Working Paper Series

Emanuela Sirtori [a, b]

[a] Maastricht Graduate School of Governance, University of Maastricht, Boschstraat 24, 6211 AX Maastricht, The Netherlands

[b] CSIL SCRL, Corso Monforte 15, 20122 Milan, Italy

Identification of multiple technology domains of LED through IPC community analysis

Working Paper N. 02/2023

Comments are welcome: sirtori@csilmilano.com

Identification of multiple technology domains of LED through IPC community analysis

Emanuela Sirtori [a, b]

[a] Maastricht Graduate School of Governance, University of Maastricht, Boschstraat 24, 6211 AX Maastricht, The Netherlands

[b] CSIL SCRL, Corso Monforte 15, 20122 Milan, Italy

Abstract

There is increasing demand for efficient methods to identify meaningful technological domains that can inform the study of technologies and their evolution. This paper shows how community detection analysis of a network of patent technology codes (IPC) can be used to classify a large and sparse patent database in a fully automated and unsupervised way. Light Emitting Diode (LED) has been selected as a test-bed of the methodology. As a multi-purpose technology, LED evolved for decades spanning several technology fields, before finding mass application in the general lighting industry. The analysis has been conducted over the largest database of patents related to LED ever used in the literature, covering over 400 thousand patent documents filed in 77 patent offices in the world between 1962 and 2018. VOS and Louvain community detection algorithms have been applied to find the technology domains around which the patent activity concentrated across the long and multi-directional historical evolution of LED. Results have been compared with other studies and approaches, in order to highlight the advantages of the proposed methodology. IPC-based community detection proves particularly useful to classify other technologies characterised by a meandering evolutionary process across several domains. It does not require particularly advanced data science skills and allows the flexibility of choosing the level of granularity in the classification by adjusting the resolution parameter.

Keywords: Patent network; Technology domains; Community detection; Multi-purpose technologies; LED

Contents

1 Introduction

There is rich literature that analyses technological evolution as a key driver of economic and societal change. Building on the initial contribution by Schumpeter (1934) on technological change and the impact of invention, innovation and diffusion processes on the economy, recent waves of empirical research have focused on the evolutionary dynamics of technological change (Nelson and Winter, 2004). These models have the concepts of technological paradigms and trajectories at the basis of their interpretation of the economic systems (Dosi, 1982). Technology develops through continuous incremental innovations, with new innovation building upon previous knowledge and extending existing ideas. Technology advance can follow different paths of engineering improvements and technical solutions, which can be partially autonomous one from the other (Dosi and Nelson, 2013; Verspagen, 2007).

The empirical research in this field has benefited from the valuable source of information provided by patent documents. The use of patents as an indicator of innovative activities is quite established in the literature (Yoon and Kim 2012; Grupp, 1990; Griliches, 1990). Information on the technological content can be extracted from patent applications and analysed to study the nature of technologies. The literature provides many examples of studies applying patent analysis to understand the development of particular technologies by cumulatively measuring the number of patents for specific technical fields (among which, Singh et al., 2021; Epicoco, 2013; Martinelli, 2012; Nomaler and Verspagen, 2019; Rizzi et al., 2014; Verspagen, 2007). Also, patent citations can be used as a proxy for knowledge flows and enable to map the diffusion of previous inventions over time, countries, and across technologies as well (Verspagen, 2007).

When technology develops along a meandering process and finds different applications in the course of its evolution, it is important to be able to identify and track the different technology domains spanned over time. A technological domain is defined as the set of artefacts (systems, processes, algorithms, devices) that achieve the same technological function using the same knowledge and scientific principles (Magee et al., 2016; Benson and Magee, 2015). The detection and analysis of these technological domains are of interest for many research areas: the historical analysis of the technology (e.g. Verspagen, 2007), the prediction of emerging technologies (Zhou et al., 2019; Érdi et al., 2013), the economics of innovation and industry dynamics (Marsili and Verspagen, 2002; Verspagen, 1991), market research and strategic management (Greve, 2000; King and Tucci, 2002; Narasimhan and Zhang, 2000; Ernst, 2003; Chang, 2012).

However, the pervasiveness of some multi-purpose or general-purpose (Bresnahan and Trajtenbergb, 1995; Lipsey et al., 2005) makes it more difficult to find meaningful and comprehensive classifications of technology domains to study (Jun et al., 2014; Bissmark and Wärnling, 2017). Light-Emitting Diode (LED) is an example of such technology. Its development builds on improvements in several technology domains, including semiconductors, chemical components, optical elements; also, since when it was discovered in the 1960s, the technology found applications in many fields, such as computer indicators, traffic lights, vehicles, printing devices, displays, lamps for general lighting (Sanderson and Simons, 2014). Identifying all the relevant technology domains associated with LED development is not trivial, not only because of ambiguous boundaries between technical fields, but also because of the large number of patents related to LED, which requires a great deal of time and manpower for data extraction and processing (Benson and Magee, 2015; Choi and Hwang, 2013). Different classifications have been produced by previous studies. All of them were based on patent analysis but adopted different classification methods. (Boyack et al., 2009; Chen et al., 2016; Choi and Hwang, 2014; Gridlogics Technologies Pvt Ltd, 2010; iRunway, 2014; Park and Jun, 2017; Simons and Sanderson, 2011).

This paper proposes an alternative classification of the technology domains related to LED, which differs from the existing literature by two main features: first, it is based on a very extensive database of over

400 thousand LED-related patent families, much larger than any previous study; second, it results from unsupervised community detection analysis applied to the network of technological codes (according to the International Patent Classification – IPC) included in the patent database. In other terms, the co-occurrence of IPC subclasses in the patent database has been analysed with community detection algorithms to find the homogeneous technology fields associated with LED. This approach is different from other grouping methods used in previous studies on LED and produces partially different results. We relied on the VOS and Louvain algorithms as two examples of community detection methods. We have chosen them not only because of their popularity and suitability to classify large networks, but also because they are embedded in existing programme packages, thus making them extremely easy-to-use also to use for researchers with no advanced data science skills.

This paper does not seek a better or more truthful classification of LED technology domains than those proposed by previous research. Instead, the objective is to show how community analysis of technology codes can achieve a meaningful and fine-grained classification of LED patents through a fully automated and data-driven classification approach. This methodology proves particularly useful to classify a very large dataset of patents referring to a heterogeneous technology, which developed through specific technology domains, whose number and definition are unknown a priori. As such, the findings of this study can be relevant when analysing other complex innovations crossing many application fields, such as software, biotechnology, telecommunication, electronic equipment or computer industries (Hall and Ziedonis, 2001; Ziedonis, 2004; Noel and Schankerman, 2013; Bessen and Hunt, 2007; Heller and Eisenberg, 1998; Martinelli, 2012; Fontana et al., 2009; Bresnahan and Greenstein, 2003). Moreover, the possibility to derive more or less detailed classifications by simply adjusting the resolution parameter of the partitioning algorithm is another key advantage offered by this methodology.

The paper is structured as follows. Section 2 reviews some previous studies of technological development and classification which rely on patent analysis, discussing the advantages and limitations of various classification methods. It also shows the classifications of LED technologies provided by previous studies, based on different methods of analysis. Section 3 outlines the overall approach and its application to LED technology, including the construction of the patent dataset and the IPC network, and the use of community detection algorithms. Section 4 provides some descriptive analysis of the results. A discussion on the robustness of results and a comparison with other classification methods are presented in Section 5. Finally, Section 6 concludes, by summarising the advantages of the proposed classification method.

2 Literature review

2.1 Identifying technology domains: why is it important?

Our research contributes to the recent streams of literature on technological change, centred around evolutionary approaches. Evolutionary economists consider technological change as a complex phenomenon, which proceeds through an evolutionary process along different trajectories (Dosi, 1982). The trajectory tends to be persistent and cumulative, since each new innovation builds upon previous knowledge and extends existing ideas. The main trajectory can also take different paths of engineering improvements and technical solutions which are partially autonomous one from the other (Nelson 1995; Verspagen, 2007). While technological trajectories correspond to incremental technological innovation, Dosi refers to the concept of technological paradigm shift to indicate a major breakthrough in knowledge development, both in the sense that it is a radical break with the past, and in terms of its reach, i.e., it affects a wide variety of research and industrial processes. The paradigm is set out by a small number of basic innovations, which eventually dominate the technological developments (or trajectory) for a long time, being constantly altered by incremental innovations (Dosi, 1982; Sahal, 1981; 1985; Dosi and Nelson, 2010).

Drawing on the concept of technological paradigm from the evolutionary literature, and from the idea of long waves and short waves to describe the frequency of major discoveries (Jovanovic and Rob, 1990), Bresnahan and Trajtenbergb (1995) introduced the notion of General-Purpose Technology (GPT). As formally defined by Lipsey et al. (2005), GPTs are characterised by some key features, the most important of which being their pervasiveness in the economy, due to the possible use in a vast number of products and applications, and their technological dynamism, reflected in several incremental improvements throughout their life time.

Examples of prevailing technological paradigms, or GPT, that are often mentioned in the literature, include electricity, semiconductors, steam power, microelectronics and information technologies (Dosi and Nelson, 2013). Even if some studies have challenged the possibility to trace clear-cut boundaries between radical and incremental innovations (Bekar et al., 2018; Moser and Nicholas, 2004; Korzinov and Savin, 2016), it is undisputed that not all technologies are alike. Some technologies show a higher degree of pervasiveness, a more complex evolution over time, and applications to multiple and diversified domains than others.

Light-emitting diode (LED) is an example of relatively more complex and pervasive technology. LED technology shows a long evolution through a continuous and incremental accumulation of innovation (Sanderson and Simons, 2014). The technology evolved thanks to significant improvements in the fields of chemistry, semiconductor materials, optical components, electronics, and simultaneously to its application in several domains (signalling, cameras, photo and printing, traffic lights, vehicles, horticulture, displays, horology, domestic and outdoor lighting, etc.). When thinking of the LED as a technology landscape populated by different innovations (Fleming and Sorenson, 2004; Kauffman et al., 2000), it is important to identify the multiple technology domains along which the LED evolved in order to investigate several aspects of its technological change. From a retrospective point of view, this is useful to analyse the breadth (different domains spanned) and depth (different innovations and novel recombinations in each domain) of technological improvements. From a forward-looking perspective, it can be used to map and examine the existing or emerging technological positions and guide the firms' decisions about where to position themselves in this landscape. In general, the distribution of innovations across the technology space is not uniform, but tends to agglomerate into clusters of adjacent technology positions (Aharonson and Schilling, 2016).

However, the question of how such domains can be identified is not trivial for technologies characterised by a long evolution across many different domains, including very specific niches. Differently from innovations deployed in a single narrow domain, LED builds on so many different technologies and its applications are so numerous that its landscape is more difficult to navigate. The methodology presented in this paper aims to achieve an accurate and meaningful mapping of the LED technology landscape so as to make its navigation easier.

2.2 Technology classification methods

Patent data provide a primary data source for scholars interested in studying the development of technological knowledge (e.g. Strumsky et al., 2012). Patents contain various types of content: the patent title, abstract, claims and description, the name of the investor, the assignee, citation information and others. A wealth of literature has taken advantage of the content of the patent applications to study the characteristics and evolution of technology over time. Two main types of information are generally used to analyse the patents: i) the description of the patent content as included in the patent's titles and abstract, or ii) the alphanumeric code assigned by the patent office examiner to classify the patent according to a specific classification system. Information from both sources can be used to classify patents.

The classification methods can be based on an expert review of patents or on automated data processing techniques. The choice of the classification approach cannot disregard the number of patents to classify and the type of technology under investigation. When dealing with small numbers of patents and with technologies with a narrow application domain, a manual classification can be feasible and produce accurate and relevant results, provided that the reviewer has sufficient expertise in the technology to be able to understand the patent documents and properly classify the technology into subgroups. Conversely, automated classification procedures are more suitable for very large databases and GPT / multi-purpose technologies. A classification algorithm, either based on clustering or community detection methods,[1] can be used to screen the content of a very large number of data and find an optimal way to partition them. In these cases, an expert can be asked to review only a sample of attributions and provide a general validation of the results, but not to classify the entire database.

In what follows, some of the main patent classification methods used in the extant literature are mentioned. They include either fully automated data analysis techniques or a mix of manual and automated procedures.

The studies that consider the patents' titles and abstracts to find agglomerations of similar technologies apply text mining and semantic analysis algorithms. A more detailed and technical review of these methods can be gathered from Tseng et al. (2007), Wang et al. (2018) and Hu et al. (2018). In general, scholars have used semantic analysis to analyse patent trends and forecast technological development in particular domains (Song et al., 2017; Park and Jun, 2017; Madani and Weber, 2016; Smith and Agrawal, 2015; Joung and Kim, 2017; Altuntas et al. 2015; Wu, 2016; Rizzi et al. 2014; Yoon and Park, 2004). The semantic analysis was also applied to help companies identify prior inventions and avoid patent infringements (Yoon and Park, 2004, 2014, Yoon and Kim, 2011) and to monitor technological development to identify novelty innovations (Bergeaud et al., 2017; Gerken and Moehrle, 2012). Arts et al. (2018) used a text-mining technique based on common keywords to develop a measure of technological similarity, thereby identifying classes of similar patents. Patent data were processed by concatenating the title and abstract and deriving a collection of unique keywords for each patent that represents its technical content. Bergeaud et al. (2017) developed a sophisticated and fully automated (unsupervised) method to classify patents according to their semantic content, using both individual keywords and multi-stems, and running multiple optimisation methods. Smith and Agrawal (2015) adopted a supervised approach through the use of textual mining and machine learning clustering techniques (k-means and k-medoids clustering) to discover meaningful associations throughout a corpus of patents and assess the accuracy of the USPTO technology classes. They found that there might be "hidden" clusters defined by textual clustering methods that offer better classification than the current USPTO system. Choi et al. (2022) used deep learning techniques for patent landscaping, by simultaneously using textual information from patent abstracts and citation-graph information to reduce the need for human resources and address the demand for automated patent classification. Zhou et al. (2019) applied a semi-automated topic clustering model to identify both old and newly-emerging technological topics and used sentence-level semantic analysis, rather than traditional keyword-based methods, to better differentiate topics in the same technological field that often contain similar vocabulary.

In general, the semantic analysis of patent data can preserve important technology content information (Tseng et al., 2007). However, patents with few keywords with little discriminating power, different spelling variants and synonym, and spelling errors increase the likelihood of false results and the need

[1] Clustering is a machine learning technique that groups data points into the same cluster based on their attributes. It can be applied to any type of database, not only on networks. Conversely, community detection is specifically tailored for network analysis and it allows discover communities inside them.

for validation by external experts. Furthermore, the semantic analysis is language-dependent: even if most of the patent documents today are available in English, some are in other languages, which could imply a significant pre-processing effort to translate all titles and abstracts into a common language. Some more sophisticated methods employing machine learning and deep learning algorithms are being developed and tested to improve the classification performance, but more work is still needed (Choi et al., 2022). Moreover, these methods require significant data science skills to be performed. Many researchers interested in technology analysis do not have these skills, which could prevent the wider adoption of these methods, at least for some time.

Other methods to identify and analyse the technology domains in a broader landscape rely on the existing national or international patent classification systems applied by the patent offices. Patent offices can employ different classification schemes, but the US (USPC) and the international (IPC) patent classification systems are the most widely used. The IPC system, established in 1972, is consistently adopted by more than a hundred countries. Its main advantages are its global and temporal coverage, as also patents earlier than the Seventies have been retrospectively classified with IPC class. Moreover, the classification is updated annually to reflect the emergence of new technology fields. This classification divides technology into eight broad sections, which are then further split according to a hierarchical structure currently made of 129 classes, 632 subclasses, 7,530 main groups, and approximately 64 thousand subgroups (Lupu et al. 2017).[2]

IPC codes (or similar standardised classification schemes) can be used to analyse technology positions on a technology landscape, measure the similarity or distance between technologies, and identify different technology domains along which technology evolves. IPC codes have been used, for instance, by Long and Ma (2015) to identify the core technologies in the metro infrastructure field, by analysing the co-occurrence of IPC codes and calculating the node importance in the overall network structure. Kim and Bae (2017) clustered the patent documents on the wellness care industry on the basis of their patent classification, examined the combination of patent codes of each formed cluster and forecasted the promising technologies on the basis of forward-citations. Ardito et al. (2018) categorised the Internet of Things (IoT) patent landscape into different subclasses based on an analysis of the IPC codes, then reviewed and validated by academic experts in the field. As one of the few examples of studies using community detection algorithms on patent data, Gao (2018) proposed a method to analyse patent classes of similar technologies as network communities: they applied the Lumped Markov Chain method and the Louvain method to extract communities from a citation network based on citations between subclasses of patent families citing each other. A review of other USPC or IPC-based measures that enable a fine-grained characterisation of a technology landscape can be found in (Aharonson and Schilling, 2016). Benson and Magee (2013, 2015) used a hybrid approach, by searching for keywords corresponding to technological domains of interest, and examining the overlap between the US patent classes and the IPC classes to identify complete and relevant list of patent classes associated to each domain.

The IPC system offers both patent examiners and other users the advantage of automated and standardised classification of all patents, made by experts with scientific or engineering background, and regularly refined to cover newly emerging technology fields. Relying on an already existing technology-based categorisation reduces the considerable amount of human effort needed to classify the applications (Bergeaud et al., 2017). Moreover, as it is terminology and language-independent, it overcomes the previously mentioned language-related challenges characterising the text-based and semantic analysis approaches. The classification can also be applied to old patent documents for which little or no searchable text is available, thus allowing a more precise and complete search. Because of these advantages, the IPC classification has been used in this study to identify the LED technology

[2] Figures are referred to the eight edition of the IPC classification.

domains (Section 3 and 4), but the pros and cons of using alternative classification methods are also discussed (Section 5).

2.3 Previous classifications of LED technology domains

Some previous studies have already provided classifications of LED into different technology domains. All these studies used patent data, but applied different grouping approaches, some of which relying on IPC codes, others considering the patents' titles and abstracts, and others using a mix of the two data sources. Moreover, some studies used manual aggregation of patents based on in-depth review of their content, while others adopted a more automated approach, based on unsupervised aggregation methods. More specifically, Gridlogcs (2010) extracted from the PatBase database the LED technology patents filed since 1927. After reducing the results to one member per family, the final dataset had 6,581 records, which were manually classified on the basis of their IPC classes and expert opinion. Park and Jun (2017) performed, instead, a mainly automated analysis. They used both patents' keywords, extracted through text-mining techniques, and IPC codes to estimate a Poisson count regression model for the analysis of smart LED systems and Bayesian networks to visualise networks of keywords and IPCs. Chen et al. (2016) adopted a different, but still automated approach. They selected the 20 companies that in Taiwan had the maximum number of LED patents until 2014 and used factor analysis to determine the key technologies that were covered, treating the different IPC codes (at eight digits) as variables. They found five main factors, corresponding to as many technological classes.

Among the studies that used semantic analysis methods, Choi and Hwang (2014) conducted a community network analysis of the keyword contained in a (small - less than 700) sample of patents of LED and wireless broadband fields between 2001 and 2011. They did so using the Label Propagation Algoritm developed by Raghavan et al. (2007). Boyack et al. (2009) worked on a larger dataset of 35,851 English-language articles and 12,420 U.S. patents published or issued during the years 1977-2004 on the domain of solid-state light and electroluminescent materials and phenomena. They computed bibliographic coupling metrics on backward and forward citations to find similar patents and publications and define technological sub-domains, through an average-link clustering algorithm that assigns each node to a cluster based on edges and distances between nodes.

Simons and Sanderson (2011) worked on an even larger dataset. They carried out a historical analysis of the emergence of the Solid-State Lighting industry (SSL), focusing on the multiple generations of technology and niche applications of LED. After extracting 185,852 patent applications on LED and SSL technology from 1937 to 2009, they classified them on the basis of both the IPC codes assigned to each application and the titles. They did so by developing and applying a taxonomy of IPC codes and keywords corresponding to each technology domain. They also used in-depth expert scrutiny of the content of a sample of patent applications to obtain final validation of the domains. The full IPC and keyword-base definition adopted to group patents is reported in their paper.

