Technological Forecasting & Social Change xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Quantifying technological change as a combinatorial process

Pedro Parraguez[a,b,*], Stanko Škec[a,c,a], Duarte Oliveira e Carmo[a], Anja Maier[a]

[a] DTU - Technical University of Denmark, DTU Management, Kongens Lyngby, Denmark
[b] Dataverz ApS, Copenhagen, Denmark
[c] University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture, Zagreb, Croatia

ARTICLE INFO

Keywords:
technological change
technological development
bioenergy
biofuel
methodology
quantitative analysis
patents and inventions
data mining

ABSTRACT

Technological change plays a critical role in industrial- and societal development. As a consequence, modelling, measuring and monitoring the rate and direction of technological change has been extensively studied. However, there is to-date no scalable, cross-domain, and data source agnostic quantitative indicator for technological change that considers a combinatorial process of invention and technology development. This paper develops and empirically tests a network-based method that takes a combinatorial view of technological development as underlying rationale and provides a quantitative indicator for technological change. Unlike prior research, the proposed method allows for the simultaneous inclusion of multiple and diverse types of data sources, i.e. publications, patents, and projects and it uses text-mining analyses based on co-occurrences of terms over time. The novel method proposed here goes beyond reliance on domain experts building custom sets of taxonomies, is applicable across different technological domains, and permits a temporal analysis of changes across years, industries, and countries. This is illustrated using a large database of worldwide bioenergy research and development (R&D) records. Findings include the detection of biofuel generations in the data before they were first mentioned as such in literature. Implications for research, industry and policy are also discussed.

1. Introduction

Quantifying the rate of technological change in a given industry or field is a prerequisite for identifying the speed and strength of technological transformations. It is also a prerequisite for supporting the management of interventions that seek to modify technological trajectories and seek to speed-up or control the degree of technological change (Guan and Liu, 2016; Phillips and Linstone, 2016). A good indicator for technological change helps to answer questions such as: Are we in a period of incremental or radical change? What patterns of technological change exist that can help forecast the future rate and degree of technological change? Despite the usefulness of indicators for technological change, there is to-date no scalable, cross-domain, and data source agnostic quantitative indicator for technological change that incorporates what is considered a crucial element of how technology changes: the combinatorial process of invention and technology development (Arthur, 2009; Youn et al., 2015).

Technologies are characterized as a combination of components or subsystems that are assembled to perform one or more functions, exploiting the structure and behavior of the parts that compose it (e.g. Arthur (2009)). In this view, an important element of technological change is the way in which technologies are configured and components or subsystems that are chosen to produce the desired function are varied (Youn et al., 2015). This has typically been associated with a combinatorial view of technological evolution. However, the underlying dynamics of technological recombination remain insufficiently explored, and a thorough understanding of underlying technological change mechanisms needs to be further developed and improved (Fleming and Sorenson, 2001).

A number of recent research studies imply this is a timely issue, for which there is no easy answer (Nosella et al., 2008). For example, monitoring and prediction of technological changes connected to changes in frequency and direction are often oversimplified, not taking into consideration various influences between existing incremental and radical technological changes (Hekkert et al., 2007). Technological change is commonly understood in terms of incremental and radical shifts, and technological diffusion is analyzed with approaches such as technological trajectories and paradigms (Dolfsma and Leydesdorff, 2009). In a more practical sense, technological change is often manifested in new products, processes, or materials (Jorgenson, 2001; Strumsky et al., 2012) and, as such, tangible traces of products, processes, or materials, e.g. in the form of patents or publications and the like can be used to identify, measure and monitor technological changes (Youn et al., 2015).

* Corresponding author.
E-mail addresses: ppru@dtu.dk, pedro@dataverz.net (P. Parraguez).

https://doi.org/10.1016/j.techfore.2019.119803
Received 1 January 2019; Received in revised form 14 October 2019; Accepted 28 October 2019

Tangible traces that hold information about technologies include patents (Benson and Magee, 2013), project repositories (Moro et al., 2018), and publications (Järvenpää et al., 2011). Such data sources are seen as well-defined objects and proxies that can provide insights into technologies within different application domains (Solé et al., 2016). As such, they codify information about technological changes and enable systematic and quantitative analyses. However, various issues may emerge due to inadequate descriptions of technologies and the complexity of mutual influences. To address this, significant effort ought to be placed on the implementation of a combinatorial perspective and the development of approaches that use combinatorial principles.

Although previous research has adopted a number of approaches to analyze and describe technological change, to this date, there is no convergence to a single commonly accepted standard approach, method or indicator. Depending on the required depth and breadth of the studied phenomenon, both qualitative and quantitative approaches have been used to explore technological changes and development (Popper, 2008). The predominant approach nowadays is based on quantitative analyses of information embedded in different digital databases (Funk and Owen-Smith, 2017). These quantitative analyses have been considered as a more systematic and objective manner for analyzing technological changes in comparison to qualitative approaches (Lee et al., 2011). By using objective information about analyzed technologies, these analyses allow a methodologically more consistent, structured and repeatable approach to study technological changes (Suominen, 2013).

The most frequent methods applied to quantify technological change can be divided into three groups: 1. Those related to changes in the characteristics (e.g. cost/performance/impact) of technology-related entities (e.g. Moore (1998)); 2. Those related to the volume of technology-related entities (such as terms, categories, term-pairs) within a collection of digital records (e.g. Dernis et al. (2015)); and 3. Those related to structural changes of networks composed by technology-related entities within a collection of digital records (e.g. Alshamsi et al. (2018)). These three groups are aligned with the categories of impact, activity and collaboration described by Moed et al. (1995) for tracking technological- and research performance.

Recently, scholars began to favor a view that acknowledges complexity using the third group of methods. These methods allow for a better understanding of the role and position of a particular technology within a given domain and enable analysis of mutual relations with other competing or supplementary technologies. Building on these premises, the main contribution of the work reported in this paper is twofold. First, by taking into consideration a wider set of data sources (publications, patents, project repositories) and introducing an improved network-based methodology to quantify a combinatorial view of technology, this paper allows for a more objective and integrative measurement of technological change. Second, by applying a novel measure that synthesizes technological change and evolution, which acknowledges the network structure of a certain technological domain, this paper provides new insights into the structure and dynamics of a technological domain and its combinatorial space.

The remainder of the paper is organized as follows. Section 2 reviews literature on the combinatorial view of technological change and on network-based approaches that have been used to measure technological change to-date. This is followed by a detailed description of the proposed methodology in Section 3, which encompasses the definition of document corpora, the creation of a dictionary of terms and the proposal of a novel indicator of technological change. The implementation of the methodology within bioenergy R&D as an application domain is demonstrated in Section 4. Section 5 includes a reflection on the findings from the example case of bioenergy, a discussion about the methodological components of the conducted research study and a comparison with existing approaches for measuring technological change. Finally, main findings are summarized and directions for future research are provided.

2. Technological change as a combinatorial process

2.1. The combinatorial view on technological change

As a driving force for technological progress, technological change has been widely understood as a process of combination and recombination (Fleming and Sorenson, 2001; Schumpeter, 1934), where different new and already existing technologies are integrated resulting in a technological novelty (Strumsky et al., 2012). As such, technological change is typically manifested through the introduction of new technological functionalities into a set of existing technologies (Youn et al., 2015).

