classification network can help researchers and analysis with the validation process of their patent query. Even though we only provide one exemple in this paper, we have been using this method for the purpose of producing strategic analyses of players and technological fields for a multiple years now.

The method can be used on any classification as long as it is present on all patents of the set. Mostly we have used the IPC and CPC classifications, the choice between the two mainly depends on the technological domain. In certain cases CPC has a more precise classification system (for fuel cells for instance).

The method reaches its limit when classification networks are very dense which happens often in chemistry related fields. A solution we have found is to use network reduction techniques such as minimum spanning trees in order to make community identification easier. This method, however still need work. We have not addressed the issue of false negatives in this paper. As is, the method does not allow to do identify them. Further work is required to see if the identified communities and nodes can be used to expand an existing query.

References

- Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987), 20120375.
- Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An open source software for exploring and manipulating networks*. Retrieved from <http://www.aiai.org/ocs/index.php/ICWSM/09/paper/view/154>
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10), P10008.
- Coates, V., Farooque, M., Klavans, R., Lapid, K., Linstone, H. A., Pistorius, C., & Porter, A. L. (2001). On the future of technological forecasting. *Technological forecasting and social change*, 67(1), 1–17.
- Ernst, H. (2003). Patent information for strategic technology management. *World patent information*, 25(3), 233–242.
- Flamand, M. (2016). *Le déploiement de l'intelligence technologique dans le processus d'innovation des firmes: quels objectifs, enjeux et modalités pratiques?: Une application à l'industrie automobile* (Unpublished doctoral dissertation). Université de Bordeaux.
- Frigant, V., & Talbot, D. (2005). Technological determinism and modularity: lessons from a comparison between aircraft and auto industries in europe. *Industry and Innovation*, 12(3), 337–355.
- Grabowski, H. G., & Mueller, D. C. (1972). Managerial and stockholder welfare models of firm expenditures. *The Review of Economics and Statistics*, 9–24.
- Han, F., & Magee, C. L. (2018). Testing the science/technology relationship by analysis of patent citations of scientific papers after decomposition of both science and technology. *Scientometrics*, 116(2), 767–796.
- Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1992). *Geographic localization of knowledge spillovers as evidenced by patent citations* (Tech. Rep.). National Bureau of Economic Research.

- Khode, A., & Jambhorkar, S. (2017). A literature review on patent information retrieval techniques. *Indian Journal of Science and Technology*, 10(36), 1–13.
- Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change*. The Belknap press of Harvard university press.
- Saint-Jean, M., Arfaoui, N., Brouillat, E., & Virapin, D. (2020). Patterns of technology knowledge in the case of ocean energy technologies. *Journal of Innovation Economics Management*, 190–33.
- Schumpeter, J. A. (1942). *Capitalism, socialism and democracy*. Routledge.
- Shalaby, W., & Zadrozny, W. (2019). Patent retrieval: a literature review. *Knowledge and Information Systems*, 1–30.
- Tijssen, R. J., Yegros-Yegros, A., & Winnink, J. J. (2016). University–industry r&d linkage metrics: validity and applicability in world university rankings. *Scientometrics*, 109(2), 677–696.
- Trajtenberg, M. (1987). Patents, citations and innovations: tracing the links. *NBER Working Paper*(w2457).
- Trippe, A. J. (2003). Patinformatics: Tasks to tools. *World Patent Information*, 25(3), 211–221.
- van der Pol, J., et al. (2018). *Explaining the structure of collaboration networks: from firm-level strategies to global network structure* (Tech. Rep.). Groupe de Recherche en Economie Théorique et Appliquée (GREThA).
- van der Pol, J., & Rameshkoumar, J.-P. (2018). The co-evolution of knowledge and collaboration networks: the role of the technology life-cycle. *Scientometrics*, 114(1), 307–323.