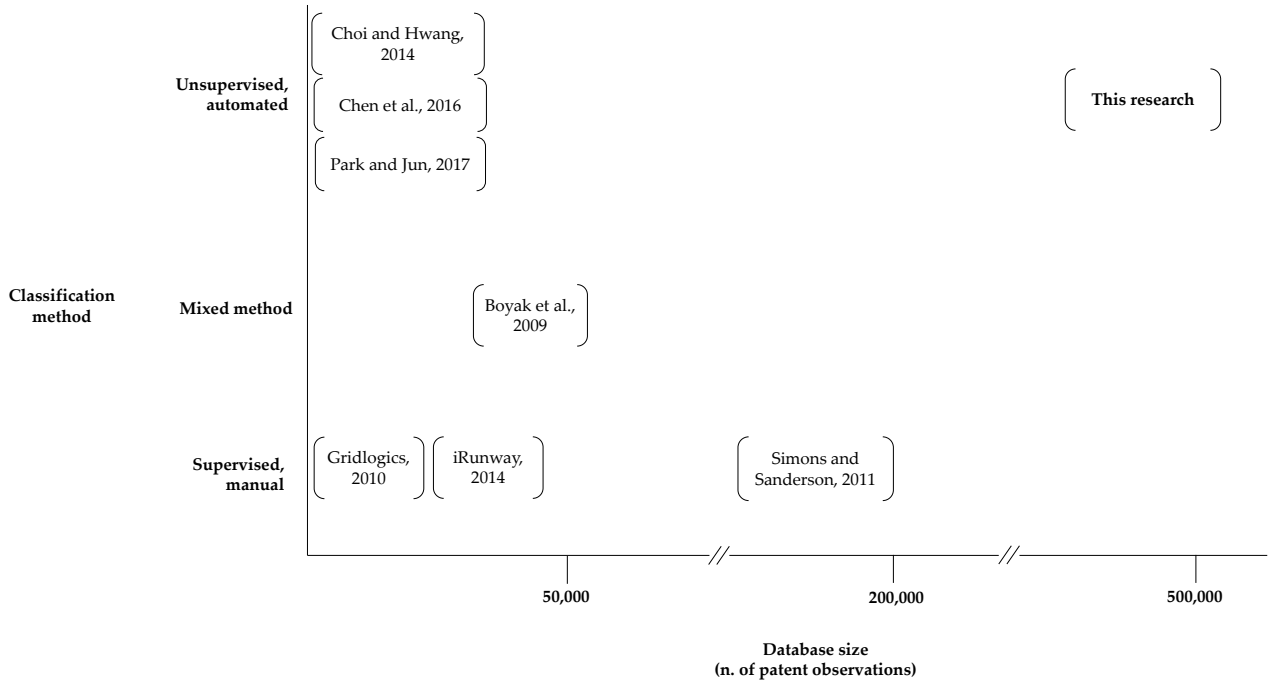
Table 1 presents the classifications proposed by previous studies to analyse the LED technology landscape and evolution. Some classifications allow the distinction between the fundamental technologies underpinning the LED (semiconductors, electronics, optical technologies, etc.) and application-related domains of LED (i.e. technologies for lighting, display, printers and scanners, vehicles, etc.). They all differ from each other due to the different methods applied, but also the different data considered, some of which limited to specific countries.

In general, previous studies show the feasibility of different methods for patent classification. The relatively small size of the patent databases handled by most of the previous literature made it possible to use either manual procedures or automated algorithms, as well as mix-methods for classification. When a larger database was analysed (Simons and Sanderson, 2011), the authors did not use any

statistical algorithm for automated data aggregation. Instead, they applied a supervised classification method, based on a definition of technology domains they had developed with the help of expert knowledge (Figure 1). The number of LED technology domains identified by these studies varied from four to 17, without any evident relationship between this number and the classification method adopted.

Against the classifications used in the literature, this paper shows the application of a different approach to classify a very large-size LED patent database, larger than those used in any previous study (nearly 500,000 patent families). The classification is conducted through a fully automated data analysis, based on community detection algorithms applied to the network of IPC codes. The methodology is presented and discussed in the next sections.

Figure 1: Mapping previous studies according to the number of patents classified and the classification method used



Source: Authors.

Table 1. LED technology domains identified in previous studies

	DATA SOURCE	SEARCH STRATEGY	COUNTRY AND TIME COVERAGE	DATABASE SIZE	METHOD OF CLUSTERING	TECHNOLOGY DOMAINS IDENTIFIED	
CHEN ET AL. (2016)	Taiwan patent search system (TWPAT)	Search based on the top 50 IPC classes related to LED technologies	Taiwan 1950-2014	4,511 patent applications (only from the 20 companies holding the largest number of patents)	Factor analysis on the top 50 IPC classes	1- Control components of light and magnetism 2- Control equipment of light source 3- Method of manufacturing component of semiconductor 4- Reflector and TV devices 5- Signalling system of visible signal	
GRIDLOGICS (2010)	PatBase	Search based on a mix of keywords and IPC codes	World 1950-2010	6,581 patents (one member per patent family)	Manual clustering, by reviewing 112 different IPC subclasses and their co-occurrence in each patent	**Application segments**: 1- Electric Light Sources 2- Semiconductor Technology 3- Audio Electronics 4- Optics 5- Medical Devices 6- Computer Peripheral Equipment 7- Traffic Control Systems 8- Scanning Equipment 9- Computer Data Transfer Devices 10- Heat Transfer & Control Systems 11- Kitchen Appliances 12- Dental Equipment 13- Automobile Lighting Systems	**Sub-technologies**: 1- Circuitry 2- Thermal control 3- Electroluminescent materials 4- Manufacturing and packaging
BOYACK ET AL. (2009)	Thomson Scientific's Science Citation Index; the US Patent and Trademark Office's database	Search based on keywords	World (but with over-representation of English-speaking countries) 1977-2004	35,851 articles 12,420 patent applications	Iterative process based on i) bibliographic coupling to compute cosine coefficients for each pair of records, and 2) clustering at increasingly higher level (from clusters of patents, to clusters of clusters) 3) final manual aggregation into "superclusters" by the Authors'	1- OLEDs 2- LEDs & Optics 3- LEDs & Heterostructures 4- Linear Arrays 5- Switches, Indicators 6- Indicators, Scanners	7- Sensors 8- Backlights 9- Panels, Phosphor 10- Portable lights Lamps, Controls
PARK AND JUN (2017)	Korea Intellectual Property Strategy Agency	Search based on a mix of keywords and IPC codes	World 1973-2015	4,226 patents applications	Combination of two methods: Method 1: Poisson hurdle regression model: keywords and IPC codes are used as explanatory variables of "Smart" and "LED" Method 2: visualisation based on Bayesian networks	**Smart LED technology:** 1- Electric lighting devices 2- Wireless control system 3- Layers and materials 4- Power signal	
IRUNWAY (2014)	USPTO	Not specified	World 1994-2012	22,662 patents applications	Property methodology, most likely based on a manual classification of IPC codes	**Technology domains:** 1- Light emissions: Materials, Front-end processing, Back-end processing	**Applications:** 1- Displays: Backlit, Active matrix

						2- Electronics: Power supply, Software 3- Light management: Directionality, Phosphors 4- Heat management: Air flow, Heat sinks	2- Lighting: Residential, Commercial/Industrial 3- Others: Indicators, Healthcare, Communication
SIMONS AND SANDERSON (2011)	USPTO	Search based on a mix of keywords and IPC codes	World 1935-2009	185,852 patent applications	Mainly manual clustering based on IPC codes and patents' titles	**Technology domains:** 1- FundSemic: Fundamental technologies for semiconductors 2- FundOptics: Fundamental technologies for optical elements 3- FundElectr: Fundamental technologies for electronic components mainly, but with some applications mixed in. 4- FundChem: Chemical fundamental technologies, including organic applications 5- FundComm: Fundamental technologies for communications, including telephonic and light beam 6- FundManuf: Manufacture of LEDs/OLEDs/PLEDs 7- FundSemiPl: Fundamental semiconductor technologies for devices with a plurality of components, most of which pertain to displays and arrays	**Applications:** 8- Lighting: Applications for primary lighting 9- Display: Applications for displays 10- PrintScan: Applications for printing and scanning use 11- Vehicle: Applications for vehicles 12- Projector: Applications for projectors 13- PhotoPrinter: Applications for taking or printing photos
CHOI AND HWANG (2014)	USPTO and World Intellectual Property Source (WIPS)	Search based on keywords	World 2000- 2011	31 LED patents and 346 wireless broadband patents	Community analysis of the network of keywords extracted from the patents' abstracts, using the Label Propagation Algorithm (Raghavan et al., 2007)	1- Inspection lamp 2- Concave mirror 3- Transparent reflective optic 4- Current regulator circuit 5- Flexible member 6- Power type 7- Power LED	8- Additional Optic 9- Package housing 10- Phosphor layer 11- Radiation pattern 12- Visible fluorescence 13- Injection moulding 14- Beam system

3 Methodology and data

3.1 Process for technology domains identification

The methodology proposed to identify the LED technology domains is based on three steps, visualised in Figure 2:. The first step relates to the construction of the patent database in the selected technology, in our case, LED. As explained earlier, we aim to compile a larger database than what was used in previous studies, in order to capture all the technological domains along which LED evolved over time. The patent search strategy and cleaning procedures are therefore designed to ensure high recall and precisions of results, as described in Section 3.2.

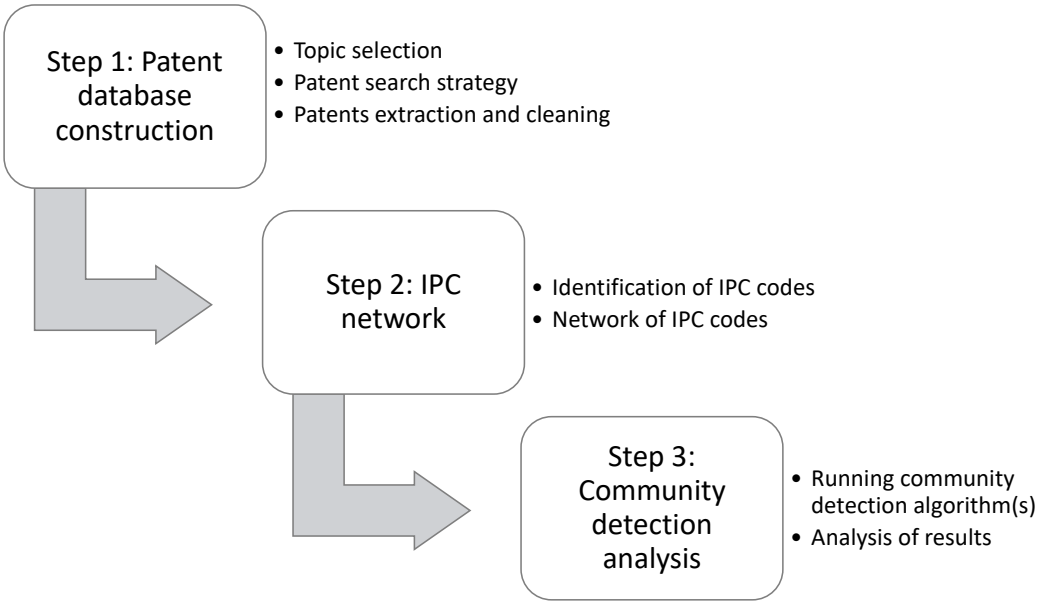
The second step is the extraction of IPC codes from each patent application previously retrieved (Section 3.3). As a unit of analysis, we have chosen the level of IPC subclasses (4 digits), which guarantees a sufficient fine-grained distinction between different technologies. Building an IPC-based network is easier than building a keyword-based network, as done for instance by Choi and Hwang (2014) and Choi et al. (2022). In the latter case, significant text mining, standardisation and cleaning work would be needed to set up the network. As discussed in Section 2.2, the IPC system offers instead the advantage of an already standardised and accepted classification of technologies. As an additional, but not strictly necessary way to increase the precision of the subsequent classification procedure, one can examine the frequency of IPC subclasses in the patent corpus and select only the most frequently used ones.

The third step consists in running the community detection algorithm using the constructed IPC network. As discussed in Sections 2.2 and 2.3, very few studies have used community detection analysis to identify technologies domains, and even less did so for the LED technology. Many algorithms have been developed and could be used to this end. They include the Edge betweenness (Girvan and Newman, 2002), Fastgreedy (Clauset et al., 2004), Walktrap (Pons and Latapy, 2005), Spinglass (Reichardt and Bornholdt, 2006), Infomap (Rosvall et al., 2007), Label propagation (Raghavan et al., 2007), VOS (van Eck and Waltman, 2007), Surprise (Aldecoa and Marín, 2013), Multilevel (Blondel et al., 2008), Louvain (Blondel et al., 2008) and the Leiden Community Detection methods (Traag, 2019).[3] In this paper, we focus on the Louvain and the VOS methods for three reasons. First, they are both popular and suitable for analysing very large networks. Second, both algorithms are already embedded in different programme packages for the analysis of network data, such as Pajek, the one used for the analysis in this paper. As such, they allow researchers to avoid computational complexity and concentrate on the interpretation of results. Third, the two methods share an interesting property: they both rely on an optimisation quality function that includes a resolution parameter, which can be arbitrarily adjusted to detect communities of different size. The largest the resolution limit, the larger the minimal size of a detected community (see Section 3.4 for more details).

We apply both algorithms on the same LED IPC-based network. The analysis of results can take different perspectives: it can compare the relation and degree of consistency and similarity of results obtained from different community detection algorithms; it can focus on the comparison between technological fields identified by one community detection method at different resolution levels; it can examine the technological fields identified by one selected algorithm at a specific resolution level, and use them to analyse the technological landscape and its evolution over time. Section 4 discusses all these types of findings.

[3] The python cdlib library contains many of these algorithms.

Figure 2:Analysis procedure for this research



Source: Authors.

3.2 Step 1: Patent database construction

A comprehensive database for the LED technology, covering relevant patents filed in any patent office of the world, since the early decades of 1900s until today, was developed for this research. LED patent applications were harvested from PATSTAT, the Worldwide Patent Statistical Database. As first step, all patent applications that respond to the following criteria were extracted: i) applications (appln_id) whose earliest filing date was from 1937 onwards[4]; ii) applications filed in any country / patent office jurisdiction; iii) all types of Intellectual Property Right types (including patents of invention and utility models); iv) both granted and non-granted applications. This search strategy led to the extraction of a total of 85,476,013 applications.

Then, a keyword search on the titles and abstracts was run to identify the LED-related patent applications. The search strategy followed Simons and Sanderson (2011). The relevant keywords included "light emitting diode(s)", "LED(s)", "OLED(s)", "semiconductor light emitting", "semiconductor lumin*", "solid state lighting", "solid-state lamp(s)", "luminescent diode(s)", and others (the full search strategy is presented in Appendix A.I). Patents referring to the semiconductor technology associated with light-emission were retained in the database, while semiconductors in general were deliberately not included. This criterion may have determined the omission of some relevant patents, but it would have more likely brought significant noise in the data.

Out of the 85.5 million patents applications extracted, 692,313 had at least one relevant keyword in either the title or the abstract, or in both. To avoid including patents where the keyword "LED" is simply used as a verb, we removed the patent applications where "LED" is used in combination with a number of prepositions/verbs, and no other relevant keywords appear in either the Title or Abstract. Over 100 thousand applications were removed through this automatic cleaning process (see Appendix A.I for more details).

[4] While the technology development of LED is generally indicated to start in 1962, some relevant research work already started in the previous years. Following Simons and Sanderson (2011), we extracted the patent applications related to LED technological development that were filed since 1937.

This strategy aimed at identifying LED-related patents with a high degree of precision. In fact, the large majority (over 90%) of patents collected include at least one relevant keyword in addition to "LED" (e.g. "light emitting diode(s)" or "solid state lighting"). The remaining patents could still include patents that are false positive, i.e. do not actually relate to the LED technology. This could happen, for instance, because the word "led" is used as a verb in combination with other prepositions that have not been taken into account in our previous cleaning step; or because the patents' titles and abstracts have been typed entirely in capital letters, which makes it difficult to distinguish with certainty whether the word "LED" was used as a verb rather than as an acronym. Therefore, the patents with the following characteristics were manually reviewed:

- 17,840 patent applications whose title or abstract contain the word "led" (in either capital or small letters), seemingly not used as a verb (i.e. not followed by any of the identified propositions or "to be"), and no other keywords;
- 2,315 applications whose title or abstract contain at least one LED-related keyword and also the word "led", but which seems to be used as a verb.

After this further cleaning step, 17,049 applications (i.e. 16,323 patent families) were dropped from the dataset. A total of 562,463 applications were retained for subsequent analysis, corresponding to 466,513 patent families (see Table 2).

While the criteria set to identify LED-related patents were sufficiently wide to return a larger number of relevant patents as compared to any other previous study (high recall), the combination of automated and manual cleaning process brought to the removal of around 20% of patent applications and families, thereby significantly increasing the precision of results.

Table 2. Results of the extraction process

		RESULTING PATENT APPLICATIONS (APPLN_ID)	RESULTING PATENT FAMILIES (DOCDB)
1)	Patent applications from 1937 to 2018	85,476,013	58,754093
2)	Patent applications where the Title contains at least one keyword	163,403	133,524
3)	Patent applications where the Abstract contains at least one keyword	674,192	581,220
4)	Patent applications where the Title or the Abstract contain at least one keyword	692,313	590,493
5)	Patent applications where the Title or the Abstract contain at least one keyword, after automatically cleaning the "LED" keywords	579,512	482,836
6)	Patent applications where the Title or the Abstract contain at least one keyword, after automatically and manually cleaning the "LED" keywords	562,463	466,513

The constructed dataset includes both granted and non-granted patents, patents of inventions and utility models. Patents span over a very long time horizon, including patents filed by virtually any countries of the world since 1962 (Figure 3: and Figure 4:). As such, to the best of our knowledge, this is the largest patent dataset that has been used to analyse LED technology.

The first patent in the database was filed in 1962 by the National Research Development Corporation (UK). It concerns a method to manufacture a light-emitting diode from a substrate of gallium arsenide or indium phosphide. It builds on earlier inventions by Nick Holonyak Jr (General Electric) and others, achieved in the context of experiments on electromagnetic radiation emitted by semiconductor devices. The dataset includes applications filed in 77 patent offices, more than two-third of which is from China, Japan and the USA (Figure 3: and Table 3). The data cover 343,993 granted patents (61.2%). 65% of all

applications refer to patents of inventions, while 35% to utility models. China has the largest number and share of utility models, being 64% of its overall patent application documents.

Figure 4: illustrates the number of patent applications by year. Having used the PATSTAT 2018 version, the extracted data are incomplete for 2017 and 2018, which explains the decreasing number of patents in those years. In 2016, there was a total of 55,234 different applications and 49,052 patent families.