Technological change emanates from recombining and synthesizing components in a novel manner (Carnabuci and Bruggeman, 2009; Fleming and Sorenson, 2001) or for a new application (Henderson and Clark, 1990; Yayavaram and Ahuja, 2008). Such combinations are considered as principal sources of technology development and progress that dominate innovation activities (Youn et al., 2015). Therefore, within the combinatorial view of technological change, new and existing technologies are considered as building blocks for other new technologies (Arthur and Polak, 2006; Fleming and Sorenson, 2001; van den Oord and van Witteloostuijn, 2018). Combinatorial processes have traditionally been described based on two criteria: 1) the familiarity of technological components being used and 2) the novelty of combinations being developed (Arts and Veugelers, 2015).

Literature on the combinatorial view of the technological progress is often based on individual case-based research (Lee et al., 2011; Nosella et al., 2008), lacking a standardized quantitative characterization and appropriate description of the underlying technological building blocks (Youn et al., 2015). The quantification of these building blocks (different technological elements) provides a more systematic approach to measuring technological changes, potentially allowing for the identification and monitoring of technology patterns and associated influencing factors (Venugopalan and Rai, 2015). Several studies have measured and employed a combinatorial search process based on insights obtained from patent and publications. Using patent data, Fleming and Waguespack (2007) explored how social interaction structures, like brokerage and cohesion, influence the creation and usage of novel combinations. With a different focus on variables including individual expertise and motivation, studies such as Arts and Veugelers (2018) and Arts and Fleming (2019) examined how such variables affect the novelty and value of inventive outputs. Although different in focus to the work reported here, these studies employ quantification strategies that can pave the way for various combinatorial process applications.

Quantitatively to describe the combinatorial process, it is necessary to establish a firm and consistent base that allows analysis of the underlying dynamics of technological changes. For that reason, technologies need to be discretized in order to map, analyze and depict a technological domain. In that way, combination traits such as the frequency and quantity of technology or component combinations may be used to recognize and demonstrate technological changes. Although Strumsky et al. (2012) expressed their empirical concerns throughout the process of discretizing and identifying technologies, there have been several attempts to develop various technological classifications that have been used for the examination of technological evolution. These classifications can be used for describing technological progress throughout a certain period and potentially forecast future trends or streams within a given application domain (Youn et al., 2015).

One of the most common classification approaches used to discretize technologies are technology codes of patents, such as International Patent Classification (IPC), the U.S. Patent Classification, or European codes (EPC), which are a way to describe a technological space. However, scholars (e.g. Venugopalan and Rai (2015)) have reported the limitations of using patent codes in connection with classifications being assigned by different patent examiners potentially

leading to inappropriate and inaccurate code assignment. In addition, patent codes do not offer enough insights about specific technological features relevant for analysis on a more detailed level. For that reason, some scholars such as Moro et al. (2018) include bibliometric analyses on project repositories and publication databases to identify and classify technologies and to tailor such analyses to the specific needs of their studies. In general, these various classification strategies result in different levels of granularity, relevance and validity of predefined categories. As a result, applicability and performance of predefined categories are often limited. Moreover, it has been reported that fixed categories or attributes that are specific to a given document type hinder the combination of different data sources (Kostoff et al., 2001) and lead to biased results (Järvenpää et al., 2011; Suominen and Seppänen, 2014). For this reason, recent studies such as Arts et al. (2018) or Arts and Veugelers (2018) adopt natural language processing techniques and text mining techniques to discretize patent technologies and as such provide first steps towards the quantification of technological changes. What is missing, however, is a quantification of technological changes that considers multiple data sources using a combinatorial view, for example, network-based approaches. Network-based approaches are varied and determine the type and characteristics of indicators of technological change that can be derived.

2.2. Overview of network-based approaches for measuring technological changes

The focus of this paper is on methods that capture structural changes of networks; networks that are composed of technology-related entities that characterize and contextualize technologies. In the case of bioenergy R&D (see Section 4 of this paper), technology-related entities would, for example, include “pyrolysis”, “catalysis”, “algae”, “microwave”, “briquette”, “biomass” within a collection of records. While, for example, algae refers to an organism, algae plus microwave refers to a technological process by which microwaves are used to process algae. Such combinations can be captured through network-based approaches. To distinguish discrete technologies, prior research has studied structural changes using various strategies such as co-occurrence of categories (Yoon and Park, 2004), references/citations (Chang et al., 2009) and keywords (Dernis et al., 2015). Such network-based approaches allow for a more encompassing and comprehensive overview of a target technological field (Choi and Hwang, 2014) and provide insights into relationships between categories or technologies analyzed (Yoon and Park, 2004). As such, network-based approaches allow for a better understanding of technological changes in an industry or technology domain (Lee et al., 2015).

Network-based approaches vary in their discretization strategies and can be divided into three groups: 1) those using predefined categories, 2) those using explicit links (references) between digital documents, and 3) those using keywords (terms) extracted from the documents.

Within the first group, we find those works that leverage both explicitly encoded categories or classes of digital records and relations between those records. Examples include technology codes in patents (e.g. Lee et al. (2015)), product classifications in countries import/export databases (e.g. Hausmann and Hidalgo (2011) and Tacchella et al. (2012)), and bibliometric studies using scientific classifications (Liu et al., 2014). These studies have allowed mapping changes in the network structure of predefined classifications within a data source of interest, relying on the granularity and comprehensiveness of the classification scheme used in the selected data source. Unfortunately, as the classification schemes are not shared across multiple data sources, these approaches are harder to scale and replicate to cover additional document sources that do not follow the same classification system.

The second group includes approaches where the primary focus is the analysis of network changes using explicit links between digital documents. These links are often interpreted as reference/citation information of digital documents. Networks built this way mostly consider digital documents as network nodes and links as citations or references of a given digital document. These citations imply the existence of knowledge diffusion between digital documents (Duguet and MacGarvie, 2005; Park and Magee, 2017), and can serve to monitor and visualize technological knowledge evolution on individual, firm, industry and national level (Kim and Magee, 2017; Sternitzke et al., 2008). Through citation-based network studies, insights have been obtained about patent and technology relationships within specific technology fields and their impact (Kim and Magee, 2017). For this reason, several attempts have been made to analyze technological development using citation networks. For example, Kajikawa et al. (2008) and Kajikawa and Takeda (2008) on several occasions have used journal citation networks to detect emerging technologies and to analyze structural network changes within sustainability technology domains such as of biofuel, solar cells etc. Utilizing patent citation networks Choe et al. (2013) explored technological knowledge flows between countries, institutions, and technologies within the domain of organic photovoltaic cells. Also, Érdi et al. (2013) conducted a study of a patent citation network to identify new technology recombination and measured temporal changes of the structure of the patent clusters. By analyzing pairwise combinations of references in scientific articles and counting frequency of co-citation pairs, Uzzi et al. (2013) explored the impact of conventional and unconventional combinations of prior work. Although these types of studies have provided a valuable contribution to the understanding of different technology domains, inherent limitations of citation-based approaches still remain (Yoon and Park, 2004). For example, these citation-based measures are often restricted to the analysis of the technology's later use without considering different combinations and configurations in which it may be included, thereby limiting a number of technological change aspects that can be monitored.