Figure 3: Share of patent applications by patent office where the application was filed (1962-2018 period)

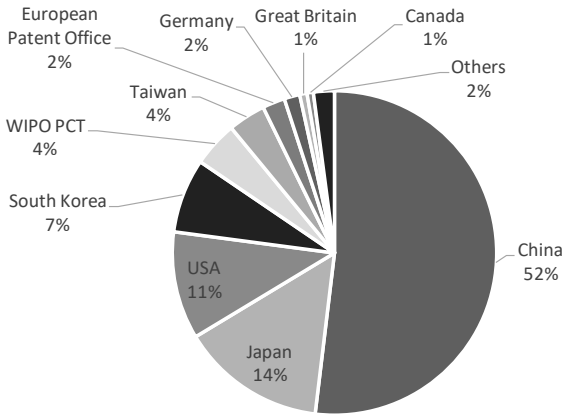
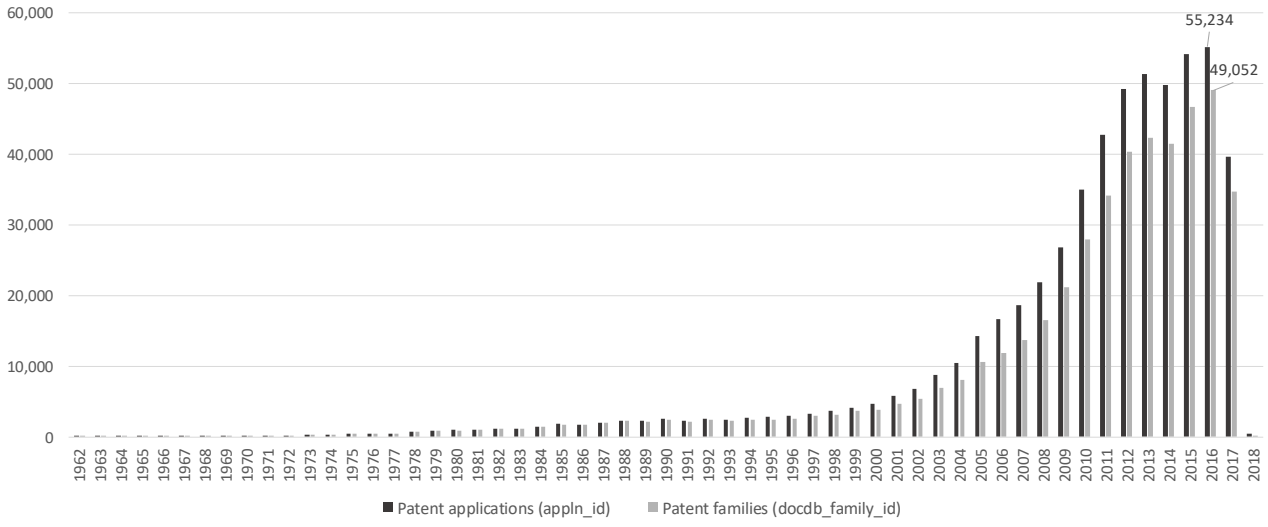


Table 3. Distribution of patent applications by patent office and type of application (patent of invention, or utility model)

	TOTAL N. OF PATENT APPLICATIONS	% OF PATENTS OF INVENTIONS	% OF UTILITY MODELS
CHINA	291,954	36%	64%
JAPAN	81,291	100%	0%
USA	60,378	100%	0%
SOUTH KOREA	41,675	97%	3%
WIPO PCT	25,410	100%	0%
TAIWAN	21,005	83%	17%
EPO	12,597	100%	0%
GERMANY	8,779	79%	21%
GREAT BRITAIN	4,104	100%	0%
CANADA	3,530	100%	0%
OTHERS	11,740	83%	17%

Figure 4: Number of patent applications (appln_id) and families (docdb_family_id) by year (1962-2018 period)



Source: Authors.

3.3 Step 2: IPC network

The dataset includes 606 different IPC subclasses (4 digits), spanning across all the eight IPC sections (A-H). If the subclasses are further broken down, we find 5,216 different IPC main groups and over 38 thousand different subgroups. Table 4 shows the twenty most frequently used IPC subclasses across the patents. They refer to different types of technologies, including those strictly related to LED applications (lighting devices, displays, photograph or projector systems, printing machines) and others related to the fundamental technologies underpinning the LED basic functioning and manufacturing, such as semiconductor devices, technologies related to electric heating, and optical elements. The most frequent IPC codes are indeed F21V "Functional features or details of lighting devices" and H01L "Semiconductor devices". These are mentioned in respectively 22% and 17% patent applications.

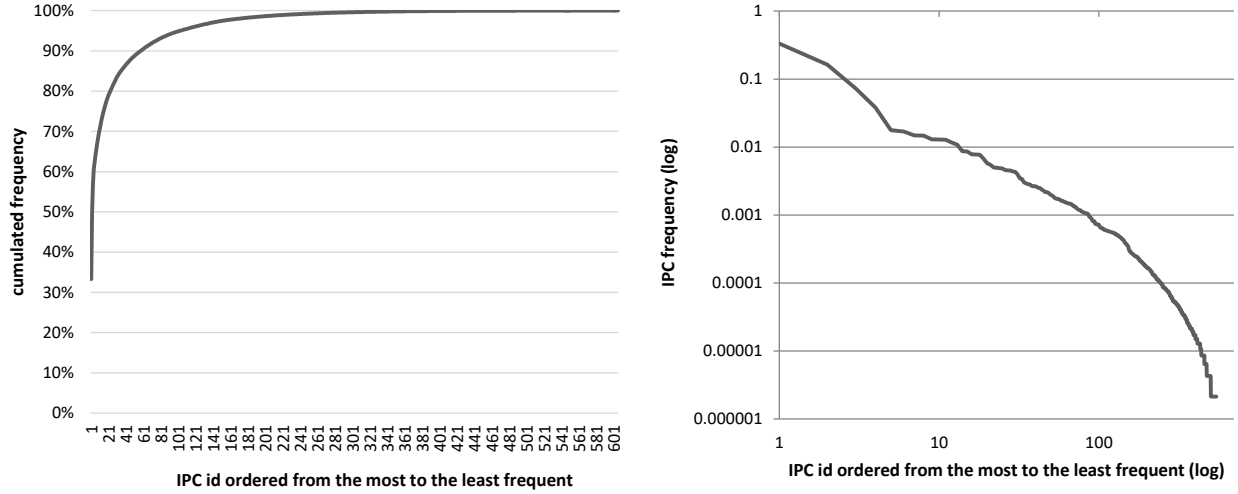
Since multiple IPC codes can occur in each patent application, and each patent family in the database includes between three and four individual applications on average, it results that each patent family has from 1 to 18 different IPC subclasses, with an average of 2. The most frequent IPC subclasses are mentioned by a large number of patent families. Specifically, 33% of patent families include at least one patent application that is assigned to the F21V subclass (Table 4).

Table 4. The twenty most frequent IPC subclasses in the database of patent applications and DOCDB families

IPC SUBCLASS CODE	IPC SUBCLASS TITLE	SHARE OF PATENT APPLICATIONS WHERE THE IPC SUBCLASS IS USED	SHARE OF PATENT FAMILIES WHERE THE IPC SUBCLASS IS USED
F21V	**MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING** - FUNCTIONAL FEATURES OR DETAILS OF LIGHTING DEVICES OR SYSTEMS THEREOF; STRUCTURAL COMBINATIONS OF LIGHTING DEVICES WITH OTHER ARTICLES, NOT OTHERWISE PROVIDED FOR	22%	33%
H01L	**ELECTRICITY** - SEMICONDUCTOR DEVICES; ELECTRIC SOLID STATE DEVICES NOT OTHERWISE PROVIDED FOR	17%	16%
F21Y	**MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING** - INDEXING SCHEME ASSOCIATED WITH SUBCLASSES F21L, F21S and F21V, RELATING TO THE FORM OF THE LIGHT SOURCES	8%	1%
F21S	**MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING** - NON-PORTABLE LIGHTING DEVICES OR SYSTEMS THEREOF	6%	1%
H05B	**ELECTRICITY** - ELECTRIC HEATING; ELECTRIC LIGHTING NOT OTHERWISE PROVIDED FOR	5%	7%
F21K	**MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING** - LIGHT SOURCES NOT OTHERWISE PROVIDED FOR	2%	0.1%
G09F	**PHYSICS** - DISPLAYING; ADVERTISING; SIGNS; LABELS OR NAME-PLATES; SEALS	2%	4%
G09G	**PHYSICS** - ARRANGEMENTS OR CIRCUITS FOR CONTROL OF INDICATING DEVICES USING STATIC MEANS TO PRESENT VARIABLE INFORMATION	2%	2%
F21W	**MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING** - INDEXING SCHEME ASSOCIATED WITH SUBCLASSES F21L, F21S and F21V, RELATING TO USES OR APPLICATIONS OF LIGHTING DEVICES OR SYSTEMS	2%	<0.1%
H04N	**ELECTRICITY** - PICTORIAL COMMUNICATION, e.g. TELEVISION	1%	2%
G02F	**PHYSICS** - DEVICES OR ARRANGEMENTS, THE OPTICAL OPERATION OF WHICH IS MODIFIED BY CHANGING THE OPTICAL PROPERTIES OF THE MEDIUM OF THE DEVICES OR ARRANGEMENTS FOR THE CONTROL OF THE INTENSITY, COLOUR, PHASE, POLARISATION OR DIRECTION OF LIGHT, e.g. SWITCHING, GATING	1%	1%
G02B	**PHYSICS** - OPTICAL ELEMENTS, SYSTEMS, OR APPARATUS	1%	1%
C09K	**CHEMISTRY; METALLURGY** - MATERIALS FOR APPLICATIONS NOT OTHERWISE PROVIDED FOR; APPLICATIONS OF MATERIALS NOT OTHERWISE PROVIDED FOR	1%	0.3%
B41J	**PERFORMING OPERATIONS; TRANSPORTING** - TYPEWRITERS; SELECTIVE PRINTING MECHANISMS, i.e. MECHANISMS PRINTING OTHERWISE THAN FROM A FORME; CORRECTION OF TYPOGRAPHICAL ERRORS	1%	0.4%
G06F	**PHYSICS** - ELECTRIC DIGITAL DATA PROCESSING	1%	1%
A61B	**HUMAN NECESSITIES** - DIAGNOSIS; SURGERY; IDENTIFICATION	1%	1%

H01S	**ELECTRICITY -** DEVICES USING STIMULATED EMISSION	1%	0.3%
G03B	**PHYSICS** - APPARATUS OR ARRANGEMENTS FOR TAKING PHOTOGRAPHS OR FOR PROJECTING OR VIEWING THEM; APPARATUS OR ARRANGEMENTS EMPLOYING ANALOGOUS TECHNIQUES USING WAVES OTHER THAN OPTICAL WAVES; ACCESSORIES THEREFOR	1%	0.5%
G01N	**PHYSICS** - INVESTIGATING OR ANALYSING MATERIALS BY DETERMINING THEIR CHEMICAL OR PHYSICAL PROPERTIES	1%	1%
H05K	**ELECTRICITY -** PRINTED CIRCUITS; CASINGS OR CONSTRUCTIONAL DETAILS OF ELECTRIC APPARATUS; MANUFACTURE OF ASSEMBLAGES OF ELECTRICAL COMPONENTS	1%	0.5%

Figure 5:Frequency of IPC subclass codes across the patent families



Source: Authors.

When the cumulated frequency of IPC subclasses in the database of patent families is computed, it is found that the two most frequent ones (F21V and H01L) are used by 50% of all patent families, the 50 most frequent IPC codes are used by 89% of patent families, and the top 100 IPC codes are used by 95% of all the patent families (see Figure 5). The remaining 506 IPC codes occasionally appear in a very small number of families. Given the high skewness of the distribution of IPC codes in the patent families, to increase precision in the classification, the next step of the analysis considers the top 100 most cited IPC codes. The Jaccard index between every pair of IPC codes was computed. The result is a 100x100 matrix of similarity coefficients, ranging between 0 and 1.[5]

3.4 Step 3: Community detection analysis

The different LED technological domains were identified by looking at the network of the 100 most frequent IPC codes at the subclass level and running community detection algorithms on them. Pajek (version 5.08) was used to analyse the Jaccard matrix. [6] The network of IPC codes has one giant component with 100 nodes.[7] The communities were searched by partitioning the net through two alternative routines: the Louvain algorithm (Blondel et al., 2008) and the VOS algorithm (van Eck and

[5] Being A and B two IPC codes, the Jaccard index is defined as the number of patent families that contain both A and B divided by the number of patent families that contain A and/or B. In notation, this is equivalent to: $J(A,B) = |A \cap B| / |A \cup B|$. The higher the index (closer to 1), the more similar the two IPC codes.

[6] The program, documentation and supporting material can be downloaded and used for free for non-commercial use from its web page: http://mrvar.fdv.uni-lj.si/pajek/.

[7] This giant component is defined in weak terms, i.e. every vertex can be reached following all edges regardless of their direction.

Waltman, 2007). With the Louvain method, communities are detected by optimising the modularity of the network, i.e. the metric of the density of links inside communities compared to links between communities (Eq. 1). The Louvain algorithm has the advantage of finding communities, even those weekly connected among each other, from very large networks in an efficient way (Traag et al., 2019). The VOS method also works by optimisation. It is generally used to find communities in patent and bibliometric networks but also to produce effective graphical representations of similarities between objects (van Eck and Waltman, 2010). Instead of optimising modularity, it optimises the VOS quality function, which is linked to the Jaccard coefficients of the network (Eq. 2).

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - r\frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad (1)$$

$$V = \frac{1}{2m}\sum_{i,j}[s_{ij} - r]\delta(c_i, c_j) \quad (2)$$

where

- Q indicates the modularity function
- V denotes the VOS quality function
- m is the total number of edges in the network
- r is the resolution parameter, whose default value if 1
- δ is a function that yields 1 if the vertices are in the same community and 0 otherwise;
- c_i, c_j are the respective communities to which i, j are assigned.
- A_{ij} represents the weight of the edge between i and j,
- s_{ij} is the association strength (i.e. the Jaccard coefficient) between vertices i, j
- $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i.

4 Results

4.1 Comparative analysis of results

Communities were searched with either the Louvain or the VOS algorithm at different resolution limits, from 0.50 to 3.50, with multi-level coarsening and multi-level refinement to obtain more stable results. The decompositions produced similar, although not identical results: three communities were identified at 0.50 resolution, and over 20 when the resolution parameter is higher than 3. The strength of the relationship between the two classifications, denoted by the Cramer's V measure, is overall good (above 0.75), but significantly high (0.93) at a resolution equal to 2. The Rajski's information indexes and the Adjusted Rand index also show that the similarity of results produced by the two methods is relatively higher at resolution around 2, and decreases as resolution increases (Table 5; see also the Appendices A.II and A.III).

Table 5. Results of the partitioning of the IPC network using the Louvain and VOS algorithms

METHOD	RESOLUTION	MODULARITY (Q) OR VOS QUALITY (V)	NUMBER OF COMMUNITIES	CHI-SQUARE	CRAMER'S V	RAJSKI (C1 <-> C2)	RAJSKI (C1 -> C2)	RAJSKI (C1 <- C2)	ADJUSTED RAND INDEX
LOUVAIN	0.5	Q = 0.630172	3	110.4624	0.7431771	0.37347009	0.51539677	0.57559328	0.6292899
VOS	0.5	V = 0.6561409612	3						
LOUVAIN	0.75	Q =0.536326	6	151.9853	0.8717378	0.42165482	0.82235387	0.46391084	0.6050236
VOS	0.75	V = 0.5465395278	4						
LOUVAIN	1	Q =0.478736	7	392.9317	0.8092504	0.49080882	0.69884102	0.62246634	0.3982234
VOS	1	V = 0.4740107147	7						
LOUVAIN	1.25	Q =0.43883	8	476.4485	0.8250095	0.56704913	0.74683862	0.70198183	0.5552085
VOS	1.25	V = 0.4361584247	8						
LOUVAIN	1.5	Q = 0.40021	9	580.3345	0.8517148	0.65315059	0.79020861	0.79016912	0.6376836

VOS	1.5	V = 0.4052666458	10						
LOUVAIN	1.75	Q = 0.367501	10	652.963	0.8517713	0.62274925	0.72937734	0.80988033	0.5118128
VOS	1.75	V = 0.3809649159	12						
LOUVAIN	2	Q = 0.340733	14	942.9544	0.9258679	0.76042865	0.89785169	0.83244672	0.7213277
VOS	2	V = 0.3629626144	12						
LOUVAIN	2.25	Q =0.316913	14	974.5993	0.8658474	0.74821453	0.87092871	0.84152738	0.6858062
VOS	2.25	V = 0.3458821204	14						
LOUVAIN	2.5	Q = 0.294606	15	1134.703	0.900279	0.79144179	0.89312429	0.87423911	0.6913192
VOS	2.5	V = 0.3295914061	15						
LOUVAIN	2.75	Q = 0.270942	17	1273.272	0.8920736	0.77723113	0.86892708	0.88045685	0.6730704
VOS	2.75	V = 0.3147972153	18						
LOUVAIN	3	Q = 0.249055	18	1272.726	0.8652527	0.73623198	0.82209004	0.87576734	0.6017696
VOS	3	V = 0.3015173093	22						
LOUVAIN	3.25	Q =0.226432	21	1484.281	0.8614758	0.75543859	0.84036124	0.88201319	0.5967126
VOS	3.25	V = 0.2915801813	25						
LOUVAIN	3.5	Q = 0.206092	23	1693.021	0.8772429	0.78577352	0.86678068	0.8937054	0.6421081
VOS	3.5	V = 0.2838107363	26						

Note: Clustering run in Pajek. For both the Louvain and VOS method, we used the standard clustering parameters, namely: Number of Restarts: 100; Maximum Number of Iterations in each Restart: 20; Maximum Number of Levels in each Iteration: 20, Maximum Number of Repetitions in each Level: 50.

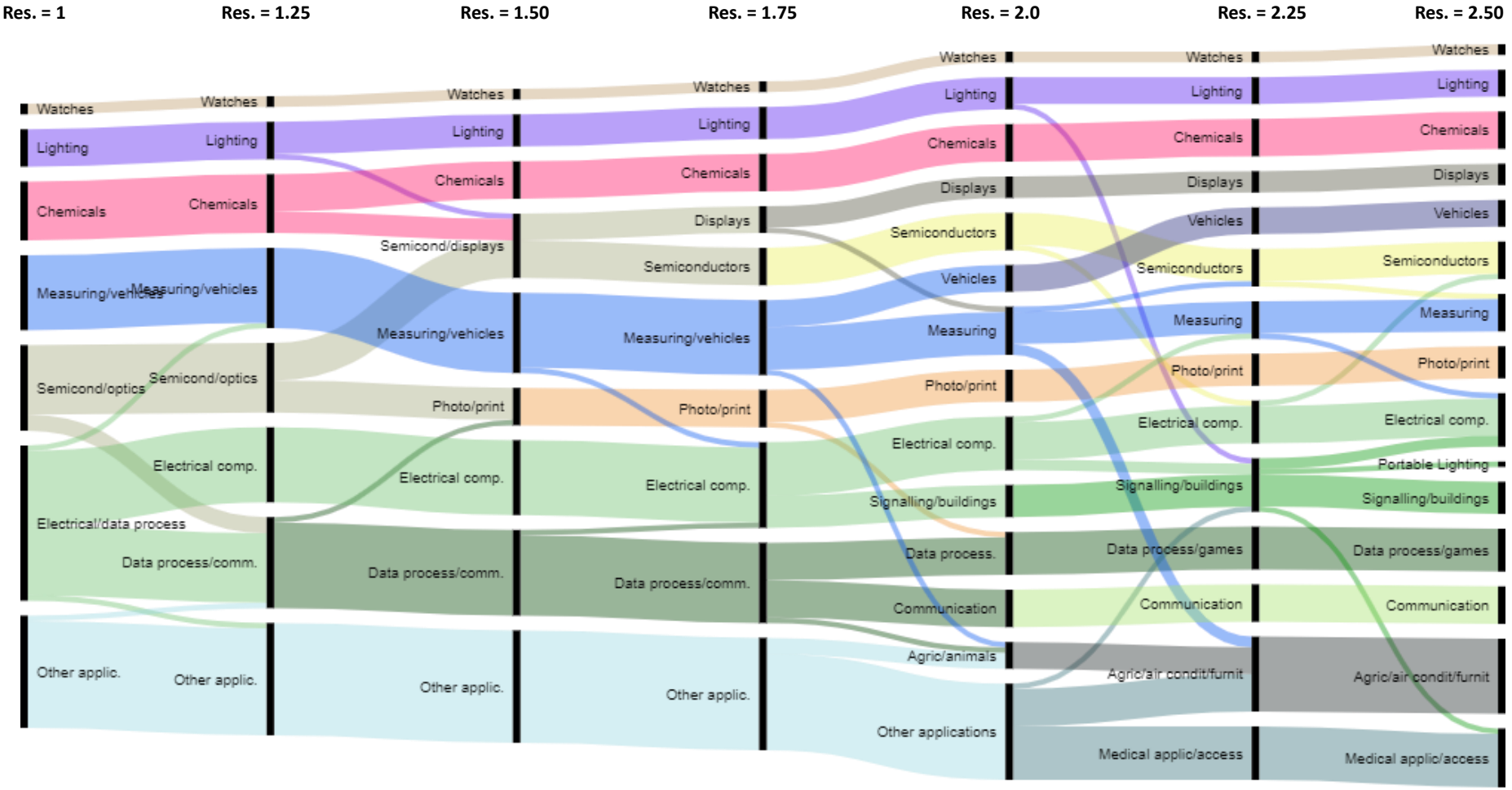
By reviewing the set of IPC codes included in each community, it is possible to understand the common technology characterising the domain. As resolution increases and a larger number of communities is formed, more specific technological domains appear from the decomposition and reassembling of other communities. The alluvial diagrams in Figure 6: and Figure 7: illustrate such a process, respectively for the Louvain and the VOS clustering, at resolutions going from 1 to 2.50. It can be observed that some specific technological domains, such as the one on watches technologies and lighting devices, appear already at a low-resolution level and remain nearly unchanged as resolution goes up. Other technological domains become evident only at a higher resolution. For instance, when the network of IPC codes is looked at low resolution, technologies for vehicles are merged with technologies related to measurement, and displays are not distinguishable from optics-related technologies.

Even when the most similar VOS and Louvain clustering results are considered, i.e. those at resolution around 2, some differences in the way how IPC codes are clustered by each algorithm can still be detected. For instance, with the Louvain method, a set of IPC codes related to chemical elements of LED (compounds, layers, coating composition, photomechanical processing and materials for semiconductor devices) stand out as a separate community. In contrast, with the VOS method, they are aggregated with semiconductor-related technologies. As another example, while the Louvain algorithm identifies a community of IPC codes related to data processing, the respective codes are split across different application communities by the VOS (namely, Communication, Photo & printing, Vehicles and Medical applications & games), as shown by Table 8.

Conversely, the VOS formula seems more suitable to detect specific application domains already at lower resolution limits than the Louvain method. When the same resolution limit is considered (2), the VOS formula separates the medical and games applications from other applications for agriculture, air conditioning and furniture. The Louvain procedure gathers, instead, all the respective IPC codes into one large community.

Tables 6 and 7 present the composition of the main technological communities identified with the Louvain and VOS algorithms respectively, at the same resolution level (2). Table 8 shows the intersection between the twelve communities resulting from the VOS method and the 14 communities found with the Louvain method at resolution equal to 2. The composition of each community at different resolution limits is presented in Appendix A.II.

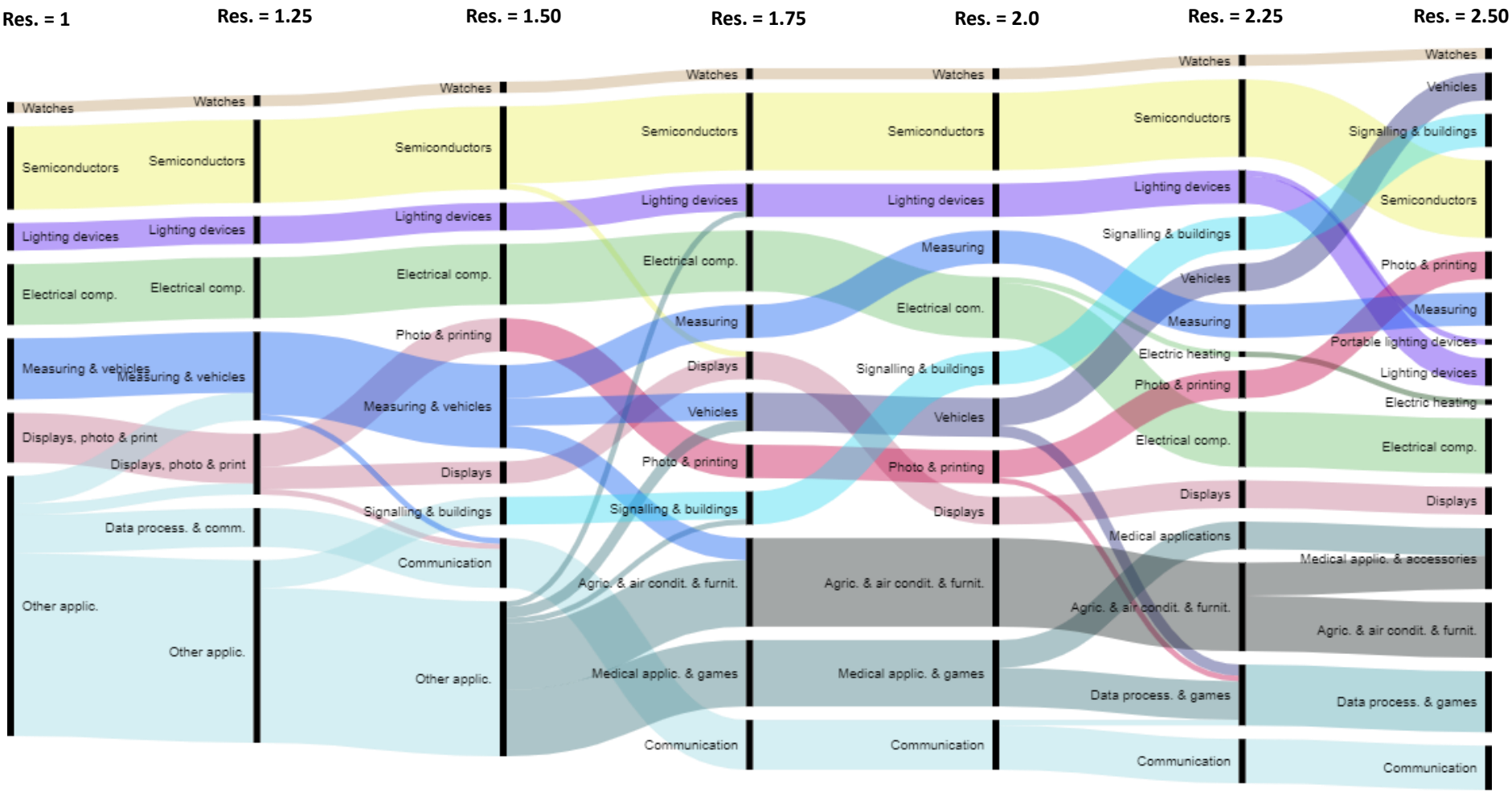
Figure 6:Louvain IPC communities at different resolutions: alluvial diagram



Note: The diagram was created in https://app.rawgraphs.io/

Source: Authors.

Figure 7:VOS IPC communities at different resolutions: alluvial diagram



Note: The diagram was created in https://app.rawgraphs.io/

Source: Authors.

Table 6. Main technological communities identified with the Louvain method, resolution = 2

LOUVAIN COMMUNITY	LOUVAIN COMMUNITY LABEL	COMMUNITY COMPOSITION: IPC CODES
1	Lighting devices	F21K F21L F21S F21V F21W F21Y
2	Semiconductors	C07D C09K C23C H01J H01L H01S H05B
3	Electrical components	G01R G05F H01H H01M H01R H02B H02H H02J H02M H03K
4	Displays	G02F G09F G09G H05K
5	Photo & printing	B41J G02B G03B G03G G06T H04N
6	Communication	G08C H04B H04L H04M H04Q H04R H04W
7	Vehicles	B60K B60Q B60R B62J G01D
8	Measuring	B23K G01B G01C G01F G01J G01K G01M G01N G01S
9	Medical applic. & games & air condit. & furniture	A41D A45B A45C A45D A47B A47F A47G A61B A61H A61L A61M A61N A63B A63H B65D F24F F25D G09B
10	Signalling & buildings	E01F E04F E04H G08B G08G H02S
11	Agric. & animals	A01G A01K A01M G05B G05D
12	Watches	G04B G04G
13	Chemicals	B29C B32B C08G C08K C08L C09D G03F
14	Data processing	A63F E05B G06F G06K G06Q G07C G07F G11B

Table 7. Main technological communities identified with the VOS method, resolution = 2

VOS COMMUNITY	VOS COMMUNITY LABEL	COMMUNITY COMPOSITION: IPC CODES
1	Lighting devices	F21K F21L F21S F21V F21W F21Y
2	Semiconductors	B23K B29C B32B C07D C08G C08K C08L C09D C09K C23C G03F H01J H01L H01S
3	Electrical components	G01R G05F H01H H01M H01R H02B H02H H02J H02M H03K H05B
4	Displays	G02B G02F G09F G09G H05K
5	Photo & printing	B41J G03B G03G G06K G06T H04N
6	Communication	G06F G08C G11B H04B H04L H04M H04Q H04R H04W
7	Vehicles	B60K B60Q B60R B62J E05B G01D G07C
8	Measuring	G01B G01C G01J G01M G01N G01S
9	Medical applic. & games	A41D A45B A61B A61H A61M A61N A63B A63F A63H G06Q G07F G09B
10	Signalling & buildings	E01F E04F E04H G08B G08G H02S
11	Agric. & air condit. & furniture	A01G A01K A01M A45C A45D A47B A47F A47G A61L B65D F24F F25D G01F G01K G05B G05D
12	Watches	G04B G04G

Table 8. Intersection between the VOS and Louvain methods (resolution = 2): number of IPC codes in each community

		VOS (res. = 2)												
		Agric. & air condit. & furniture	**Communication**	**Displays**	**Electrical components**	**Lighting devices**	**Measuring**	**Photo & printing**	**Semiconductors**	**Signalling & buildings**	**Vehicles**	**Watches**	**Medical applic. & games**	**Total n. of IPC codes**
LOUVAIN (res. = 2)	Agric. & animals	5												**5**
	Chemicals								7					**7**
	Communication		7											**7**
	Data processing		2					1			2		3	**8**
	Displays			4										**4**
	Electrical components				10									**10**
	Lighting devices					6								**6**
	Measuring	2					6		1					**9**
	Medical applic. & games & air condit. & furniture	9											9	**18**
	Photo & printing			1				5						**6**
	Semiconductors				1				6					**7**
	Signalling & buildings									6				**6**
	Vehicles										5			**5**
	Watches											2		**2**
	Total n. of IPC codes	**16**	**9**	**5**	**11**	**6**	**6**	**6**	**14**	**6**	**7**	**2**	**12**	**100**

4.2 Examining the LED technology domains

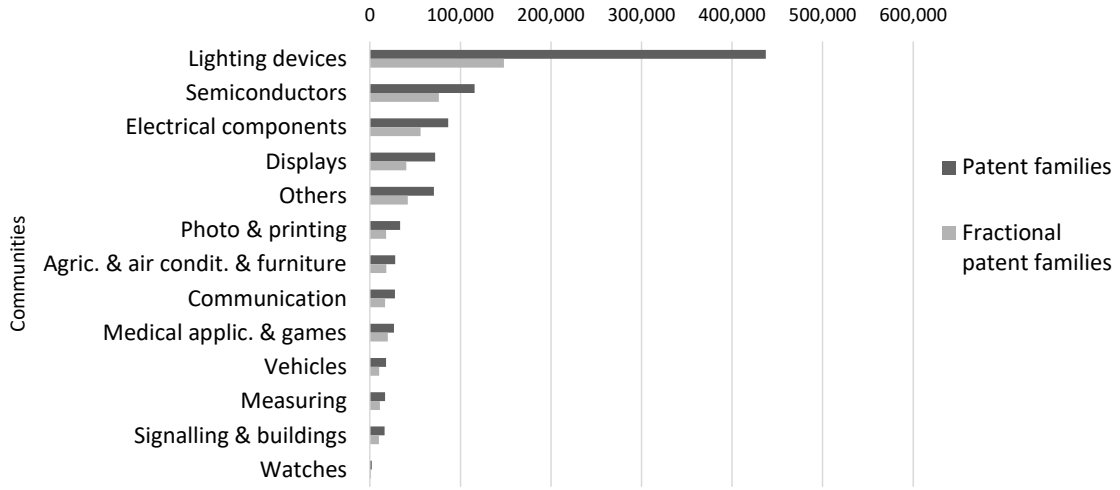
The different technological domains of LED identified by the community detection methods can be used to inform the analysis of the evolution of the technology. This section provides some insights that are gathered when examining the different technology domains of the LED patent landscape. Among the results obtained from either the Louvain and the VOS method and the different resolution levels, the focus is put on the communities resulting from the VOS clustering at a resolution equal to 2. The reasons for selecting these results can be derived from the previous discussion. First, the resolution limit equal to 2 is chosen because, at this level, the results from both the VOS and the Louvain methods are the most similar with each other, as denoted by the Cramer's V and the Adjusted Rand Index. Second, while having a smaller number of communities (12 with the VOS method vs 14 with the Louvain method, holding the resolution level fixed at 2), the VOS method enables to distinguish different applications of LED better. Table 9 presents a non-technical definition of the LED technology domains identified with the VOS community analysis.

Table 9. Definition of the LED technology domains

LED TECHNOLOGY DOMAIN	DESCRIPTION
SEMICONDUCTORS	Semiconductors light-emitting devices (diodes, chips, wafers, modules) based on several compounds. It includes methods to manufacture them.
ELECTRICAL COMPONENTS	Circuits, switching power supply, control systems applied to a LED device.
LIGHTING DEVICES	It includes two main types of devices: - LED light bulb (or lamp): Device that uses LEDs to produce lights and fits in standard screw-in connections. Instead of a filament, the LED bulb contains a number of LEDs and electronic components, hidden inside a shell that looks exactly like an incandescent bulb. - LED lighting fixture or luminaire: Lighting device around the light source. It consists of several components such as mounting, lamp holder, reflector, share or glass cover, and illuminant.
DISPLAYS	Panel displays that use an array of LEDs (backlighting) as pixels for a video display. They can be used indoor (televisions) as well as for outdoor applications (signs and billboards). They can also be used for smaller devises such as phone displays.
PHOTO & PRINTING	LEDs used in devices for taking photographs or projecting images, printing and scanning devices, typewriters. LED can be used for flash light and image processing (similar to laser light).
AGRICULTURE, AIR CONDITIONING, FURNITURE	LEDs used for: horticulture, LED devices incorporated into pieces of furniture, air conditioning and ventilation systems, refrigerators and other applications for "human necessity" (IPC Section A).
COMMUNICATION	LEDs in communication, transmitting and receiving devices. It includes LEDs to illuminate radio/telephone systems, liquid crystal display and keypad in mobile phones, and for wireless data transmission.
MEDICAL APPLICATIONS AND GAMES	LEDs used for medical treatments (e.g. phototerapy, light surgery) or for medical devices and instruments. This domains also includes LEDs for gaming products (video games) and toys. They share with medical application the use of LEDs mainly for status indicators on circuit boards or control panels.
VEHICLES	LEDs for front and rear lights, position detectors, vehicle headlight and safety light, electronic indicators in motor vehicles
MEASURING	Sensor, optical systems, photometers to measure the chemical or physical properties of LED and its parameters quantify and evaluate the parameters of LED (lighting output, luminous intensity, chromaticity, temperature, light and colour uniformity, contrast, …).
SIGNALLING AND BUILDING	LED lights used for signalling purposes, especially including traffic signal lights, landing zones, railway crossing, construction site signals, building decoration.
WATCHES	LEDs in clocks or watches, including smart watches

The analysis of patent diffusion in the network shows that 46% of LED patent families have IPC codes belonging to one community only; the remainder has a mix of IPC codes that fall under two or more different communities. The community of Lighting devices is the largest one in terms of patent families, with over 430 thousand families having at least one application assigned to the IPC code related to the lighting technology. Lighting technology is the most diffuse also when computing the fractional number of patent families, i.e. the patents divided by the number of communities to which their IPC codes belong, thus avoiding double counting (Figure 8:).

Figure 8:Number of patent families and fractional patent families by community (VOS, res.=2)

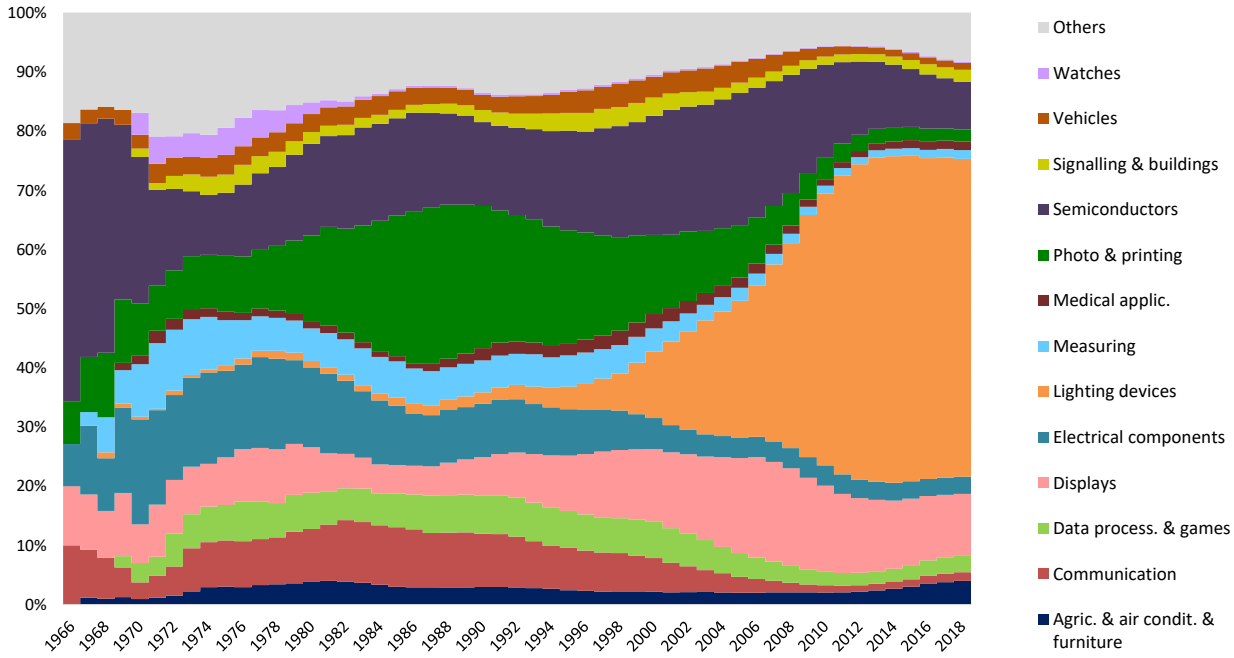


Note: The community labelled as "Others" refers to the residual IPC codes that have not been considered in the analysis, because they are not in the top 100 most frequently used IPC codes.

Source: Authors.

The number of patents by community and year can be looked at to have a preliminary understanding of how the technology has evolved over time (Figure 9:). Research on semiconductors started in the very earlier years of the LED technology evolution, but it continued over the following years and gained high relevance again in recent years, with the development of new organic semiconductor substrates for OLED display applications. Electrical/electronic technologies have been predominant until the early Nineties. Technologies for photo and printing devices were significant in the Eighties and early Nineties. Afterwards, the innovative effort focused on technologies for displays and lighting devices. These findings are coherent with other empirical studies on LED (Sanderson and Simons, 2014; Simons and Sanderson, 2011).

Figure 9:Share of patent families by community and year (VOS, res.=2)



Note: 5-years moving average from 1962 to 2018

Source: Authors.