The third group of approaches is comprised of those that analyze the structural changes of networks using technology-related keywords or terms. These approaches provide several advantages, going beyond approaches based on fixed classifications or citation-networks. The most important advantage is that they allow for a more inclusive and granular analysis using actual terms used within the documents (keywords) instead of a predefined set of fixed categories or citations (Yoon and Park, 2004). An additional advantage of the term-based methods is that since they do not rely on fixed categories, they enable analysis of a combination of heterogeneous data sources without a prior common categorization system or formal citation structure. These terms are then commonly explored by various techniques such as co-word and co-occurrence analyses (Joung and Kim, 2017). Such analyses are in turn often coupled with other complementary approaches that either improve their input (Yoon et al., 2011) or output (Chang et al., 2010). The generated matrix of keywords or terms can be represented as a network graph and as such, various network metrics such as density and centrality together with procedures like clustering can be utilized to analyze a given technology domain. Many different applications of keyword-based approaches can be found in the technology monitoring and measurement literature. For example, Engelsman and van Raan (1994) used co-word and co-classification maps to represent relations between and within different technological fields. Chang et al. (2010) emphasized the role of visual representation of patent keyword-based network analysis to explore the research field of carbon nanotube field emission display. In order to improve co-word analysis, Yoon et al. (2011) coupled it with the property-function based approach and social network analysis. Recently, several examples can be found where scholars suggest using SAO-based (subject-action-object) and TF-IDF (term frequency-inverse document frequency) approaches mostly focused on the improvement of the identification of emerging technologies, e.g. Joung and Kim (2017).

Despite the promise that the previously introduced third group of

approaches represents, to this date there is no method specifically designed to calculate an indicator of technological change that captures in one measure overall technological change over time based on a combinatorial view of technology. In what follows, we offer and test such a method.

3. Methods and data

Consistent with a combinatorial view on technological change (Youn et al., 2015), the proposed method in this paper uses the occurrence and co-occurrence of terms within a corpus of documents (Feldman and Sanger, 2007) to describe different combinatorial configurations within a document corpus of R&D-related records. In our approach, terms are text strings of one or more words, also called n-grams (Dale et al., 2000) that represent technology-relevant entities. Within the domain bioenergy chosen here, such technology-relevant entities are production inputs (e.g. barley straw), processing technologies (e.g. pyrolysis) and outputs (e.g. biogas). More specifically, we use changes over time in the occurrences and co-occurrences of such terms as a proxy for technological changes. In this approach, the occurrences and co-occurrences of selected terms within documents are used to build adjacency matrices that store the weighted combinations of those terms and term-pairs for each time period. Such matrices serve 1) as a description of the combinations of terms that have been explored in a given period of time and 2) to calculate configurational changes in the matrix from one period to the next. An overview of the key steps in the process is provided in Fig. 1 below.

3.1. Data sources and creation of the document corpus

The document corpus is built of documents extracted from sources that record the results of research and development (R&D) activities over time. Example sources for these document records include patents, scientific publications, industry databases of R&D pilots and facilities, as well as descriptions of research and innovation projects (e.g. R&D projects funded by the European Commission). In the analyses here, such a range of document sources is integrated with the purpose of including a more diverse set of sources which allows for a more inclusive representation of the state of the art in a given technological area (Guthrie et al., 2013; National Research Council, 2014; O'Keeffe and McCarthy, 2012) and helps to embrace different knowledge types related to these documents records. In fact, the proposed method is not limited to a predefined set of sources, as it can include as a document source any date-stamped collection of text-based records that is considered relevant for a given technological field or application domain. This combination of different and complementary types of records provides an integrated representation of technological changes within a given domain.

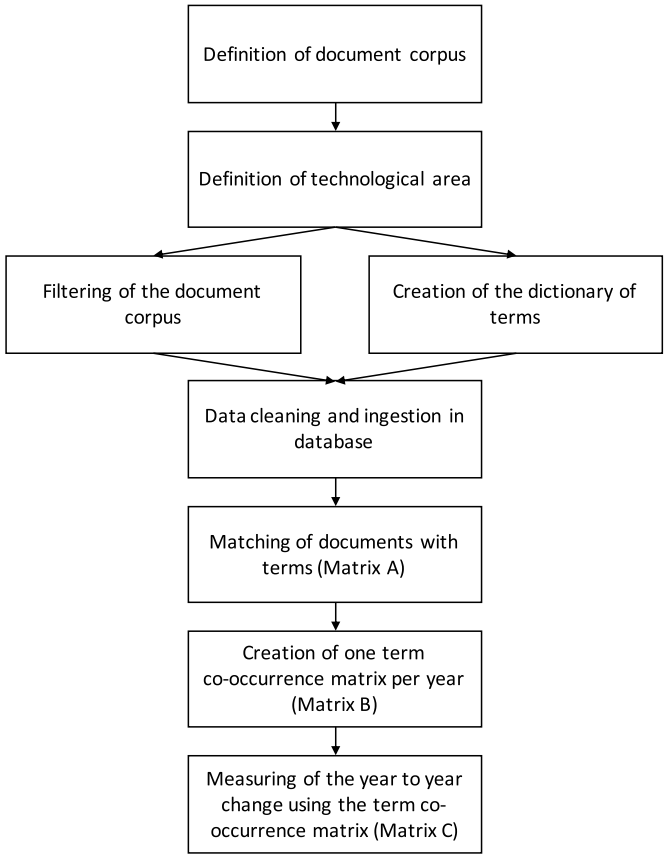

Fig. 1. Overview of the key steps in the proposed method to quantify technological change.

The minimum criteria for inclusion of a document source is the existence of an abstract, date stamps containing at least the publication year, and means to ensure the relevance of the documents included in the data source (e.g. peer-review in the case of scientific articles and application processes for patents and research projects). Additional data about each document record can be used to filter or interpret results, including information on geographical location, organizational affiliation or authorship.

Depending on the application domain or technological area of interest, filters are used to narrow down the records extracted from each source. The objective of filtering is to facilitate the analysis and interpretation of results (Wachsmuth, 2015), restricting the combinatorial space of possibilities to the technological areas most relevant to understand the technology domain of interest. Technological areas of interest can be defined at a very broad level, e.g. “Energy” or narrowed down to increasingly specific areas of interest e.g. “Renewable Energy”, “Solar energy”, or “Solar Photovoltaic Energy”. To operationalize the filter, a set of keywords is defined to extract a broad set of document records from each source connected to the technological area of interest. The search strings used for filtering are selected following a strategy that seeks high recall and precision, while recognizing the inherent trade-off between these attributes (Buckland and Gey, 1994). The selected strategy is to divide the search process into two steps, a first step that seeks to maximize recall (maximum coverage) and a second step focused on maximizing precision (removing false positives) (Buckland and Gey, 1994; Wachsmuth, 2015).

3.2. Creation of the dictionary of terms

There are multiple alternatives to generate a dictionary of terms to build the term co-occurrence matrices (Alghamdi and Alfalqi, 2015; Cook and Jensen, 2019; Feldman and Sanger, 2007; Noh et al., 2015). These alternatives can be divided into three broad groups: 1) Methods that generate a reference dictionary by tapping directly into knowledge from domain experts. These methods create an ad-hoc set of terms for the required analyses. In that way, a rich and highly contextualized dictionary is created. Yet, one that is hard to scale and hard to update. 2) Methods that extract terms directly from the corpus, employing information retrieval techniques, such as term frequency–inverse document frequency (TF-IDF) that identifies keywords and ranks them based on their relative frequency. These methods are highly scalable and easy to update, but they are unable to capture attributes specific to the technological domain and classify the terms accordingly. 3) Hybrid methods, like the ones applied in this paper, that mine pre-existent expert knowledge from large open datasets, including structured dictionaries, ontologies and taxonomies to extract terms associated with the inputs, processing technologies and outputs. A key characteristic of these methods is that they do not build the dictionary using the main document corpus that will be analyzed to quantify technological

change. Instead, they use a separate set of data sources specifically chosen for the purpose of building a reference dictionary. Such methods combine the advantages of gathering a temporal structured knowledge from domain experts with the scalability of computational text mining approaches. In these hybrid methods, the key criteria to determine the inclusion or exclusion of a given term is their existence in the curated set of data sources. In this way, the list of terms is the result of the aggregated expert knowledge distributed in authoritative repositories as opposed to being based on a discretionary decision.