When analysing different technological domains, an interesting question can arise, namely to what extent these domains are similar or different from each other. The concepts of technological similarity, proximity or relatedness are widely used in the literature to study knowledge recombination (Nakamura et al., 2015; Nelson and Winter, 2004) and the mechanisms for new knowledge generation (Joo and Kim, 2010). Different metrics and methods can be used to assess how close or distant two technological domains are from each other (Alstott et al., 2017; Aharonson and Schilling, 2016). Conducting a fully-fledged assessment of technological similarity, connectivity or relatedness is out of the scope of this paper. However, some insights can be derived through some descriptive analysis and simple statistics on the network of IPC codes and their diffusion in the patent database.

If the tree structure of the IPC nomenclature is exploited to analyse the combination of IPC codes across the 12 technological domains at section (1 IPC digit, from A to H) or class (3 digits) level, it can be observed how sparse each technological domain is across different types of technologies (Table 10). The domains of lighting devices, measurement technologies and watches are the most homogenous ones, as all the IPC subclasses codes composing these domains belong to one IPC section and class only. The lighting devices domain is composed of IPC codes related to the F section and F21 (Lighting) class. The measuring and watches domains are made of IPC codes falling into the G section (Physics) and, respectively, the G01 (Measuring; testing) and G04 (Horology) classes. All other domains are more heterogeneous in their composition, being characterised by IPC codes that belong to different IPC classes and sections. Conversely, some technological domains are more diffuse across different IPC codes. For instance, the community related to the design and manufacture of semiconductor devices results from the aggregation of IPC codes related to chemistry (section C), electricity (section H), and performing operations (section B). As another example, the electrical components community is mainly composed by IPC codes referring to electricity (section H), but also to power control and measuring properties (included under Section G).

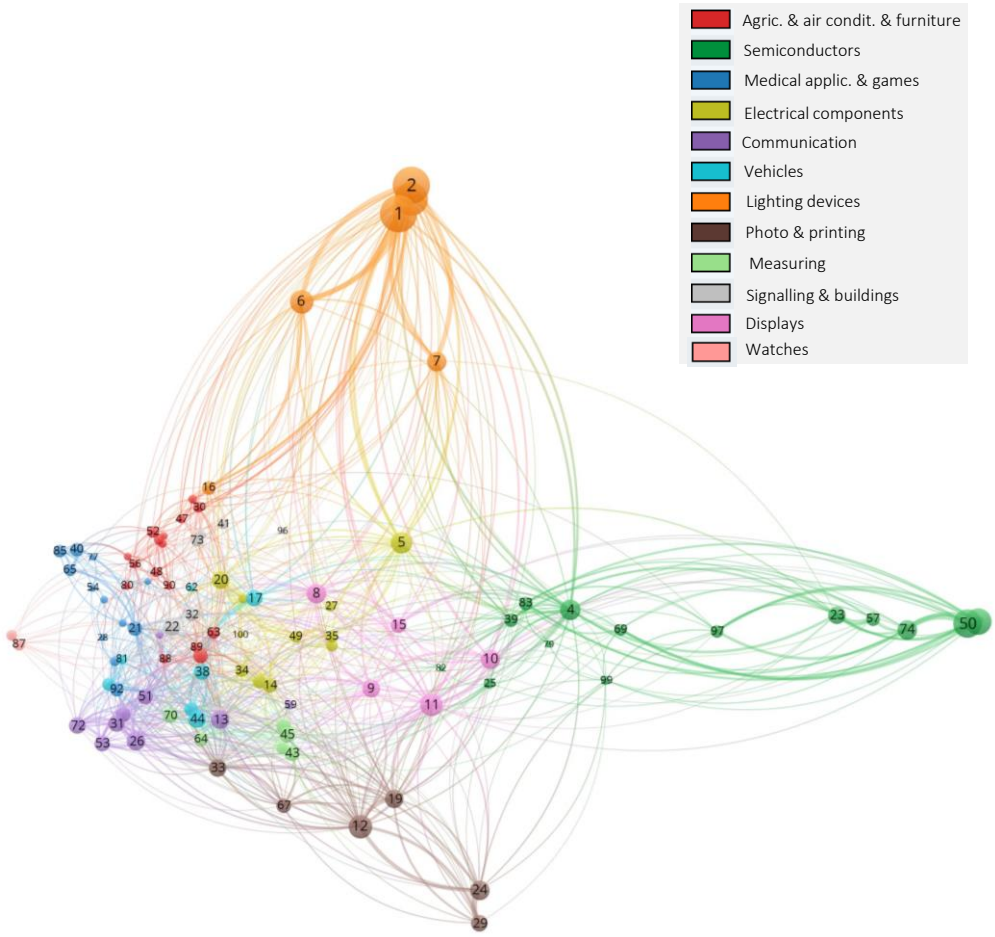
Table 10. Distribution of IPC classes of each technological domain across the IPC classes

		TECHNOLOGICAL DOMAINS (VOS COMMUNITIES, RES.=2)											
IPC CLASS	IPC LABEL	LIGHTING DEVICES	SEMICONDUCTORS	ELECTRICAL COMPONENTS	DISPLAYS	PHOTO & PRINTING	AGRIC. & AIR CONDIT. & FURNITURE	COMMUNICATION	MEDICAL APPLIC. & GAMES	VEHICLES	MEASURING	SIGNALLING & BUILDINGS	WATCHES
A01	AGRICULTURE; FORESTRY; ANIMAL HUSBANDRY …						19%						
A41	WEARING APPAREL								8%				
A45	HAND OR TRAVELLING ARTICLES						13%		8%				
A47	FURNITURE; DOMESTIC ARTICLES OR APPLIANCES …						19%						
A61	MEDICAL OR VETERINARY SCIENCE …						6%		33%				
A63	SPORTS; GAMES; AMUSEMENTS								25%				
B23	MACHINE TOOLS …		7%										
B29	WORKING OF PLASTICS …		7%										
B32	LAYERED PRODUCTS		7%										
B41	PRINTING; LINING MACHINES; TYPEWRITERS; STAMPS					17%							
B60	VEHICLES IN GENERAL									43%			
B62	LAND VEHICLES FOR TRAVELLING OTHERWISE THAN ON RAILS									14%			
B65	CONVEYING; PACKING; STORING, …						6%						
C07	ORGANIC CHEMISTRY		7%										
C08	ORGANIC MACROMOLECULAR COMPOUNDS …		21%										
C09	DYES; PAINTS; POLISHES; NATURAL RESINS …		14%										
C23	COATING METALLIC MATERIAL …		7%										
E01	CONSTRUCTION OF ROADS, RAILWAYS, OR BRIDGES											17%	
E04	BUILDING											33%	
E05	LOCKS; KEYS; WINDOW OR DOOR FITTINGS; SAFES									14%			
F21	LIGHTING	100%											
F24	HEATING; RANGES; VENTILATING						6%						
F25	REFRIGERATION OR COOLING …						6%						
G01	MEASURING; TESTING			9%			13%			14%	100%		
G02	OPTICS				40%								
G03	PHOTOGRAPHY …		7%			33%							
G04	HOROLOGY												100%
G05	CONTROLLING; REGULATING			9%			13%						
G06	COMPUTING; CALCULATING; COUNTING					33%		11%	8%				
G07	CHECKING-DEVICES								8%	14%			
G08	SIGNALLING							11%				33%	
G09	EDUCATING; CRYPTOGRAPHY; DISPLAY; ADVERTISING; SEALS				40%				8%				
G11	INFORMATION STORAGE							11%					
H01	BASIC ELECTRIC ELEMENTS		21%	27%									
H02	GENERATION, CONVERSION, OR DISTRIBUTION OF ELECTRIC POWER			36%								17%	
H03	BASIC ELECTRONIC CIRCUITRY			9%									

H04	**ELECTRIC COMMUNICATION TECHNIQUE**					17%		67%					
H05	**ELECTRIC TECHNIQUES NOT OTHERWISE PROVIDED FOR**			9%	20%								
	TOTAL	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**
	VARIANCE	**0**	**0.003**	**0.012**	**0.009**	**0.006**	**0.003**	**0.059**	**0.009**	**0.013**	**0**	**0.006**	**0**

Figure 10: illustrates the network of IPC codes, where the distance between the nodes reflects their similarity calculated based on the IPC co-occurrence matrix and the maximisation of the weighted sum of the squared Euclidean distances between all pairs of items (van Eck and Waltman, 2007). The network map shows that the IPC codes related to lighting devices, semiconductors, photo and printing, and watches are the most different from the other communities, being positioned at the margins of the map. These are also the communities that the community detection algorithms can identify already at low-resolution levels. At the centre of the graph, there is a core network of technologies that are relatively less different from each other and mainly relate to LED technologies for various applications (vehicles, communication, medical applications, etc.). It is interesting to notice that, based on the co-occurrence of IPC codes, the display-related technologies are positioned between the general semiconductor technologies, the photo and printing and other electrical components technologies. This location is coherent with the fact that the application of LED to displays historically originated from the evolution of semiconductor, electrical and image data processing technologies, the latter characterising the photo and printing devices as well (Sanderson and Simons, 2014).

Figure 10: Network visualisation of IPC communities (VOS, res.=2)



Note: created with VOS Viewer, 1.6.14 (27 Jan 2020).

Source: Authors.

5 Discussion on the robustness of results and comparison with other classification methods

The discussion in Section 4.1 shows that the results of the community analysis partially change depending on the algorithm used, both in terms of the number of communities identified and in terms of their composition. Moreover, different results are obtained when changing the resolution parameter of the algorithm. These issues are not considered limitations for this analysis. On the contrary, the possibility to get results at different levels of granularity depending on the chosen resolution, as well as to compare results from different algorithms, is regarded as an advantage of this method. By looking at the patent landscape from multiple perspectives and from different distances, the researcher can get a richer understanding of the technology domains and more flexibility in the description and configuration of the landscape, so as to better accommodate his/her specific research needs.

Still, some actions have been taken to check the stability of results deriving from each community detection algorithm. For both the Louvain and the VOS clustering, the standard parameters proposed by Pajek were applied, namely: number of restarts: 100; maximum number of iterations in each restart: 20; maximum number of levels in each iteration: 20; maximum number of repetitions in each level: 50. The communities resulting from running either one or the other community detection algorithm do not significantly change when changing any of these parameters. For instance, the same results are obtained when only 10 restarts are allowed.

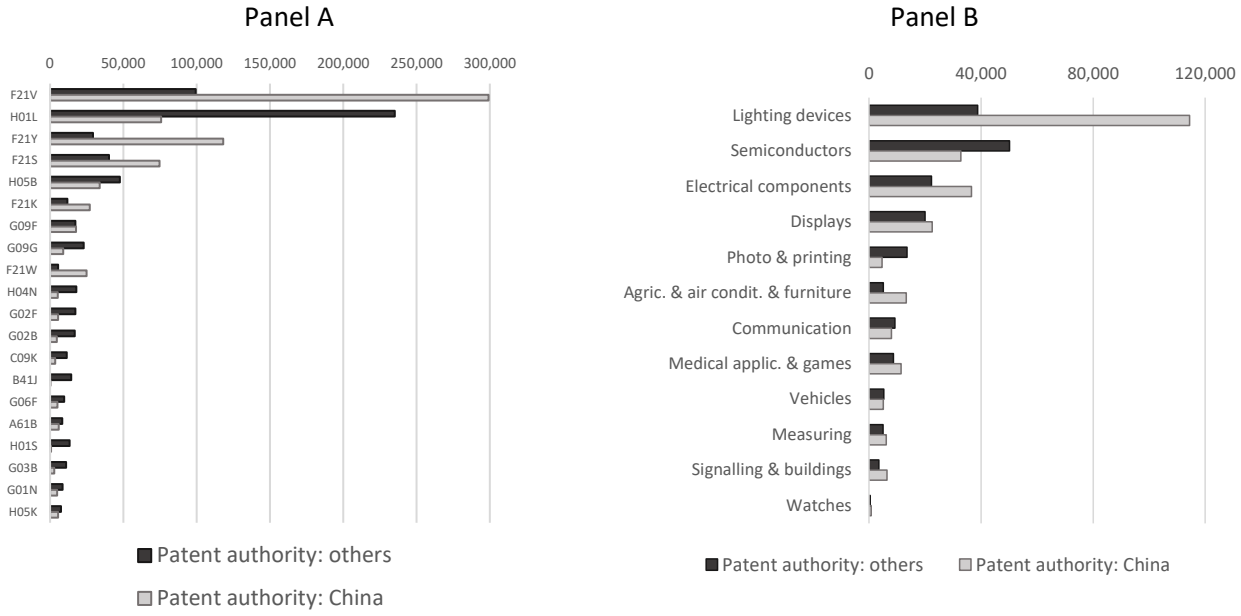
Considering the large number of Chinese patent applications and utility models present in the database, it is important to know if their inclusion or exclusion affects the results of the community analysis and, if so, to what extent. The Squared Chi test rejects the hypothesis that IPC codes are equally distributed in the total patent database and in the sample of Chinese patents, or of utility models. However, there is a positive probability that IPC subclasses assigned to a larger number of patent applications filed in any patent office worldwide are also the most frequent ones among Chinese applications; and that IPC subclasses to which larger number of patents of inventions are assigned are also the ones to which larger number of utility models are attributed.[8] Some notable divergences between the two distributions are worth mentioning (Figure 11: and Figure 12:): the share of Chinese applications in the subclasses F21V and F21Y are 22% and 10% higher than the share of applications filed in any other patent office, which denote a stronger specialisation of Chinese patent applicants in these technologies. On the contrary, the share of Chinese applications focused on technologies classified in the H01L code is 20% less than the share of applications in other countries. Similar divergences are found when considering the subsample of utility models and patent of inventions, also due to the fact that 64% of all Chinese applications are utility models: utility models are more frequent in the F21V subclass (+23%) and F21Y (+10%), while the share of patents of invention in the H01L subclass is 22% higher than the share of utility models in the same class. In all the other IPC subclasses, the discrepancy in the two distributions is lower than 2%.

Because of these differences in the distribution of IPC codes, the results of the community analysis partially change when excluding either Chinese applications, or utility models. More specifically, the VOS community detection method applied to the database of patents of inventions (thus excluding utility models) still finds 12 different communities at resolution equal to 2. The composition of these communities in terms of IPC codes largely overlaps with the results obtained on the full patent database:

[8] The correlation in the distribution of IPC codes between the number of Chinese applications and the number of applications at other patent authorities is 59%; the correlation in the distribution of IPC codes between the patents of inventions and the utility models is 61%. Both the correlation coefficients are statistically significant at 1% confidence level.

93% of IPC codes are clustered in the same communities as those found on the full dataset. When reducing the dataset to patents filed in non-Chinese patent offices, the VOS clustering finds 12 communities, which are similar by 86% in their composition to the results obtained from the full dataset. In this case and at this resolution, the algorithm cannot clearly distinguish among the LED technologies related to vehicles, measuring and testing, and other applications. It is sufficient to increase the resolution by 0.2 points to regroup the IPC codes into 14 communities and better distinguish the LED technologies for vehicles from other applications.

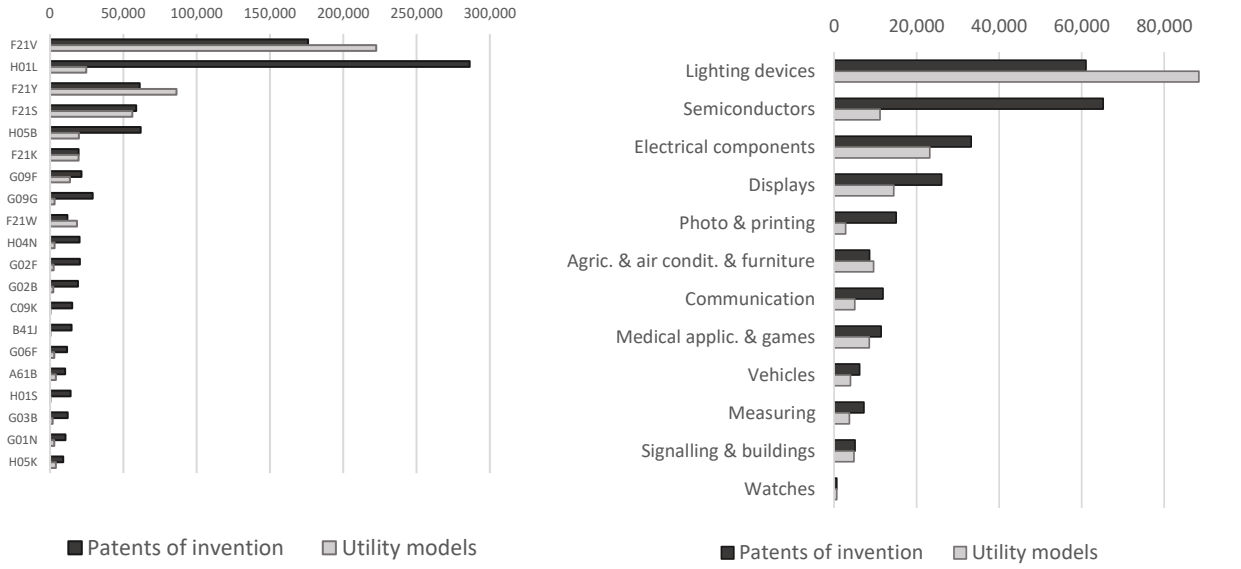
Figure 11: Diffusion of patent applications in the top 20 IPC subclass (panel A) and the VOS communities (patent B) by patent authority



Note: The fractional number of patent applications by patent authority and technology community has been calculated and displayed in panel B chart.

Source: Authors.

Figure 12: Diffusion of patent applications in the top 20 IPC subclass (panel A) and the VOS communities (patent B) by IPR type



Note: The fractional number of patent applications by patent authority and technology community has been calculated and displayed in panel B chart.

Source: Authors.

The classification produced by this study differs from other literature because of the different grouping method used, but also by the different search criteria considered to build the LED patent database (Section 3.2). This study considered a larger database both in terms of time and country coverage (any patent office worldwide). Table 11 compares the composition of technology domains resulting from our study (according to the VOS community detection method, with resolution equal to 2) and the technology classification developed by Simons and Sanderson (2011) through manual scrutiny and attribution of IPC codes. These authors developed their classification on the basis of the second largest database of LED patents ever used in previous literature, extracted with the same keyword search criteria, so comparing the results of the two studies is quite interesting.

In short, the following similarities and differences can be highlighted.

First, each study included in its classification a bunch of IPC codes that are not considered by the other study (see last row and last column of Table 11). For instance, Simons and Sanderson did not account for any IPC code of section A. These codes, however, are useful to identify the technology domains related to applications such as agriculture, air conditioning and furniture, and medical applications and games. Conversely, Simons and Sanderson used many codes of section B to detect patents related to the manufacture of semiconductors; these codes are not accounted for in this study because they are not among the 100 most frequently assigned to LED patents. Focusing on the most used and relevant IPC codes allowed this study to detect meaningful technology classes that have been overlooked in the previous study (another example is the watches domain), and at the same time to drop those codes which, albeit potentially relevant for the LED technology, are assigned to a much smaller fraction of patents.

Second, Simons and Sanderson's classification provides a more detailed characterisation of the fundamental technologies related to semiconductor devices, by distinguishing between fundamental semiconductor technology, plural semiconductors (e.g. arrays), manufacturing techniques of semiconductors, chemical components, organic LED. All these domains are grouped under the same category of Semiconductors in this study. As a matter of fact, the community detection algorithm finds this domain at low resolution level (VOS, res.=1) and its composition does not change as resolution increases up to 2.5 (see Figure 7:). While further domains appear from the decomposition of the semiconductors group at higher resolution levels (see Appendix A.II), the frequent co-occurrence of the underlying IPC codes suggests that this domain, albeit large in size, is in fact highly homogenous and therefore less prone to be split into smaller groups. A similar consideration applies to the photo & printing community, whose IPCs are distributed across three different classes by Simons and Sanderson, namely projector, printing and scanning, photo and printing.