Generating a list of terms for a dictionary is sometimes referred to as named-entity extraction and recognition (Nadeau, 2007). Entities can be extracted for example from, Wikidata, DBpedia, open classifications such as the European "Reference and Management of Nomenclatures" system, and from patent claims and patent categories. This method provides access to structured lists of terms covering, among other things, all known chemical elements, molecules, organisms, products and commodities. Such extensive sources of organized entities allow for comprehensive coverage of the potential combinatorial space and have the advantage of being a list that is constantly updated by a community of supporters to reflect new discoveries and inventions allowing to maintain the dictionary up-to-date. The extraction of terms to build the dictionary from these sources can be challenging because many terms have multiple synonyms and variations. To deal with these challenges, term normalization techniques are applied to merge conceptually equivalent term duplicates (e.g. Cho et al. (2017)) and a final step of manual verification of the results by the researchers provides a qualitative validation of the dictionary.

3.3. Creation of term co-occurrence adjacency matrices

To build the adjacency matrices that store the co-occurrences of terms on each time period, first, the occurrences of each of the terms in each of the documents in the corpus per time period need to be collected. The occurrences of terms within documents can be stored in what is known as a bi-adjacency matrix. Here, this matrix is a representation of a bipartite network, where the first type of nodes are the terms in the created dictionary and the second type of nodes are each of the documents within the corpus. In this bipartite network, an edge exists between a given term and a given document if the term is found at least once within the document. In matrix representation, this takes the form of an unweighted rectangular Matrix A of size d x t, where the rows in the matrix list all documents (d) and the columns lists all terms (t). Within this Matrix A each cell represents the existence or absence of a match term-document.

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1t} \\ \vdots & \cdots & \vdots \\ a_{d1} & \cdots & a_{dt} \end{bmatrix} \quad (1)$$

In Matrix A, the sum of the values in a column provides the number of matches for that term in the document corpus. The column projection of Matrix A is calculated as A^TA (Everett and Borgatti, 2013) and allows to transform the original matrix from a two-mode network representation "document-term" to a one-mode network representation that is contained on a weighted and undirected (symmetric) adjacency term-term matrix. In this new weighted Matrix B, each cell counts the number of documents in which a given pair of terms co-occur. This procedure is run for the documents that occur in each year and as a result creating one weighted adjacency matrix per period (B_y).

To account for the different amount of document records that exist per year, the matrix (B_y) is normalized by the total number of the document records per year. The normalization approach divides each entry in the matrix (B_y) by the total number of documents in the year "y".

3.4. Calculation of year-to-year changes in the normalized matrix of term co-occurrences

In order to measure year-to-year changes in the normalized term co-occurrence matrix, an algorithm is needed that summarizes the relationship between pairs of matrices representing different years. More generically, this is equivalent to quantifying the relation between two pairs of high-dimensional datasets in matrix form. Among a diverse set of matrix correlation algorithms (for a review see (Ramsay et al., 1984)), the R Vector (RV) coefficient (Robert and Escoufier, 1976) is often considered as the most appropriate metric to quantify the similarity between squared symmetric matrices (Abdi, 2007; Josse et al., 2008; Smilde et al., 2009).

Mathematically, the RV coefficient is a measure that takes values between 0 and 1, where 0 indicates no similarity between two matrices and 1 indicates that the matrices are structurally equivalent. For Matrix B_i (B in time i) and Matrix B_j (B in time j), their RV coefficient is:

$$RV(\mathbf{B}_i, \mathbf{B}_j) = \frac{trace(\mathbf{B}_i\mathbf{B}_{i'}\mathbf{B}_j\mathbf{B}_{j'})}{\sqrt{(trace[(\mathbf{B}_i\mathbf{B}_{i'})^2]trace[(\mathbf{B}_j\mathbf{B}_{j'})^2])}} \quad (2)$$

In the context of a combinatorial perspective on technological change, the RV coefficient measures the extent of the configurational change from one year to the next, in the form of a similarity measure which fits the requirement of examining not only the evolution of individual terms or categories but also the way in which they are combined.

An overview of the three groups of matrices used in the method for quantifying technological change proposed here is provided in Fig. 2.

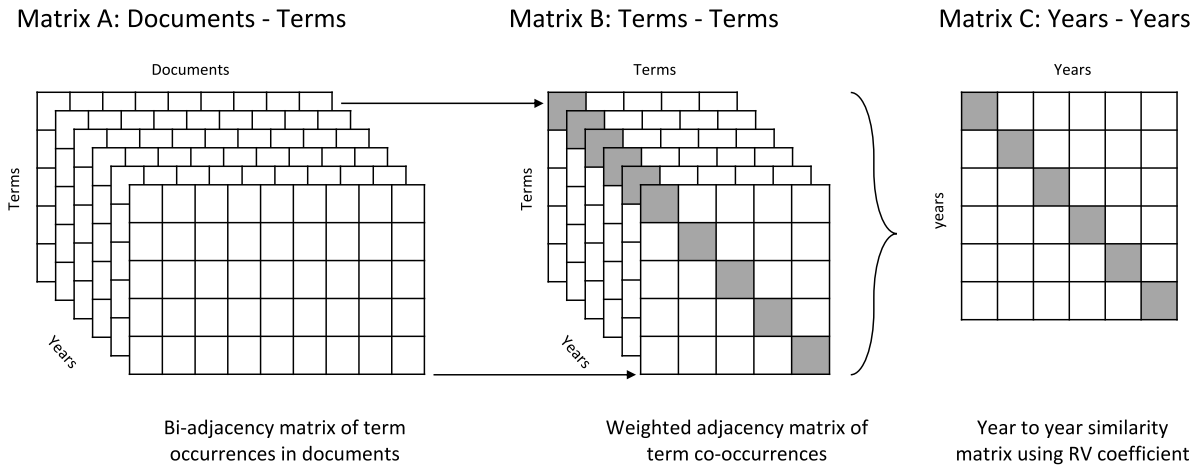

Fig. 2. Overview of the three groups of matrices used in the proposed method for quantifying technological change.

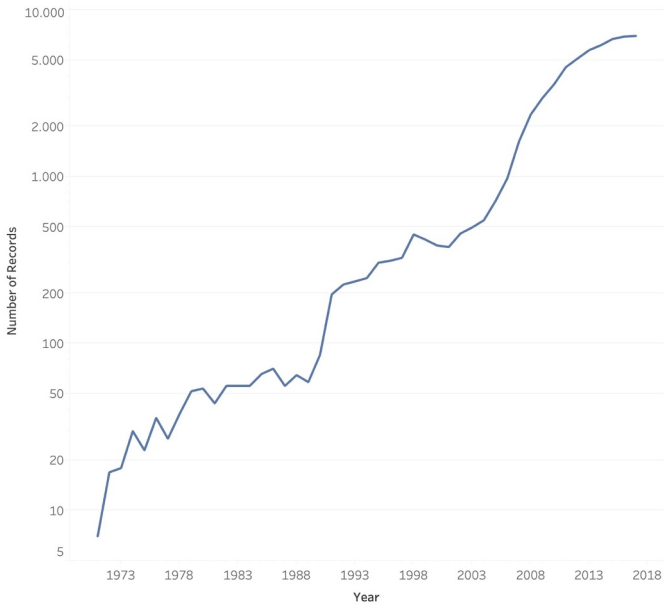

Fig. 3. Volume of R&D documents within the bioenergy field collected for this research (logarithmic scale).

4. Results: technological change in bioenergy R&D worldwide

To demonstrate the method for quantifying technological change proposed here, we applied it to measure technological changes in the context of research and development (R&D) in the field of bioenergy solutions. Research and development of bioenergy solutions is an active technological field, with R&D-related records starting from the mid-seventies (Clarivate, 2018a; Gupta et al., 2014), and a field that has in the past two decades experienced a significant increase in the volume of R&D document records (see Fig. 3, logarithmic scale).

Bioenergy R&D is concerned with the generation of renewable energy using materials derived from biological sources such as biomass (Gupta et al., 2014). One of the main areas within this technological space is the production of biofuels (Pandey et al., 2011). During the lifetime of this field, there have been several technological changes related to aspects such as the production inputs used (the feedstocks) and the processing technologies utilized (Ferreira et al., 2013). The main drivers for these changes are connected to the social, environmental and economic sustainability of biofuels. These drivers have translated into pressures to increase the speed and volume of the production of greener and cheaper biofuels that can become viable alternatives to fossil fuels (Pandey et al., 2011). Retrospectively, such changes have been characterized in terms of what is now known as four biofuel generations (Aro, 2016).

Although not all researchers agree on the descriptions of these four generations, and these descriptions have changed over time, in general terms, first-generation biofuels, also known as conventional biofuels, are characterized by being produced using food crops and processes such as fermentation. Second-generation biofuels are produced using non-food crops as well as agricultural waste and are often processed using thermochemical and biochemical approaches. Searching scientific publications (Scopus) and the indexed corpus of Google Books, the first formal reference to “second generation biofuel” appears in the year 2006, introduced to describe the notion of a generational distinction in the technological landscape of bioenergy R&D. Third-generation biofuels are mostly algae-based and are often processed using oil-extraction methods (the first formal reference to them is in 2008). Finally, fourth-generation biofuels represent a wide-range of approaches currently in development and the first formal reference to them appears in 2010. This fourth generation of biofuels, in conjunction with sustainability benefits associated with second- and third-generation biofuels, is characterized by technological advancements such as the integration of CO_2 capture and storage processes, the use of synthetic biology (e.g. designer photosynthetic microorganisms) and the use of cyanobacteria. Fourth-generation biofuels seek to go beyond being carbon neutral to effectively providing net carbon-negative solutions.

As the above description suggests, from a combinatorial perspective the technological landscape of bioenergy R&D is a rich space of study, where a large number of combinatorial possibilities have been explored connecting multiple sets of inputs, processing technologies and outputs over time. This, in addition to the societal relevance of bioenergy R&D, has made this field a good test ground for the method of quantifying technological change proposed here.

An important challenge within this technological field relates to the difficulty of mapping the rate of technological changes (Chuck, 2016; Curci and Mongeau Ospina, 2016). Given the combinatorial nature of technology (Youn et al., 2015), technological change needs to consider not only individual trends based on the volume of documents with specific terms or categories but also the relative changes in the explored combinations of key inputs, processing technologies, and outputs.

4.1. Data sources and creation of the document corpus

Following the method proposed here, to achieve a broad coverage of relevant research and development document sources, we have text-mined a diverse set of historical records of technology-related R&D activity. These records include patents, scientific publications, official EU project descriptions (the CORDIS database), as well as databases that include Bioenergy2020+, ETIP Bioenergy, Biofuel Digest and Genscape which contain descriptions about biofuel-specific projects and worldwide biofuel facilities.

To focus on those results that are most clearly associated to bioenergy or biofuels, we have used the following text string as filter for all the data-sources: [“biofuel* OR bio-fuel* OR "bio fuel*" OR bioenerg* OR "bio energ*" OR bio-energ*”]. This text string was selected based on the amount and quality of the results it provided, which after manual examination showed a good balance between recall and precision. After applying the filter, the number of documents that are per data source included in the analyses is:

- Scientific publications (Clarivate, 2018a): 58.239 documents
- Patents (Clarivate, 2018b): 6.570 documents
- Official EU project descriptions (European Commission, 2018): 692 documents
- Biofuel facilities and projects (Biofuel Digest, 2018; COMET Centre, 2018; European Technology and Innovation Platform Bioenergy, 2018; Genscape, 2018): 1.647 documents

Although the majority of the records come from scientific publications, the additional coverage in terms of patents, projects, and biofuel facilities, allows the capture of terms and term-pairs that are widely used outside academic circles and to a lesser extent represented in scientific publications. In addition, since the year-to-year analyses are not affected by absolute volume but rather by the overall configuration of the term matrix, yearly variations in volume between document sources are less problematic.

4.2. Creation of the dictionary of terms

To build a dictionary of relevant terms, we followed the hybrid method described in Section 3.2, mining comprehensive open taxonomies in English that have been developed in the bioenergy field with the explicit objective of cataloguing all known feedstocks, processing technologies and biofuel outputs. The taxonomies we mined are Reegle's “Renewable Energy Glossary” (REEEP, 2018), “NREL - The Biofuels Atlas” (NREL, 2018), the “Advanced Biofuels & Biobased materials

Project Database" (Biofuel Digest, 2018), and the "Bioenergy Feedstock Library Idaho National Laboratory" (Idaho National Laboratory, 2018). Once terms using name entity extraction were elicited (Nadeau, 2007) and duplicates removed, a list of 208 entities was obtained (mostly made of one or two terms each). For example, entities classified as inputs (feedstocks) include strings such as "corn", "algae" and "bagasse". Entities within processing technologies include strings such as "transesterification", "pyrolysis" and "enzymatic hydrolysis". Entities classified as outputs include strings such as "biobutanol", "biodiesel" and "methanol". Although new terms might be introduced over time, the analysis can be replicated and kept up to date by adding terms when a new one is listed in one of the mined sources. This is possible as increasing the size of the co-occurrence matrix to include new terms does not affect the analysis for the previous years when the term has not yet appeared in that year's corpus.

4.3. Creation of term co-occurrence adjacency matrices

To obtain the term-term co-occurrence Matrix B, we first build the bi-adjacency matrix of documents-terms per year. In this case, the bi-adjacency matrix contains 208 entity terms (t) and 67.148 documents (d). Within this matrix (Matrix A), each cell represents the existence or absence of a match term-document.

The projection of Matrix A from documents-terms to terms-terms (Matrix B) provides one weighted adjacency matrix of size 208 by 208 that contains 21.528 unique term-pairs for each year. Each term-pair cell in the matrix stores the number of times a set of two terms co-occurred in the same document in a given year. In addition, we store within the diagonal of Matrix B the number of times each term appears on a given year. The combination of term-pairs co-occurrences and the matrix diagonal that stores the volume of individual terms on a given year, integrates information of both enacted combinatorial possibilities and individual term usage. For example, Fig. 4 shows Matrix B for one year with the terms grouped into feedstocks, processing technologies and outputs. Matrix B provides a consolidated view of the relative usage of each term and all the connections between terms as well as their volume.

To test if there were systematic differences in the results between the two main datasets, scientific publications and patents, a parallel analysis was run creating an equivalent adjacency matrix for each of these datasets.