Third, the IPC community detection method provides a more precise definition of the LED lighting technology, by rearranging the IPC codes that Simons and Sanderson used to define the LED lighting technology into more specific domains. Code G09F13 (illuminated signs), which they included in the definition of the Lighting domain, in fact is more often used in combination with other IPC codes related to the display domain, as reflected by the results of the community analysis. Likewise, it makes sense to separate the general lighting application of LED from LEDs used in traffic control systems (codes G98G), road and traffic signs and railway crossing (E01F/9 and E01F/13). The community detection analysis enables to make this distinction since a very low-resolution level.

Table 11. Comparison of the technology domains between Simons and Sanderson (2011) and this study (VOS method, res. = 2)

		TECHNOLOGY DOMAINS FROM THIS STUDY												OTHER IPC CODES CONSIDERED BY SIMONS AND SANDERSON BUT NOT IN THIS STUDY
		LIGHTING DEVICES	DISPLAYS	PHOTO & PRINTING	VEHICLES	SEMICONDUCTORS	ELECTRICAL COMPONENTS	SIGNALLING & BUILDINGS	MEASURING	AGRIC. & AIR CONDIT. & FURNITURE	COMMUNICATION	WATCHES	MEDICAL APPLIC. & GAMES	
TECH. DOMAINS FROM SIMONS AND SANDERSON	LIGHTING	F21K F21L F21S F21V F21W F21Y	G09F13				H05B31-43	E01F13 E01F9 G08G						F21H E01C17
	DISPLAY		G09G3/03-38 G09G5 G09F9	H04N3-17							G06F3/147 G06F3/033	G04G9		
	OPTICS		G02B G02F											
	PROJECTOR		G02B27/18	G03B21 H04N5/74 H04N9/31										
	PRINTING AND SCANNING			H04N1 B41J G03G15 G06K7 G06K15										Other subclasses of B41
	PHOTO AND PRINTING			G03B G03G		G03F								B27, other subclasses of B03
	VEHICLE				B60Q B60R B60K B62J									Other subclasses of B60 and B62. B61-B64
	FUNDAMENTALS OF SEMICONDUCTORS					G03F H01L21 H01L23 H01L25 H01L29 H01L31 H01L51 H01L33/00								G11C
	FUNDAMENTALS OF SEMICONDUCTORS WITH PLURALITY OF COMPONENTS					H01L27								
	MANUFACTURE OF LEDS					B23K B29C B32B				B65D				B0 (except B01D35-53 and B01J) B03 B04 B05 B07 B08 B21 B22 B24 B25 B26 B27 B28 B30 B42 B43 B66 B67; other subclasses of B23 B29 B32 B65
	CHEMICALS					C07D C08G C09D C09K C08K C08L C09K11 C23C								C01 C02 C03 C04 C09C C12 C22 C23 C25 C30; Other subclasses of C07 C08 C09 and C23
	ORGANIC					C07D C08G C08K C08L								
	COMMUNICATIONS										H04B10 H04M H04Q			
	ELECTRICAL TECHNOLOGY		H05K			H01J H01S	H01R H03K							
OTHER IPC CODES USED IN THIS STUDY BUT NOT BY SIMONS AND SANDERSON, BY TECHNOLOGY DOMAIN			Others groups in G09F and G09G	G06T, other groups in G06K and in H04N	E05B G01D G07C		Other groups in H05B; G01R G05F H01H H01M H02B H02H H02J H02M	Other groups in E01F; E04F E04H G08B G08G H02S	G01B G01C G01J G01M G01N G01S	A01G A01K A01M A45C A45D A47B A47F A47G A61L F24F F25D G01F G01K G05B G05D	Other groups in G06F and H04B, G08C G11B H04L H04R H04W	G04B, other groups in G04G	A41D A45B A61B A61H A61M A61N A63B A63F A63H G06Q G07F G09B	

Finally, it is worth mentioning that alternative approaches, besides the community detection method, were considered to identify the technological domains in the LED patent database. More specifically, the following methods were also tested in the framework of this research: i) text analysis of titles and abstracts of patent applications and clustering based on keywords instead of IPC codes; ii) clustering of patent applications (or families) on the basis of their IPC codes; iii) manual clustering of patents based on both IPC codes and keywords (following Simons and Sanderson, 2011). Presenting and discussing the results of all these other methods is out of the scope of this paper. However, some lessons can be drawn from those tests, which are useful to appreciate the differences and comparative advantages of the method chosen and presented in this paper (Table 12).

Table 12. Advantages and limitations of different methods applied to identify the LED technology community in the same patent database

METHOD	ADVANTAGES	LIMITATIONS
CLUSTERING OF PATENTS BASED ON IPC CODES AND KEYWORDS, ACCORDING TO A PRE-DEFINED CLASSIFICATION DERIVED FROM PREVIOUS STUDIES (MAINLY SIMONS AND SANDERSON, 2011)	• Top-down classification, which does not require automated data processing • Easy to apply	• The database contains some relevant IPC codes that are not classified by previous studies • Previous classifications may fail to identify relevant technology domains
TEXT-BASED CLUSTERING OF PATENTS	• Mix between automated analysis and manual review and agglomeration • Able to identify technology domains at very high-degree of granularity	• Applicable only for patents with titles and abstracts in English • High number of clusters to review in-depth to identify the underlying technology domain • Difficult interpretation of technology domains: the keywords clusters does not immediately point to a common underlying technology
CLUSTERING PATENTS BASED ON THEIR IPC CODES	• Automated analysis, enables to agnostically classify patents by looking at IPC subclasses as attributes • Easy interpretation of results based on the combination of IPC technology classes	• High number of clusters to review in-depth to identify the underlying technology domain • Highly data-intensive computing due to high data sparseness
COMMUNITY DETECTION OF IPC CODES (100 MOST FREQUENT ONES, OCCURRING IN 95% OF PATENT FAMILIES)	• Automated analysis, enables to agnostically classify patents by looking at IPC subclasses as attributes • Easy interpretation of results based on the combination of IPC technology classes • Focusing on the 100 most frequent IPC subclasses and clustering IPC codes, instead of patents, enable to reduce the data sparseness and make the clustering more manageable • Possibility to detect communities at different levels of granularity, by changing the resolution parameter	• One patent family may fall under different technology domains, depending on the co-occurrence of IPC subclasses in the patent applications

The text analysis of patent titles and abstracts was performed on 552,193 LED patent applications having a title and abstract in English (more than 98% of the initial extraction). Following Arts et al. (2018), Bergeaud et al. (2017) and Choi and Hwang (2014), over 6,461 1-gram stemmed keywords were extracted to construct the Term Frequency-Inverse Document Frequency matrix. The co-occurrence of keywords was analysed through the k-means clustering method (Sinanga and Yang, 2020), leading to the identification of 98 clusters. In each cluster, the top keywords were examined, a sample of abstracts looked in more detail and the distribution of IPC codes in the applications inspected in order to rearrange the 98 clusters in a smaller number of "super clusters" containing similar types of technologies.[9] This

[9] The Elbow method has always been used in combination with the k-means clustering in order to find the optimal number of clusters.

approach enabled us to find some relevant technology domains related to both fundamental technologies (semiconductors, electrical components, optical components, materials and chemical components) and application-related technologies (lamps, displays, advertising boards, camera, scanning, vehicles, air conditioning, computer indicators, phone, game machines, others). This method proved able to identify very specific technology domains largely similar to those resulting from clustering the IPC codes. However, this outcome depends primarily on the reaggregation of clusters that was performed manually. As a matter of fact, the keywords characterising each cluster were often not sufficient to enable the detection of the underpinning technology domain. Compared to that, the fully automated community detection from the network of IPC codes, that is discussed in this paper, produced much more self-explanatory results.

Another attempt consisted in clustering the 562,463 LED patent applications based on the diffusion of the IPC codes, either at subclass or group level. A varying number of clusters, from 10 to more than 80, was identified, depending on the clustering approach. Both the k-means and the hierarchical method (Ward, 1963; Kanungo et al., 2002; Cheung, 2003; Celebi et al., 2013; Sinanga and Yang, 2020) were tested. The analysis of IPC codes led to an easier interpretation of the results and attribution of each cluster to a technology domain. However, it required significant computation power because of the large size of the applications-IPC network. The high skewness of the frequency distribution of IPC codes (many IPC codes used in a small number of patents) produced a lot of noise in the data, making the database sparse and patents extremely difficult to cluster in a completely automated manner (Jun et al., 2014). Data sparsity is widely considered as a key cause for unsatisfactory classification accuracy (Bissmark and Wärnling, 2017). The problem arises when the matrix of data to classify (such as the IPC frequency matrix) has many missing values, which prevents from producing accurate predictions. Data sparseness had two implications: first, it determined the creation of a residual cluster of patent applications characterised by a combination of many different and not clearly interpretable keywords or IPC codes;[10] second, it prevents from properly checking the robustness of results, because of the high-computing power and time required to run more than one clustering iteration.[11] The top-down grouping of patents according to the IPC-based LED classification developed by Simons and Sanderson (2011) provides an easier way to split the database into technology domains. However, their classification considers only a subset of IPC codes, missing many relevant and highly diffused codes.

Against these alternative methods, the advantages of the community detection method presented in this paper are noticeable. The challenge of clustering a large and sparse database is addressed by grouping the most frequent IPC codes, instead of patent applications directly. Once the technology domains are identified, the distribution of IPC codes in the patent applications is analysed to attribute patents to one or more domains. The analysis leads to a manageable number of communities, easily interpretable, without any human action being necessary to adjust the clustering. Being a fully data-driven exercise, the classification is done in an agnostic way, on the basis of IPC codes co-occurrences only. The possibility to detect communities at different resolution levels is another advantage of this method over others. A possible drawback is that each patent may be assigned to more technology domains, depending on the multiple IPC codes assigned to each application. As explained in Section 4.2, around half of the patent families are classified under one technology domain only, the remainder falls into two or more domains. On the one hand, clustering IPC codes can cause some difficulties in the conceptualisation of domains

[10] This residual cluster is made of around 5% of patent applications, when the text-based clustering is performed. It has a varying size up to 20% approximately when the IPC clustering is applied, depending on the clustering procedure and parameters selected.

[11] In the database of 466,513 patent families, there are about 110 billion possible pairs of IPC subclasses. Running the clustering code on such network takes several days. The time increases as more iterations are run and different starting points are set to get more robust results.

when patent applications or families are the unit of analysis. On the other hand, having multiple technology domains assigned to patents is a property that can be exploited and used for ad hoc analysis, e.g. on the degree of relatedness of technological domains (Joo and Kim, 2010).

5.1 Conclusions

This paper contributes to the literature on technological change and evolutionary economics, by answering the question of how to identify the different technology domains of innovation, especially if characterised by multiple applications and a meandering evolutionary process. Community detection algorithms have been used for this aim. LED has been selected as an example of multi-purpose technology and used to test the feasibility of the community detection analysis. Indeed, while this methodology is consolidated in the literature and applied in several other research fields, its application for the analysis of LED and, in particular, for the identification of its technology domains, is new.

The co-occurrence of technological codes included in over 400 thousand LED patent families filed since 1962 across the world has been analysed. Depending on the partitioning method and the resolution parameters, the 100 most frequent IPC codes, assigned to 95% of all LED-related patents families, can be grouped into a number of communities. Each community of IPC codes can be interpreted as a technology domain. More detailed communities can be detected when the resolution parameter of the algorithm increases. The VOS method provides a more precise identification of specific technology domains of LED already at low-resolution limits, as compared to the Louvain algorithm. At resolution equal to 2, the Louvain and the VOS algorithms produce very similar results, both in terms of the number of communities detected and their size and composition. Twelve different communities are identified at this resolution limit with the VOS method. The largest ones are related to the application of LED to lighting devices, displays and printing, and the general semiconductor technologies and other electrical/electronic components. Other specific technology domains refer to the use of LED for watches and computer indicators, which are among the first applications of LED, but also traffic lights, vehicles, medical applications and others.

Both the visualisation of the network of IPC communities and the analysis of the distribution of IPC codes across the patent families and over time provide an indication of how the technology domains have emerged, which is coherent with the historical development of LED discussed in the literature. The proposed classification of LED technology domains is therefore meaningful, easy to interpret, and can be exploited for a richer and insightful analysis of the technology.

As compared to the results of previous studies on LED, the classification presented in this paper relies on a much larger number of patents, focuses on the most diffused IPC codes in the database and their actual co-occurrence in the patent corpus, and is based on unsupervised network partitioning with no need for direct inspection into the patent documents or manual classification to refine the identified technology domains. The list of technology domains partially differs from the classifications produced by other studies. The comparison with the results previously obtained by Simons and Sandersons (2011) is particularly interesting. The authors developed a classification of LED domains based on a database of patents which, although smaller in size, was built with the same keywords search strategy adopted for this paper. The automated IPC-based community detection analysis enables to detect technology domains related to different applications of LED (e.g. LED technologies for watches, signalling, medical and games technologies, etc.), which are not accounted for by Simons and Sanderson. Conversely, the classification developed by those researchers is more oriented at providing a finer-grained distinction of different fundamental technologies related to LED (especially on semiconductors). This more detailed classification can be achieved by the community analysis too, if the resolution parameter is increased.

Overall, being entirely data-driven, based on the co-occurrence of IPC codes, the community detection methodology provides a more agnostic classification of the technology, but at the same time allows flexibility in the choice of the level of detail: in fact, the algorithm can be adjusted to detect either more general communities, or smaller and more specific domains. On these grounds, this method is considered particularly suitable when there is the need to study the evolution of complex technologies, such as GPT or other multi-purpose technologies, which evolved over a long time period and across several multiple directions that may not be entirely known ex-ante.

Finally, the methodology has two additional advantages worth highlighting. First, it is efficient: by clustering IPC codes included in patent applications, rather than clustering patent applications based on their technology codes, it requires much less computing power. Second, it is easy-to-use: the most common community detection algorithms today are embedded in different applications, which means that the analyst does not require particularly advanced data science skills.

Two possible avenues for future research can be envisaged. On the one hand, the methodology can be further explored and tested with other technologies, and its results can be compared with alternative methods. Community detection can be especially useful when the focus is on multi-purpose or GPT technologies, or when the technology landscape to investigate is not pre-defined around an individual technological field (in line with recent research by Nomaler and Verspagen, 2019). When the scope of analysis is very broad and the interest is in looking at a multitude of technological trajectories and a very large set of patents, community detection analysis can help discover the different technology fields “hidden” in the patent landscape. On the other hand, the results obtained on the LED technology can be brought forward and used for a deeper analysis of LED. The relatedness of different domains to each other can be examined more in-depth. How the technology evolution moved from one technology domain to another in the course of its changing process, how different domains interacted with each other to push the technology advancement forward, how companies positioned in the landscape, which technology domains they occupied, are all other interesting questions for future study.

Appendixes

A.I Patent search strategy

The identification of LED-related patents relied on the search of a number of keyword in the patents' title and /or abstract. The keyword selection was based on Simons and Sanderson (2011). The search required whole words to be identified separated by spaces or non-alphanumeric characters such as punctuation marks. An "(s)" or "(n)" denotes an optional s or n, and an asterisk allows any characters. Searches were not case-sensitive.

The following English-language keyword strings were used to search titles and/or abstract (all with a space both before and after): "light emitting diode(s)", "light-emitting diode(s)", "LED(s)", "OLED(s)", "PLED(s)", "L.E.D.(s)", "LED-based", "semiconductor light emitting", "semiconductor light-emitting", "semiconductor light emission", "semiconductor lighting", "semiconductor lumin*", "solid state lighting", "solid-state lighting", "solid state light(s)", "solid-state light(s)", "solid state lamp(s)", "solid-state lamp(s)", "micro-LED(s)", "light emitting die(s)", "light-emitting die(s)", "luminescent diode(s)", "light emittingdiode(s)", "lightemitting diode(s)", "lightemittingdiode(s)".

The following basic non-English-language keyword strings were also used: "lichtemittierende Diode(n)", "Leuchtdiode(n)", "diode(s) luminescente(s)", "diodo(s) electroluminoso(s)", "diodo(s) luminoso(s)", "diodo luminescente", "diodi luminescenti", "luminescente diodo", "luminescenti diodi", "diodos emissores de luz", "diodo(s) emitindo-se claro(s)".

The resulting database was then cleaned by removing the patent applications where the word "LED" used in combination with the following prepositions/verbs, and no other relevant keywords appear in either the Title or Abstract: " led to "," led in "," led by "," led with "," led up to "," led on "," led into "," led onto "," led from "," led over "," led through "," is led "," are led "," be led "," being led "," LED TO "," LED IN "," LED BY "," LED WITH "," LED UP TO "," LED ON "," LED INTO "," LED ONTO "," LED FROM "," LED OVER "," LED THROUGH "," IS LED "," BE LED "," ARE LED "," BEING LED ". The software STATA was used to perform this cleaning process. Since STATA is case-sensitive, it allowed the word "led" (more likely used as a verb) to be distinguished from "LED" (more likely the acronym of Light-Emitting Diode).

A.II Results of the community analysis of IPC codes

Table 13. Table 1. Results of the community analysis of IPC codes with Louvain and VOS methods at different resolution limits (0.2 – 1.75)

LABEL IPC	IPC CODE	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=0.500000, Q=0.630172, NC=3)	MULTI-LEVEL VOS CLUSTERING (100, RES=0.500000, VOS=0.6561409612, NC=3)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=0.750000, Q=0.536326, NC=6)	MULTI-LEVEL VOS CLUSTERING (100, RES=0.750000, VOS=0.5465395278, NC=4)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=1.000000, Q=0.478736, NC=7)	MULTI-LEVEL VOS CLUSTERING (100, RES=1.000000, VOS=0.4740107147, NC=7)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=1.250000, Q=0.438830, NC=8)	MULTI-LEVEL VOS CLUSTERING (100, RES=1.250000, VOS=0.4361584247, NC=8)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=1.500000, Q=0.400210, NC=9)	MULTI-LEVEL VOS CLUSTERING (100, RES=1.500000, VOS=0.4052666458, NC=10)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=1.750000, Q=0.367501, NC=10)	MULTI-LEVEL VOS CLUSTERING (100, RES=1.750000, VOS=0.3809649159, NC=12)
V1	F21V	1	1	1	1	1	1	1	1	1	1	1	1
V2	F21Y	1	1	1	1	1	1	1	1	1	1	1	1
V3	F21S	1	1	1	1	1	1	1	1	1	1	1	1
V4	H01L	1	2	2	2	2	2	2	2	2	2	2	2
V5	H05B	1	3	1	2	1	3	1	3	2	3	2	3
V6	F21W	1	1	1	1	1	1	1	1	1	1	1	1
V7	F21K	1	1	1	1	1	1	1	1	1	1	1	1
V8	G09F	2	3	2	2	2	4	2	4	2	4	3	4
V9	G09G	2	3	2	2	2	4	2	4	2	4	3	4
V10	G02F	2	2	2	2	2	4	2	4	2	4	3	4
V11	G02B	2	2	2	2	2	4	2	4	3	4	4	4
V12	H04N	2	2	2	3	2	4	2	4	3	5	4	5
V13	G06F	2	3	3	4	2	5	3	4	4	6	5	6
V14	G01R	2	3	3	4	3	3	4	3	5	3	6	3
V15	H05K	2	3	2	2	2	2	2	2	2	2	3	4
V16	F21L	1	3	1	4	1	5	1	5	1	7	1	1
V17	B60Q	2	3	3	4	4	6	5	6	6	8	7	7
V18	G01N	2	3	3	4	4	6	5	6	6	8	7	8
V19	G03B	2	2	2	3	2	4	2	4	3	5	4	5
V20	H02J	2	3	3	4	3	3	4	3	5	3	6	3
V21	A61B	2	3	4	4	5	5	6	5	7	7	8	9
V22	G08B	2	3	3	4	3	5	3	5	4	7	6	10
V23	C09K	3	2	5	2	6	2	7	2	2	2	2	2
V24	B41J	2	2	2	3	2	4	2	4	3	5	4	5
V25	H01S	2	2	2	2	2	2	2	2	2	2	2	2
V26	H04B	2	3	3	4	3	5	3	7	4	6	5	6
V27	H01R	2	3	3	4	3	3	4	3	5	3	6	3
V28	A63F	2	3	3	4	5	5	3	5	4	7	5	9
V29	G03G	2	2	2	3	2	4	2	4	3	5	4	5
V30	A01G	2	3	3	4	5	5	6	5	7	7	8	11
V31	H04M	2	3	3	4	3	5	3	7	4	6	5	6
V32	G08G	2	3	3	4	3	5	5	5	6	9	6	10
V33	G06K	2	3	3	4	2	5	3	4	3	5	4	5
V34	H01H	2	3	3	4	3	3	4	3	5	3	6	3
V35	H02M	2	3	3	4	3	3	4	3	5	3	6	3
V36	G09B	2	3	3	4	5	5	6	5	7	7	8	9
V37	G05B	2	3	3	4	3	5	3	6	4	8	5	11

V38	B60R	2	3	3	4	4	6	5	6	6	8	7	7
V39	H01J	3	2	5	2	6	2	7	2	2	2	2	2
V40	A61N	2	3	4	4	5	5	6	5	7	7	8	9
V41	E01F	2	3	3	4	3	5	4	5	5	9	6	10
V42	H03K	2	3	3	4	3	3	4	3	5	3	6	3
V43	G01B	2	3	3	4	4	6	5	6	6	8	7	8
V44	G01D	2	3	3	4	4	6	5	6	6	8	7	7
V45	G01J	2	3	3	4	4	6	5	6	6	8	7	8
V46	A01M	2	3	3	4	5	5	6	5	7	7	8	11
V47	A01K	2	3	3	4	5	5	6	5	7	7	8	11
V48	A47G	2	3	3	4	5	5	6	5	7	7	8	11
V49	H02H	2	3	3	4	3	3	4	3	5	3	6	3
V50	C08L	3	2	5	2	6	2	7	2	8	2	9	2
V51	G08C	2	3	3	4	3	5	3	6	4	8	5	6
V52	A61L	2	3	3	4	5	5	6	5	7	7	8	11
V53	H04L	2	3	3	4	3	5	3	7	4	6	5	6
V54	A63B	2	3	3	4	5	5	6	5	7	7	8	9
V55	F24F	2	3	3	4	5	5	6	5	7	7	8	11
V56	A47B	2	3	3	4	5	5	6	5	7	7	8	11
V57	C07D	3	2	5	2	6	2	7	2	2	2	2	2
V58	H04R	2	3	3	4	3	5	3	7	4	6	5	6
V59	G11B	2	3	3	4	2	5	3	7	4	6	5	6
V60	G01M	2	3	3	4	4	6	5	6	6	8	7	8
V61	C08K	3	2	5	2	6	2	7	2	8	2	9	2
V62	B62J	2	3	3	4	4	6	5	6	6	8	7	7
V63	G05D	2	3	3	4	4	5	5	6	6	8	7	11
V64	G01C	2	3	3	4	4	6	5	6	6	8	7	8
V65	A61M	2	3	4	4	5	5	6	5	7	7	8	9
V66	H04Q	2	3	3	4	3	5	3	7	4	6	5	6
V67	G06T	2	3	2	3	2	4	2	4	3	5	4	5
V68	A63H	2	3	3	4	5	5	6	5	7	7	8	9
V69	B29C	3	2	5	2	6	2	7	2	8	2	9	2
V70	G01S	2	3	3	4	4	6	5	6	6	8	7	8
V71	G05F	2	3	3	4	3	3	4	3	5	3	6	3
V72	H04W	2	3	3	4	3	5	3	7	4	6	5	6
V73	H02S	2	3	3	4	3	5	4	5	5	9	6	10
V74	C08G	3	2	5	2	6	2	7	2	8	2	9	2
V75	E04H	2	3	3	4	3	5	4	5	5	9	6	10
V76	F25D	2	3	3	4	5	5	6	5	7	7	8	11
V77	A45B	2	3	3	4	5	5	6	5	7	7	8	9
V78	B60K	2	3	3	4	4	6	5	6	6	8	7	7
V79	C23C	3	2	5	2	6	2	7	2	2	2	2	2
V80	A45C	2	3	3	4	3	5	6	5	7	7	8	11
V81	E05B	2	3	3	4	3	5	3	5	4	7	5	7
V82	B23K	2	2	2	2	2	2	2	2	2	2	3	2
V83	B32B	3	2	5	2	6	2	7	2	8	2	9	2
V84	G07C	2	3	3	4	3	5	3	5	4	7	5	7
V85	A61H	2	3	4	4	5	5	6	5	7	7	8	9
V86	A41D	2	3	3	4	5	5	6	5	7	7	8	9
V87	G04G	2	3	6	4	7	7	8	8	9	10	10	12
V88	G01F	2	3	3	4	4	5	5	6	6	8	7	11

V89	G01K	2	3	3	4	4	5	5	6	6	8	7	11
V90	B65D	2	3	3	4	5	5	6	5	7	7	8	11
V91	G07F	2	3	3	4	3	5	3	5	4	7	5	9
V92	G06Q	2	3	3	4	3	5	3	5	4	7	5	9
V93	H01M	2	3	3	4	3	3	4	3	5	3	6	3
V94	G04B	2	3	6	4	7	7	8	8	9	10	10	12
V95	A47F	2	3	3	4	5	5	6	5	7	7	8	11
V96	E04F	2	3	3	4	3	5	4	5	5	9	6	10
V97	C09D	3	2	5	2	6	2	7	2	8	2	9	2
V98	A45D	2	3	4	4	5	5	6	5	7	7	8	11
V99	G03F	3	2	5	2	6	2	7	2	8	2	9	2
V100	H02B	2	3	3	4	3	3	4	3	5	3	6	3

Note: Clustering run in Pajek. For both the Louvain and VOS method, we used the standard clustering parameters, namely: Number of Restarts: 100; Maximum Number of Iterations in each Restart: 20; Maximum Number of Levels in each Iteration: 20, Maximum Number of Repetitions in each Level: 50.

Table 14. Table 2. Results of the community analysis of IPC codes with Louvain and VOS methods at different resolution limits (2 - 3.5)

LABEL IPC	IPC CODE	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=2.000000, Q=0.340733, NC=14)	MULTI-LEVEL VOS CLUSTERING (100, RES=2.000000, VOS=0.3629626144, NC=12)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=2.250000, Q=0.316913, NC=14)	MULTI-LEVEL VOS CLUSTERING (100, RES=2.250000, VOS=0.3458821204, NC=14)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=2.500000, Q=0.294606, NC=15)	MULTI-LEVEL VOS CLUSTERING (100, RES=2.500000, VOS=0.3295914061, NC=15)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=2.750000, Q=0.270942, NC=17)	MULTI-LEVEL VOS CLUSTERING (100, RES=2.750000, VOS=0.3147972153, NC=18)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=3.000000, Q=0.249055, NC=18)	MULTI-LEVEL VOS CLUSTERING (100, RES=3.000000, VOS=0.3015173093, NC=22)	MULTI-LEVEL LOUVAIN COMMUNITIES (100, RES=3.250000, Q=0.226432, NC=21)	MULTI-LEVEL VOS CLUSTERING (100, RES=3.250000, VOS=0.2915801813, NC=25)	MULTI-LEVEL LOUVAIN COMMUNITIES 1 (100, RES=3.500000, Q=0.206092, NC=23)	MULTI-LEVEL VOS CLUSTERING (100, RES=3.500000, VOS=0.2838107363, NC=26)
V1	F21V	1	1	1	1	1	1	1	1	1	1	1	1	1	1
V2	F21Y	1	1	1	1	1	1	1	1	1	1	1	1	1	1
V3	F21S	1	1	1	1	1	1	1	1	1	1	1	1	1	1
V4	H01L	2	2	2	2	2	2	2	2	2	2	2	2	2	2
V5	H05B	2	3	3	3	2	3	3	3	3	3	3	3	3	3
V6	F21W	1	1	1	1	1	1	1	1	1	1	1	1	1	1
V7	F21K	1	1	1	1	1	1	1	1	1	1	1	4	1	4
V8	G09F	3	4	4	4	3	4	4	4	4	4	4	5	4	5
V9	G09G	3	4	4	4	3	4	4	4	4	4	4	5	4	6
V10	G02F	3	4	4	4	3	4	4	4	4	4	5	5	5	6
V11	G02B	4	4	5	4	4	4	4	5	4	4	5	5	5	6
V12	H04N	4	5	5	5	4	5	5	5	5	5	6	6	6	7
V13	G06F	5	6	6	6	5	6	6	6	6	6	7	7	7	8
V14	G01R	6	3	7	7	6	7	7	7	7	7	8	8	8	9
V15	H05K	3	4	4	4	3	4	4	4	4	8	4	9	4	5
V16	F21L	1	1	8	1	7	8	8	8	8	9	9	10	9	10
V17	B60Q	7	7	9	8	8	9	9	9	9	10	10	11	10	11
V18	G01N	8	8	7	9	9	10	10	10	10	7	8	8	8	9
V19	G03B	4	5	5	5	4	5	5	5	5	5	6	6	6	7
V20	H02J	6	3	8	7	6	7	7	7	7	9	11	10	9	10
V21	A61B	9	9	10	10	10	11	11	11	11	11	12	12	11	12
V22	G08B	10	10	8	11	11	12	12	12	12	12	13	13	12	13
V23	C09K	2	2	2	2	2	2	2	13	2	13	2	14	2	14
V24	B41J	4	5	5	5	4	5	5	5	5	5	6	6	6	7
V25	H01S	2	2	2	2	2	2	2	2	2	2	2	2	2	2
V26	H04B	11	6	11	12	12	13	13	14	13	14	14	15	13	15
V27	H01R	6	3	3	7	6	7	7	7	7	8	15	9	14	5
V28	A63F	5	9	6	6	5	6	6	15	6	15	16	16	15	16
V29	G03G	4	5	5	5	4	5	5	5	5	5	6	6	6	7
V30	A01G	12	11	12	13	13	14	14	16	14	16	17	17	16	17
V31	H04M	11	6	11	12	12	13	13	14	13	14	14	15	13	15
V32	G08G	10	10	8	11	11	12	12	12	12	17	11	18	17	18
V33	G06K	5	5	6	6	5	6	6	6	6	5	7	7	18	8
V34	H01H	6	3	3	7	6	7	7	7	7	18	15	19	14	19
V35	H02M	6	3	3	7	6	7	7	7	7	18	15	19	14	19
V36	G09B	9	9	10	6	10	6	11	15	11	15	16	16	15	16
V37	G05B	12	11	12	13	13	14	15	16	15	19	18	20	19	20
V38	B60R	7	7	9	8	8	9	9	9	9	10	10	11	10	11
V39	H01J	2	2	2	2	2	2	2	2	2	2	2	2	2	2
V40	A61N	9	9	10	10	10	11	11	11	11	11	12	12	11	12

V41	E01F	10	10	8	11	11	12	12	12	12	17	11	18	17	18
V42	H03K	6	3	3	7	6	7	7	7	7	18	15	19	14	19
V43	G01B	8	8	7	9	9	10	10	10	10	20	19	21	20	21
V44	G01D	7	7	9	8	8	9	9	9	9	10	10	11	10	11
V45	G01J	8	8	7	9	9	10	10	10	10	7	8	8	8	9
V46	A01M	12	11	12	13	13	14	14	16	14	16	17	17	16	17
V47	A01K	12	11	12	13	13	14	14	16	14	16	17	17	16	17
V48	A47G	9	11	12	13	13	11	14	11	14	16	17	22	16	22
V49	H02H	6	3	3	7	6	7	7	7	7	18	15	19	14	19
V50	C08L	13	2	13	2	14	2	16	17	16	21	20	23	21	23
V51	G08C	11	6	11	12	12	13	13	14	13	19	14	20	13	20
V52	A61L	9	11	12	13	13	14	14	16	14	16	17	22	16	22
V53	H04L	11	6	11	12	12	13	13	14	13	14	14	15	13	15
V54	A63B	9	9	10	6	10	6	11	15	11	15	16	16	15	16
V55	F24F	9	11	12	13	13	14	14	16	14	16	17	22	16	22
V56	A47B	9	11	12	13	13	14	14	16	14	16	17	22	16	22
V57	C07D	2	2	2	2	2	2	2	13	2	13	2	14	2	14
V58	H04R	11	6	11	12	12	13	13	14	13	14	14	24	13	24
V59	G11B	5	6	6	12	5	13	6	6	6	14	7	24	7	24
V60	G01M	8	8	7	9	9	10	10	10	10	7	8	8	8	9
V61	C08K	13	2	13	2	14	2	16	17	16	21	20	23	21	23
V62	B62J	7	7	9	8	8	9	9	9	9	10	10	11	10	11
V63	G05D	12	11	12	13	13	14	15	16	15	19	18	20	19	20
V64	G01C	8	8	7	9	9	10	10	10	10	20	19	21	20	21
V65	A61M	9	9	10	10	10	11	11	11	11	11	12	12	11	12
V66	H04Q	11	6	11	12	12	13	13	14	13	14	14	15	13	15
V67	G06T	4	5	5	5	4	5	5	6	5	5	6	7	6	8
V68	A63H	9	9	10	6	10	6	11	15	11	15	16	16	15	16
V69	B29C	13	2	13	2	14	2	16	2	16	2	20	2	21	2
V70	G01S	8	8	7	9	9	10	10	10	10	20	19	21	20	21
V71	G05F	6	3	3	7	6	7	7	7	7	18	15	19	14	19
V72	H04W	11	6	11	12	12	13	13	14	13	14	14	15	13	15
V73	H02S	10	10	8	11	11	12	12	12	12	17	11	18	17	18
V74	C08G	13	2	13	2	14	2	16	17	16	21	20	23	21	23
V75	E04H	10	10	8	11	11	12	12	12	12	17	11	18	17	18
V76	F25D	9	11	12	13	13	14	14	16	14	16	17	22	16	22
V77	A45B	9	9	10	6	10	6	11	15	11	15	16	16	15	16
V78	B60K	7	7	9	8	8	9	9	9	9	10	10	11	10	11
V79	C23C	2	2	2	2	2	2	2	2	2	2	2	2	2	2
V80	A45C	9	11	8	13	10	11	11	11	11	16	13	22	12	22
V81	E05B	5	7	6	6	5	6	6	15	6	15	7	16	18	25
V82	B23K	8	2	2	2	9	2	2	2	17	2	19	2	22	2
V83	B32B	13	2	13	2	14	2	16	2	16	2	20	2	21	2
V84	G07C	5	7	6	6	5	6	6	15	6	15	7	16	18	25
V85	A61H	9	9	10	10	10	11	11	11	11	11	12	12	11	12
V86	A41D	9	9	10	10	10	11	11	11	11	11	13	12	12	12
V87	G04G	14	12	14	14	15	15	17	18	18	22	21	25	23	26
V88	G01F	8	11	12	13	13	11	9	11	15	19	18	20	19	20
V89	G01K	8	11	12	13	13	11	9	11	15	19	18	20	19	20
V90	B65D	9	11	12	13	13	11	14	11	14	16	17	22	16	22
V91	G07F	5	9	6	6	5	6	6	15	6	15	7	16	18	16

V92	G06Q	5	9	6	6	5	6	6	15	6	15	7	16	18	16
V93	H01M	6	3	8	7	6	7	7	7	7	9	11	10	9	10
V94	G04B	14	12	14	14	15	15	17	18	18	22	21	25	23	26
V95	A47F	9	11	12	13	13	14	14	16	14	16	17	22	16	22
V96	E04F	10	10	8	11	11	12	12	12	12	17	11	18	17	18
V97	C09D	13	2	13	2	14	2	16	2	16	2	20	23	21	23
V98	A45D	9	11	10	13	10	11	11	11	11	16	12	22	11	22
V99	G03F	13	2	13	2	14	2	16	2	16	2	20	2	22	2
V100	H02B	6	3	3	7	6	7	7	7	7	18	15	19	14	19

Note: Clustering run in Pajek. For both the Louvain and VOS method, we used the standard clustering parameters, namely: Number of Restarts: 100; Maximum Number of Iterations in each Restart: 20; Maximum Number of Levels in each Iteration: 20, Maximum Number of Repetitions in each Level: 50.

A.III Relation between the Louvain and VOS clustering methods

Figure 13: Cramers' V index at different resolution levels

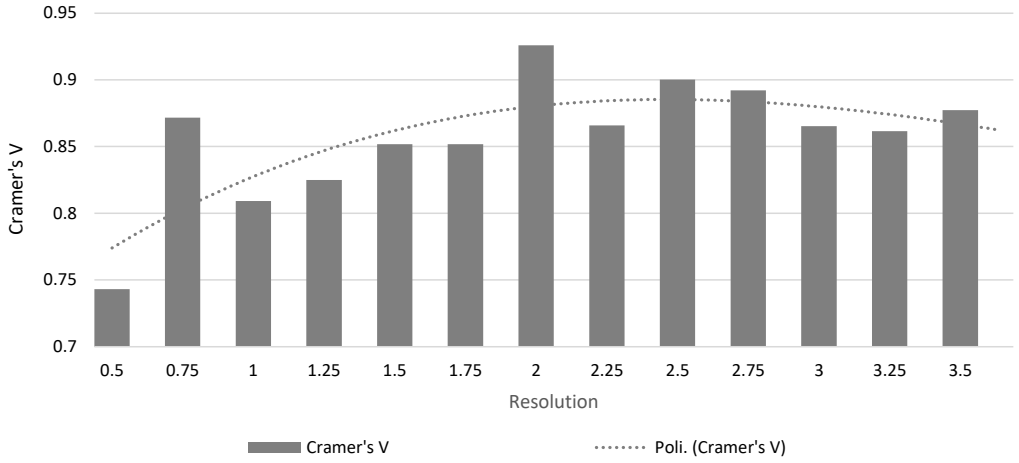


Figure 14: Rajski indexes at different resolution levels

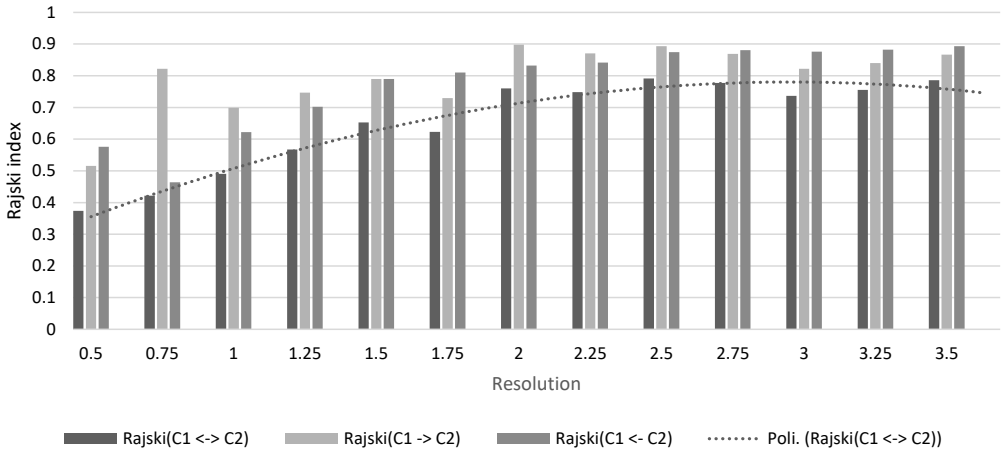
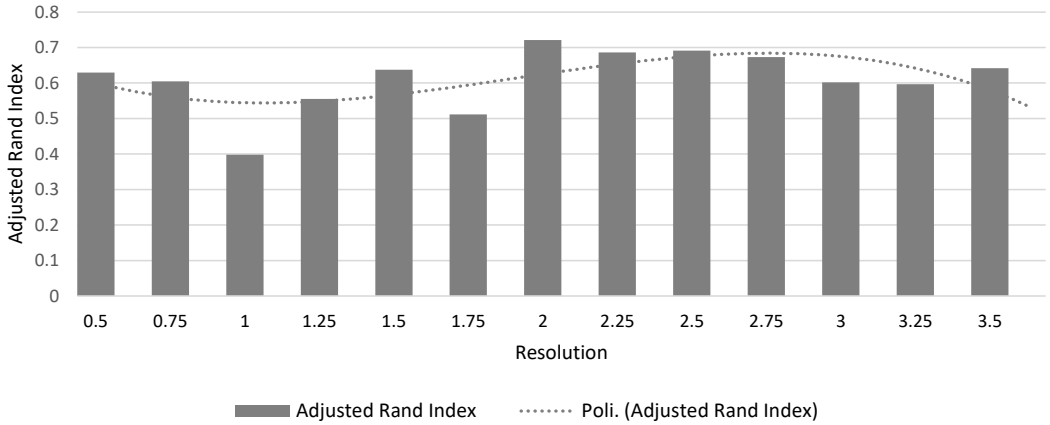


Figure 15: Adjusted Rand Index (ARI) at different resolution levels



List of references

Aharonson, B.S., Schilling, M.A., 2016. Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution. Research Policy 45, 81–96. https://doi.org/10.1016/j.respol.2015.08.001

Aldecoa, R. and Marín, I., 2013. Surprise maximization reveals the community structure of complex networks. Scientific Reports, vol. 3, no. 1030. doi: 10.1038/srep01060

Alstott, J., Triulzi, G., Yan, B., Luo, J., 2017. Mapping technology space by normalizing patent networks. Scientometrics 110, 443–479. https://doi.org/10.1007/s11192-016-2107-y

Altuntas, S., Dereli, T., Kusiak, A. 2015. Forecasting technology success based on patent data. Technological Forecasting and Social Change 96, 202-214. https://doi.org/10.1016/j.techfore.2015.03.011

Ardito, L., D'Adda, D., Messeni Petruzzelli, A., 2018. Mapping innovation dynamics in the Internet of Things domain: Evidence from patent analysis. Technological Forecasting and Social Change 136, 317–330. https://doi.org/10.1016/j.techfore.2017.04.022

Arts, S., Cassiman, B., Gomez, J.C., 2018. Text matching to measure patent similarity. Strategic Management Journal 39, 62–84. https://doi.org/10.1002/smj.2699

Bekar, C., Carlaw, K., Lipsey, R., 2018. General purpose technologies in theory, application and controversy: a review. J Evol Econ 28, 1005–1033. https://doi.org/10.1007/s00191-017-0546-0

Benson, C.L., Magee, C.L., 2013. A hybrid keyword and patent class methodology for selecting relevant sets of patents for a technology field. Scientometrics 96, 69–82.