4.4. Measuring year-to-year changes in the matrix of term co-occurrences

The combinatorial space of possibilities for any given year in the bioenergy R&D example case here holds a total of 21.528 term-term pairs. One approach to quantify the degree of technological change is to measure the configurational similarity of each pair of year matrices (Matrix C). The higher the configurational similarity between one year and another, the lower the overall technological change between those years. Hence, year-to-year changes in the method proposed here account both for the changes in the relative amount of individual terms between one year and another and for changes in the configuration of the matrix that describes the combinatorial space actually explored between one year and another.

The RV coefficient is used to measure the configuration similarity

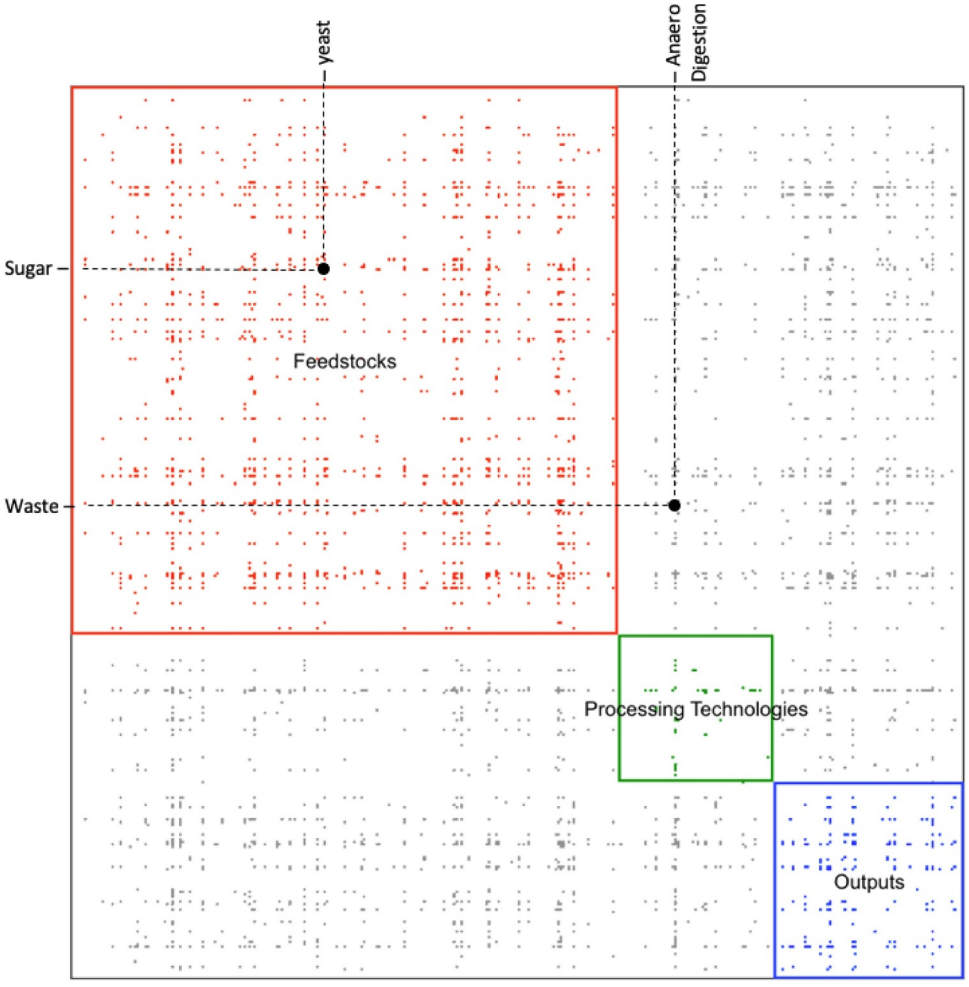

Fig. 4. Term-term Matrix B. The matrix shows the relative usage of each term and all the connections between terms as well as their volume.

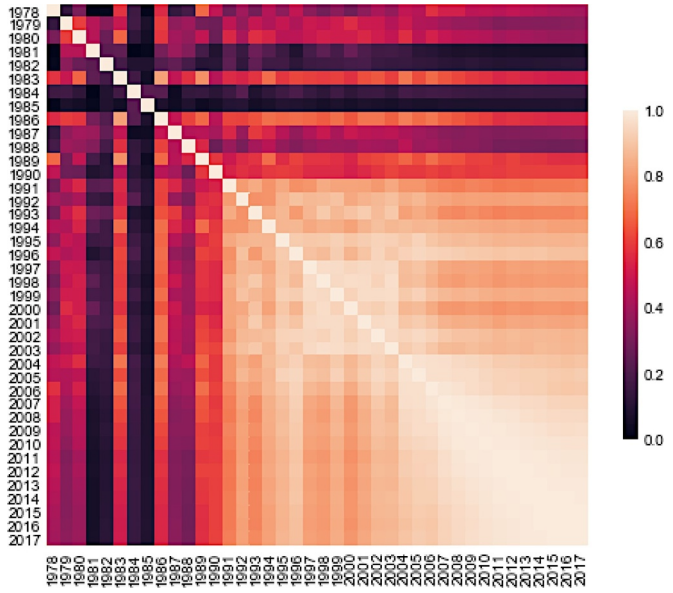

Fig. 5. Matrix C, Years-years matrix storing RV coefficients that measure the configuration similarity between any two-year pairs. The lower the value, the greater the change.

between any two given matrices B for all year-pairs. These similarity measures are stored in a weighted years-years adjacency Matrix C. The chronologically ordered years-years Matrix C for the period 1978–2017 is presented in Fig. 5. The top-left to bottom-right diagonal in the matrix stores the similarity of a year with itself, hence that value is always one. Off-diagonal cells range from zero (minimum similarity value between two given years) and one (complete equivalence).

The chronological results for the year-to-year technological similarity measures are presented in Fig. 6 below. The results can be understood as a technological change indicator where a low RV coefficient means higher technological change and high coefficient means lower technological change.

Despite the different sizes of the patent and publication databases, parallel analyses of patents and scientific publications in isolation showed no statistically significant difference to the year-to-year technological similarity measures observed for the aggregated dataset.

4.5. Interpretation of the indicator of technological change applied to the bioenergy R&D case

Results shown in the matrix that stores the RV coefficients for all year pairs (Fig. 5) point to the following four main findings:

1) As intuition would suggest, in general, when the time between two years increases, their similarity decreases. This is an indication that our method is able to capture the theoretically expected macro-behavior of technological change, which predicts that over time the accumulation of year-to-year changes (both incremental and radical) should lead to an increase in the accumulation of technological changes over time (Parayil, 1993). For example, an examination of the RV coefficient curve for the year 2017 in comparison to all other years, see Fig. 7, shows that with few exceptions, the farther we move from 2017, the lower the RV coefficient becomes.
2) As previously shown in Figs. 6 and 7, year-to-year similarity measures are relatively low from one year to the next in earlier periods and are higher in later periods. This indicates that in earlier periods, i.e. 1978–1990, year-to-year configurational changes in the matrix that stores the combinatorial possibilities are of larger magnitude and more frequent. As time passes, i.e. 1990 onwards, year-to-year changes become smaller, which can be interpreted as a sign that overall bioenergy R&D is settling into more stable technological configurations. This is consistent with previous research on technological breakthroughs by means of niche accumulation and co-evolution of technologies (Geels, 2005), which suggest that early on in a technological area we should observe the emergence of several niches that take time before breaking out to whole systems level and becoming mainstream.
3) Year-to-year similarity measures are generally much lower and more rapidly changing in earlier years than in later years. We see this as a marked difference in RV coefficients in the period pre-1991 (average RV coefficient of 0,31) and post year 1991 (average RV coefficient of 0,88). One element that influences this behavior is the lower volume of documents in earlier years, which means that the chances available to explore all potential combinatorial possibilities are reduced. However, this is not an artefact of the sampling method or the data-sources available. Instead, it shows that as the volume of R&D activity increases exploration of a wider set of combinatorial possibilities is possible, which in turn translates into more stable year-to-year similarity measures. A second element that influences this

Fig. 6. Chronological results for the year-to-year technological similarity measures.