Benson, C.L., Magee, C.L., 2015. Technology structural implications from the extension of a patent search method. Scientometrics 102: 1965-1985. DOI 10.1007/s11192-014-1493-2

Bergeaud, A., Potiron, Y., Raimbault, J., 2017. Classifying patents based on their semantic content. PLOS ONE 12, e0176310. https://doi.org/10.1371/journal.pone.0176310

Bessen, J., Hunt, R.M., 2007. An Empirical Look at Software Patents. Journal of Economics and Management Strategy 16, 157-189. https://doi.org/10.1111/j.1530-9134.2007.00136.x

Bissmark, J., Wärnling, O., 2017. The sparse data problem within classification algorithms: The effect of sparse data on the naïve Bayes algorithm. Bachelor's Thesis in Computer Science at KTH, School of Computer Science and Communication (CSC), DD142x, Stockholm, Sweden. Available at https://www.diva-portal.org/smash/get/diva2:1111045/FULLTEXT01.pdf Last access on 30/01/2022

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

Boyack, K.W., Tsao, J.Y., Miksovic, A., Huey, M., 2009. A recursive process for mapping and clustering technology literatures: Case study in solid-state lighting. International Journal of Technology Transfer and Commercialisation 8, 51–87. https://doi.org/10.1504/IJTTC.2009.023434

Bresnahan, F., Trajtenbergb, M., 1995. General purpose technologies 'Engines of growth'? Journal of Econometrics 65, 83–108. https://doi.org/10.1016/0304-4076(94)01598-T

Bresnahan, T.F., Greenstein, S., 2003. Technological Competition and the Structure of the Computer Industry. The Journal of Industrial Economics 47, 1–40. https://doi.org/10.1111/1467-6451.00088

Celebi, M. E., Kingravi, H.A., Vela, P.A. 2013. A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm. Expert Systems with Applications 40 200-210. 10.1016/j.eswa.2012.07.021

Chang, S. B., 2012. Using patent analysis to establish technological position: Two different strategic approaches. Technological Forecasting and Social Change 79, 3-15. https://doi.org/10.1016/j.joi.2011.09.001

Chen, C., Fang, W., Hsu, S.S., 2016. A study on technological trajectory of light emitting diode in Taiwan by using patent data. International Journal of Technology Management 72, 83-104. https://doi.org/10.1504/IJTM.2016.080548

Cheung, Y-M. 2003. K-Means: A new generalized k-means clustering algorithm. Pattern Recognition Letters, 24: 2883-2893. https://doi.org//10.1016/S0167-8655(03)00146-6

Choi S., Lee., H., Park, E., Choi, S., 2022. Deep learning for patent landscaping using transformer and graph embedding. Technological Forecasting and Social Change 175, 121413. 10.1016/j.techfore.2021.121413

Choi, J., Hwang, Y.-S., 2014. Patent keyword network analysis for improving technology development efficiency. Technological Forecasting and Social Change 83, 170–182. https://doi.org/10.1016/j.techfore.2013.07.004

Clauset, A., Newman, M. E. and Moore, C., 2004. Finding community structure in very large networks. Physical Review, E 70, 066111

Dosi, G., 1982. Technological paradigms and technological trajectories. A Suggested Interpretation of the Determinants and Directions of Technical Change. Research Policy 11, 147 – 162. https://doi.org/10.1016/0048-7333(82)90016-6

Dosi, G., Nelson, R.R., 2010. Technical change and industrial dynamics as evolutionary processes. In: Bronwyn H. H., Rosenberg, N. (Eds) Handbook of the Economics of Innovation, 2010, vol. 1. Amsterdam: North-Holland, 51-128. https://doi.org/10.1016/S0169-7218(10)01003-8

Dosi, G., Nelson, R.R., 2013. The evolution of technologies: an assessment of the state-of-the-art. Eurasian Business Review 3, 3–46. https://doi.org/10.14208/BF03353816

Epicoco, M., 2013. Knowledge patterns and sources of leadership: Mapping the semiconductor miniaturization trajectory. Research Policy 42, 180–195. https://doi.org/10.1016/j.respol.2012.06.009

Érdi, P., Makovi, K., Somogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P., Zalányi, L., 2013. Prediction of emerging technologies based on analysis of the US patent citation network. Scientometrics 95, 225–242. https://doi.org/10.1007/s11192-012-0796-4

Ernst, H., 2003. Patent information for strategic technology management. World patent information 25, 233-242. https://doi.org/10.1016/S0172-2190(03)00077-2

Fleming, L., Sorenson, O., 2004. Science as a map in technological search. Strategic Management Journal, 25, 909-928. https://doi.org/10.1002/smj.384

Fontana R., Nuvolari A., Verspagen B., 2009. Mapping technological trajectories as patent citation networks. An application to data communication standards. Economics of Innovation and New Technology 18, 311-336. https://doi.org/10.1080/10438590801969073

Gao, Y., Zhu, Z., Kali, R., Riccaboni, M., 2018. Community evolution in patent networks: technological change and network dynamics. Applied network science 3, 1-23. https://doi.org/10.1007/s41109-018-0090-3

Gerken, J. M., Moehrle, M. G., 2012. A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. Scientometrics 91, 645-670. https://doi.org/10.1007/s11192-012-0635-7

Girvan, M. and Newman, M. E., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences 99, 7821–7826.

Greve, H.R., 2000. MARKETING NICHE ENTRY DECISIONS: COMPETITION, LEARNING, AND STRATEGY IN TOKYO BANKING, 1894-1936. Academy of Management Journal 43, 816–836. https://doi.org/10.2307/1556412

Gridlogics Technologies Pvt Ltd, 2010. Technology Insight Report. LEDs in Lighting. Available at http://www.patentinsightpro.com/techreports/0410/Gridlogics%20Technology%20Insight%20Report-LEDs%20in%20Lighting.pdf Last access on 30/01/2022

Griliches, Z., 1990. Patent statistics as economic indicators: A survey. Journal of Economic Literature 28, 1661–1707

Grupp, H., 1990. The concept of entropy in scientometrics and innovation research: An indicator for institutional involvement in scientific and technological developments. Scientometrics 18, 219-239. https://doi.org/10.1007/bf02017763

Hall, B. H., Ziedonis, R. H., 2001. The patent paradox revisited: an empirical study of patenting in the US semiconductor industry, 1979-1995. RAND Journal of Economics 32, 101-128. https://doi.org/10.2307/2696400

Heller, M.A., Eisenberg, R.S., 1998. Can Patents Deter Innovation? The Anticommons in Biomedical Research. Science 280, 698–701. https://doi.org/10.1126/science.280.5364.698

Hu, Jie, Li, S., Yao, Y., Yu, L., Yang, G., Hu, Jianjun, 2018. Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification. Entropy 20, 104-123. https://doi.org/10.3390/e20020104

iRunway, 2014. LED Patent Landscape. Report prepared by iRunway. Available at https://pdf4pro.com/view/led-patent-landscape-irunway-56e835.html Last access on 30/01/2022

Joo, S.H., Kim, Y., 2010. Measuring relatedness between technological fields. Scientometrics 83, 435–454. https://doi.org/10.1007/s11192-009-0108-9

Joung, J., Kim, K. 2017. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technological Forecasting and Social Change 114, 281-292. https://doi.org/10.1016/j.techfore.2016.08.020

Jovanovic, B., Rob, R., 1990. Long Waves and Short Waves: Growth Through Intensive and Extensive Search. Econometrica 58, 1391. https://doi.org/10.2307/2938321

Jun, S., Park, S.-S., Jang, D.-S., 2014. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Systems with Applications 41, 3204–3212. https://doi.org/10.1016/j.eswa.2013.11.018

Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., 2002. An Efficient K-Means Clustering Algorithm Analysis and Implementation. IEEE Transactions On Pattern Analysis and Machine Intelligence, 24: 881-892. https://doi.org/10.1109/TPAMI.2002.1017616

Kauffman, S., Lobo, J., Macready, W.G., 2000. Optimal search on a technology landscape. Journal of Economic Behaviour & Organization, 43, 141-166

Kim, G., Bae, J., 2017. A novel approach to forecast promising technology through patent analysis. Technological Forecasting and Social Change 117, 228–237. https://doi.org/10.1016/j.techfore.2016.11.023

King, A.A., Tucci, C.L., 2002. Incumbent entry into new market niches: The role of experience and managerial choice in the creation of dynamic capabilities. Management science 48, 171–186. https://doi.org/10.1287/mnsc.48.2.171.253

Korzinov, V., Savin, I., 2016. Pervasive enough? General purpose technologies as an emergent property. KIT Working Paper Series in Economics No. 95. Karlsruher Institut für Technologie (KIT), Institut für Volkswirtschaftslehre (ECON), Karlsruhe. http://hdl.handle.net/10419/147981

Lipsey, R. G., Carlaw, K. I., Bekar, C. T., 2005. Economic transformations: general purpose technologies and long-term economic growth. New York, NY: Oxford University Press. ISBN 9780199285648

Long, M., Ma, T., 2015. On a patent analysis method for identifying the core technologies of metro in China, in: Proceedings of the 5th International Conference on Civil Engineering and Transportation 2015. Presented at the 5th International Conference on Civil Engineering and Transportation, Atlantis Press, Guangzhou, China. https://doi.org/10.2991/iccet-15.2015.317

Lupu, M., Mayer, K., Kando, N., Trippe, A. J. (Eds.), 2017. Current challenges in patent information retrieval (Vol. 37). Heidelberg: Springer. ISBN: 978-3-662-53817-3

Madani, F., Weber, C., 2016. The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. World Patent Information 46, 32–48. https://doi.org/10.1016/j.wpi.2016.05.008

Magee, C.L., Basnet, S., Funk, J.L., Benson, C.L., 2016. Quantitative empirical trends in technical performance. Technol. Forecast. Soc. Change 104, 237–246. https://doi. org/10.1016/j.techfore.2015.12.011.

Marsili, O., Verspagen, B., 2002. Technology and the dynamics of industrial structures: an empirical mapping of Dutch manufacturing. Industrial and corporate change 11, 791–815. https://doi.org/10.1093/icc/11.4.791

Martinelli, A., 2012. An emerging paradigm or just another trajectory? Understanding the nature of technological changes using engineering heuristics in the telecommunications switching industry. Research Policy 41, 414–429. https://doi.org/10.1016/j.respol.2011.10.012

Moser, P., Nicholas, T., 2004. Was electricity a general purpose technology? Evidence from historical patent citations. American Economic Review 94, 388-394. https://doi.org/10.1257/0002828041301407

Nakamura, H., Suzuki, S., Sakata, I., Kajikawa, Y., 2015. Knowledge combination modeling: The measurement of knowledge similarity between different technological domains. Technological Forecasting and Social Change 94, 187–201. https://doi.org/10.1016/j.techfore.2014.09.009

Narasimhan, C., Zhang, Z.J., 2000. Market Entry Strategy Under Firm Heterogeneity and Asymmetric Payoffs. Marketing Science 19, 313–327. https://doi.org/10.1287/mksc.19.4.313.11790

Nelson, R.R., 1995. Why should managers be thinking about technology policy? Strategic Management Journal 16, 571-588. https://doi.org/10.1002/smj.4250160802

Nelson, R.R., Winter, S.G., 2004. An evolutionary theory of economic change, digitally reprinted. ed. Cambridge, MA: The Belknap Press of Harvard University. ISBN 0-674-27228-5 (paper)

Noel, M., Schankerman, M., 2013. Strategic Patenting and Software Innovation: Strategic Patenting and Software Innovation. The Journal of Industrial Economics 61, 481–520. https://doi.org/10.1111/joie.12024

Nomaler, Ö., Verspagen, B., 2019. Greentech homophily and path dependence in a large patent citation network. Maastricht Economic and Social Research Institute on Innovation and Technology (UNU-MERIT) Working Papers N. 36.

Park, S., Jun, S., 2017. Technology Analysis of Global Smart Light Emitting Diode (LED) Development Using Patent Data. Sustainability 9, 1363. https://doi.org/10.3390/su9081363

Pons, P. and Latapy, M., 2005. Computing communities in large networks using random walks. Physics and Society. arXiv:physics/0512106v1 [physics.soc-ph] https://doi.org/10.48550/arXiv.physics/0512106

Raghavan, U. N., Albert, R. and Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Physical Review, E 76 (3), 036106.

Reichardt, J. and Bornholdt, S., 2006. Statistical mechanics of community detection. Physical Review E 74, 016110.

Rizzi, F., Annunziata, E., Liberati, G., Frey, M., 2014. Technological trajectories in the automotive industry: are hydrogen technologies still a possibility? Journal of Cleaner Production 66, 328–336. https://doi.org/10.1016/j.jclepro.2013.11.069

Rosvall, M. and Bergstrom, C. T., 2007. An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences 104, 7327–7331

Sahal D., 1981. Patterns of technological innovation. Reading, MA: AddisonWesley Pub. Co. ISBN 0201066300 9780201066302

Sanderson, S.W., Simons, K.L., 2014. Light emitting diodes and the lighting revolution: The emergence of a solid-state lighting industry. Research Policy 43, 1730–1746. https://doi.org/10.1016/j.respol.2014.07.011

Schumpeter, J.A., 1934. The Theory of Economic Development. London: Oxford University Press. ISBN 9780674879904

Simons, K.L., Sanderson, S.W., 2011. Global Technology Development in Solid State Lighting. International Journal of High Speed Electronics and Systems 20, 359–382. https://doi.org/10.1142/S0129156411006647

Sinanga, K.P., Yang, M-S., 2020. Unsupervised K-Means Clustering Algorithm. IEEE Access PP(99):1-1. 10.1109/ACCESS.2020.2988796

Singh, A., Triulzi, G. and Magee, C.L., 2021. Technological improvement rate predictions for all technologies: Use of patent data and an extended domain description.

Smith, M., Agrawal, R., 2015. A Comparison of Patent Classifications with Clustering Analysis, in: Wang, J., Cellary, W., Wang, D., Wang, H., Chen, S.-C., Li, T., Zhang, Y. (Eds.), Web Information Systems Engineering – WISE 2015. Cham, CH: Springer International Publishing, Cham, pp. 400–413. https://doi.org/10.1007/978-3-319-26187-4_38

Song, K., Kim, K.S., Lee, S., 2017. Discovering new technology opportunities based on patents: Text-mining and F-term analysis. Technovation 60–61, 1–14. https://doi.org/10.1016/j.technovation.2017.03.001

Strumsky, D., Lobo, J., Van der Leeuw, S., 2012. Using patent technology codes to study technological change. Economics of Innovation and New technology 21, 267-286. https://doi.org/10.1080/10438599.2011.578709

Traag, V. A., Waltman, L. and van Eck, N. J., 2019. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports, vol. 9, no. 1: 1–12, 2019, doi: 10.1038/s41598–019–41695-z

Tseng, Y.-H., Lin, C.-J., Lin, Y.-I., 2007. Text mining techniques for patent analysis. Information Processing & Management 43, 1216–1247. https://doi.org/10.1016/j.ipm.2006.11.011

van Eck, N. J., Waltman, L., 2007. Bibliometric mapping of the computational intelligence field. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 15, 625-645. https://doi.org/10.1142/S0218488507004911

van Eck, N. J., Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84: 523-538.

Verspagen, B., 1991. A new empirical approach to catching up or falling behind. Structural change and economic dynamics 2, 359-380

Verspagen, B., 2007. Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. Advances in complex systems 10, 93–115. https://doi.org/10.1142/S0219525907000945

Wang, H., Chi, Y., Hsin, P., 2018. Constructing Patent Maps Using Text Mining to Sustainably Detect Potential Technological Opportunities. Sustainability 10, 3729. https://doi.org/10.3390/su10103729

Ward, J.H. 1963. Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association 58, 236-244. https://doi.org/10.1080/01621459.1963.10500845

Wu, C. C., 2016. Constructing a weighted keyword-based patent network approach to identify technological trends and evolution in a field of green energy: a case of biofuels. Quality & quantity 50, 213-235. https://doi.org/10.1007/s11135-014-0145-1

Yoon, B., Park, Y. 2004. A text-mining-based patent network: Analytical tool for high-technology trend. The Journal of High Technology Management Research 15, 37-50. https://doi.org/10.1016/j.hitech.2003.09.003

Yoon, J., Kim, K., 2012. TrendPerceptor: A property–function based technology intelligence system for identifying technology trends from patents. Expert Systems with Applications 39, 2927-2938. https://doi.org/10.1016/j.eswa.2011.08.154

Zhou, Y., Lin, H., Liu, Y., Ding, W., 2019. A novel method to identify emerging technologies using a semi-supervised topic clustering model: a case of 3D printing industry. Scientometrics 120, 167–185. https://doi.org/10.1007/s11192-019-03126-8

Ziedonis, R.H. (2004), Don't Fence Me In: Fragmented Markets for Technology and the Patent Acquisition Strategies of Firms, in MANAGEMENT SCIENCE, vol. 50, 6, 804-820

Compliance with Ethical Standards

Funding: This research has not received any funding.

Conflict of Interest: There is no conflict of interest to declare.

Ethical Conduct: The research has been conducted in compliance with the Ethical requirements of the Maastricht University, School of Business and Economics. The objective, scope and methodology of this research did not require the formal approval by the Ethical Approval Committee.

Data Availability Statements: The patent data associated with this paper are available upon reasonable request.