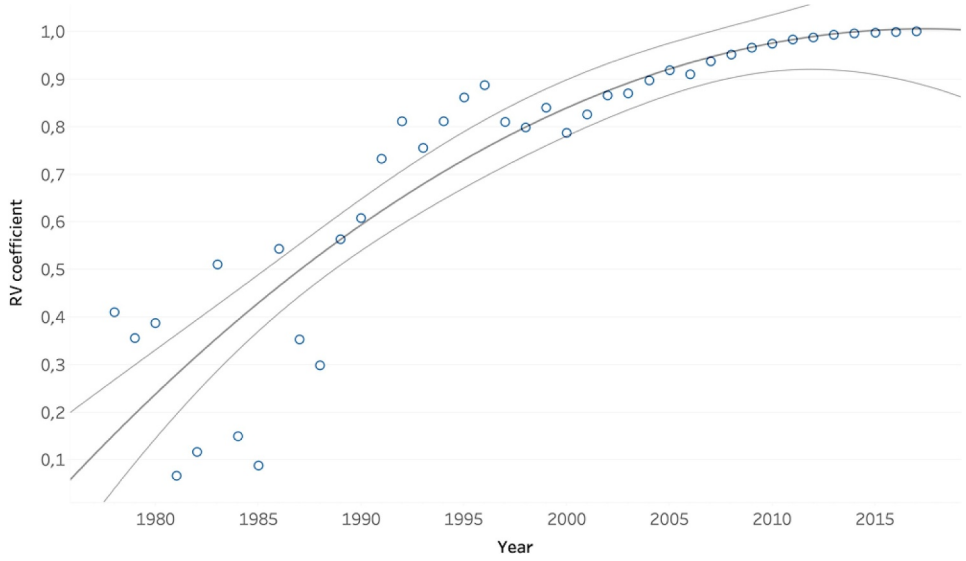

Fig. 7. RV coefficient time series for year 2017.

behavior is the more experimental and uncertain nature of early explorations of a new technological field. This means that shifts in early periods will translate into larger configurational changes when compared to later periods. One driver for this is the relatively small volume of overall R&D activity earlier on, which makes each change a more significant percentage of total R&D activity. In contrast, in later periods, the larger volume of R&D activity can be distributed into multiple parallel areas of research.

4) A clustering analysis on the full matrix including all year-pairs was conducted as an additional step to complement the separate analysis of the year-to-year similarity measures. The results of this cluster analysis permit identification of blocks/groups of years with higher and more stable year-to-year similarity that are interrupted by changes that break the high similarity found within the block. After such change, a new block emerges over time. Using a hierarchical clustering analysis (Johnson, 1967) of Matrix C, shown in Fig. 8, we can identify the following tree structure describing the year intervals that have high similarity within interval and low similarity outside of the interval:

Results of the cluster analysis of the Matrix C allow the observation of two large year intervals; before and after 1990, which is the year that marks the largest configurational difference between any two large intervals. In the period until 1990, we observe two sub-intervals

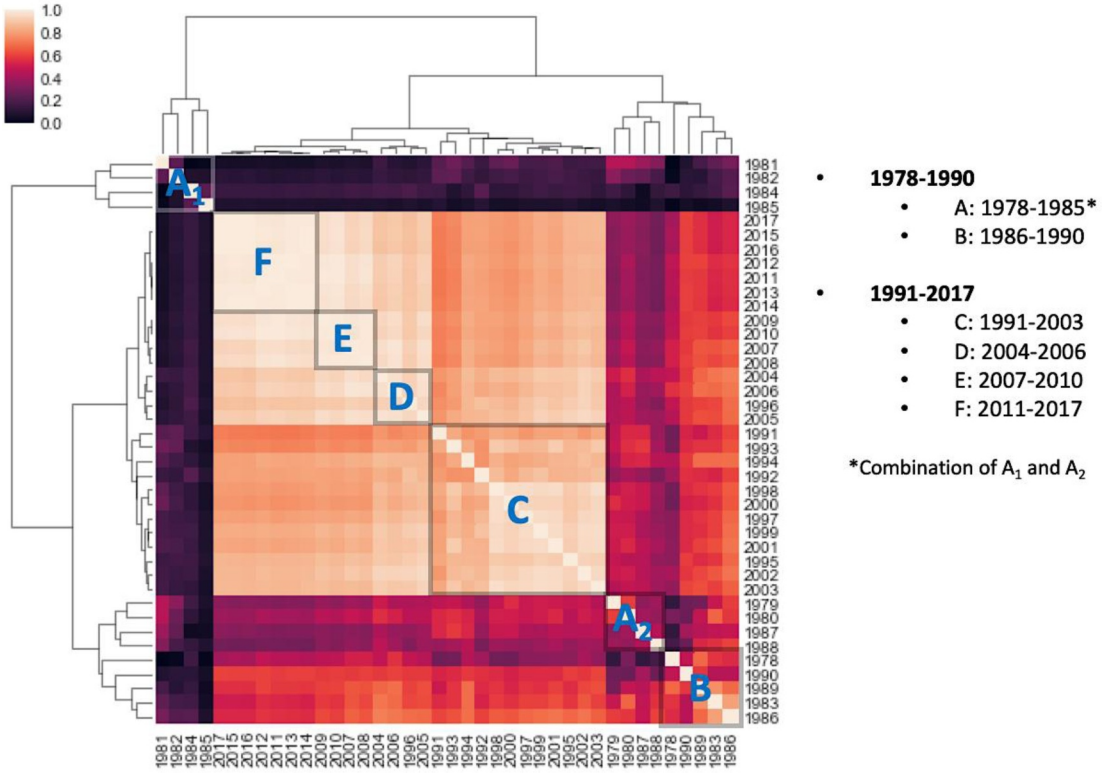

Fig. 8. Hierarchical clustering analysis of years.

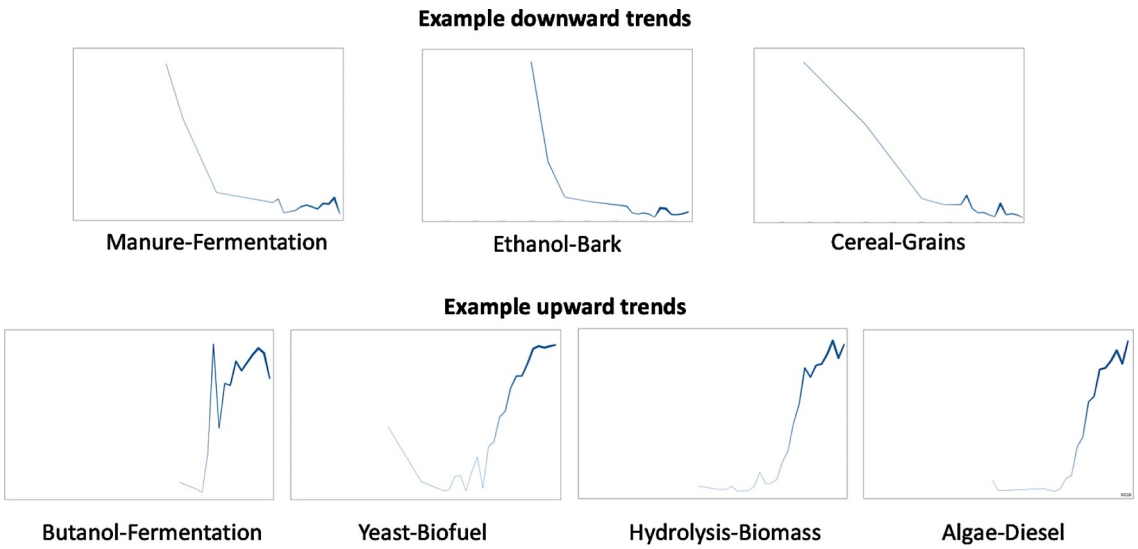

Fig. 9. Example term-pairs trends. The vertical axis (independent for each plot) shows the volume of records. The horizontal axis shows years from 1975 to 2018. The thickness of the line is a representation of the number of records.

1978–1985 and 1986–1990. In the period after 1990, we observe five sub-intervals 1991–2003, 2004–2006, 2007–2010 and 2011–2017. In addition, using principal component analysis (PCA), we found a similar structure to the one revealed by the clusters with two large groups of years, one before and one after 1990.

To evaluate the meaningfulness of the results obtained, and to provide an interpretation of the changes over time of the indicator developed here, reference points against which observed patterns can be validated are used.

A first reference point is provided by an examination of volume-normalized trends for individual terms and term-pairs, focusing on trends that show the largest increase or decrease over time. The objective is to identify patterns within those trends that help to contextualize the obtained year-to-year similarity results. As shown in Fig. 9, term-pairs such as manure-fermentation, ethanol-bark, and cereal-grains show a consistent downward trend from the nineties, which becomes a stable flat line from around 1995 onwards. In turn, term-pairs such as butanol-fermentation, yeast-biofuel, hydrolysis-biomass, and algae-biodiesel all show a sharp increase around the year 2005, and with the exception of butanol-fermentation, which exhibits a sharp increase much later, all these combinations appear for the first time between 1990 and 1995.

What these trends show are examples of technological replacement that the method developed here captures first as two large technologically distinct groups, one before and one after 1990. Before 1990 the total list of terms used is short and some of the top terms include waste, fat, cereal, molasses, and gasification. After 1990, the following highlights for each of the identified subperiods are found: 1) 1991–2003 acts as transition between the terms used in the 1980s (e.g. grain and corn) which are now in decline and a wide range of new terms such as pyrolysis, hydrolysis, and liquefaction; 2) 2004–2006, where the terms first introduced during the previous stage now experience rapid growth; 3) 2007–2010, where there is rapid growth of terms associated with third-generation biofuels (for example algae); and 4) 2011–2017, where the configuration of terms is stable and mostly focused on terms connected to second and third-generation biofuels.

A second reference point is provided by the time when each generation of biofuel was formally described for the first time. The left panel of Fig. 10 shows a plot with the number of documents mentioning each of the generations between 2006 and 2018, and the right panel shows the year-to-year similarity matrix with marks when the first published mention of each generation occurred. The figure shows that the first formally published description of the notion of a second generation of biofuels comes significantly after the first quantitative signs

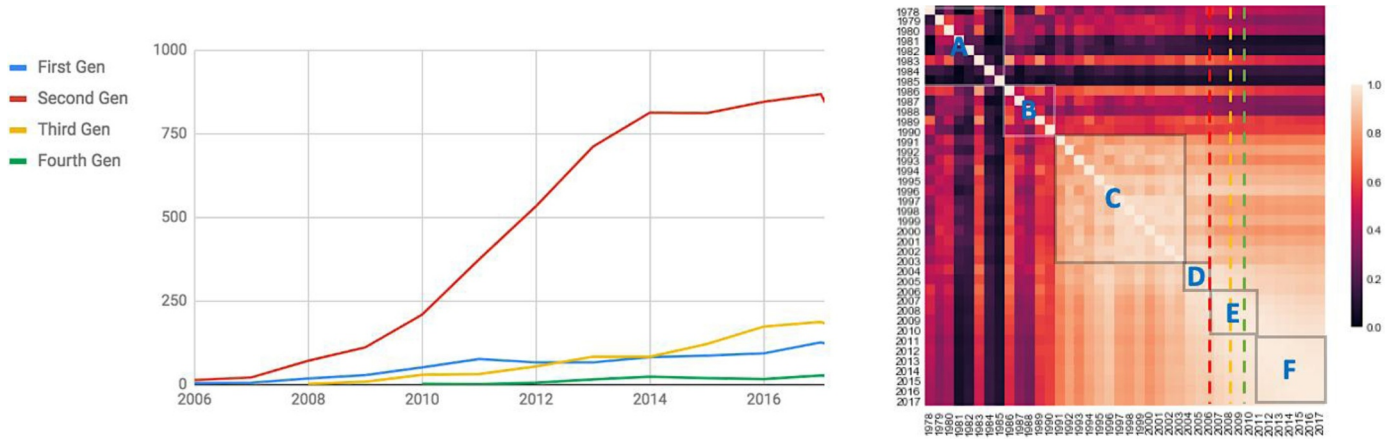

Fig. 10. Left panel, number of records containing references to each of the four biofuel generations. Right panel, year-to-year similarity matrix between 2006 and 2018, with colored vertical lines for when the first mentions to each generation occurred and the previously described clusters A to F as reference points (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

are identified with the proposed method. For example, the second generation of biofuels, formally described for the first time in 2006, is associated with terms such as “cellulosic ethanol”, but early signs can be found in publications from 1990 onwards.

Likewise, the third generation of biofuels, formally described for the first time in 2008, is often associated with algal biofuels, but early signs of activity in algal biofuels can be found before 1990. Furthermore, in 2004, two years before the first mention of a third generation of biofuels, there was already a high growth trajectory for algae-related term-pairs. We can also identify signs of high growth connected to fourth-generation biofuels, which was formally mentioned for the first time in 2010, at least two years after we observe the high growth in the usage of terms such as “capture and storage”, “synthetic biology”, and “cyanobacteria”.

The third reference point is provided by historical milestones in the development of modern biofuels, which can be divided in two waves: 1) milestones related to energy security shocks that triggered governmental initiatives seeking to replace traditional fossil fuels (economic sustainability) and 2) milestones driven by the negative effects of first-generation biofuels in food supplies (social sustainability) and by the desire to introduce carbon-neutral alternatives to fossil fuels (environmental sustainability) (see Gupta et al. (2014) and Pandey et al. (2011) for an overview of these milestones). The first wave of milestones created a surge in the development of modern first-generation biofuels. This surge is reflected in Fig. 11 in the period 1978–1990 with significant technological changes as measured by the low year-to-year similarity measures in clusters A and B. In turn, the second wave of milestones influenced the development of biofuels from non-food crops (second-generation biofuels) as well as the development of additional alternatives that minimized environmental and social externalities (third- and fourth-generation biofuels). This second wave is reflected in Fig. 11 in the period 1990–2010 where the increased number of combinatorial alternatives generated in cluster C (1991–2003) led to a growth in the year-to-year similarity measures that can be interpreted as a sign of growing maturity in the configurational possibilities explored.

5. Discussion and conclusions

From a research implication point of view, the proposed method permits the quantification of technological change by focusing on a combinatorial, temporal perspective. Although alternative metrics to quantify technological change that are “objective and reproducible” have been developed (e.g. review by Suominen (2013)), such metrics are often tailored to the needs of specific application domains without providing general guidelines for multiple domains (e.g.

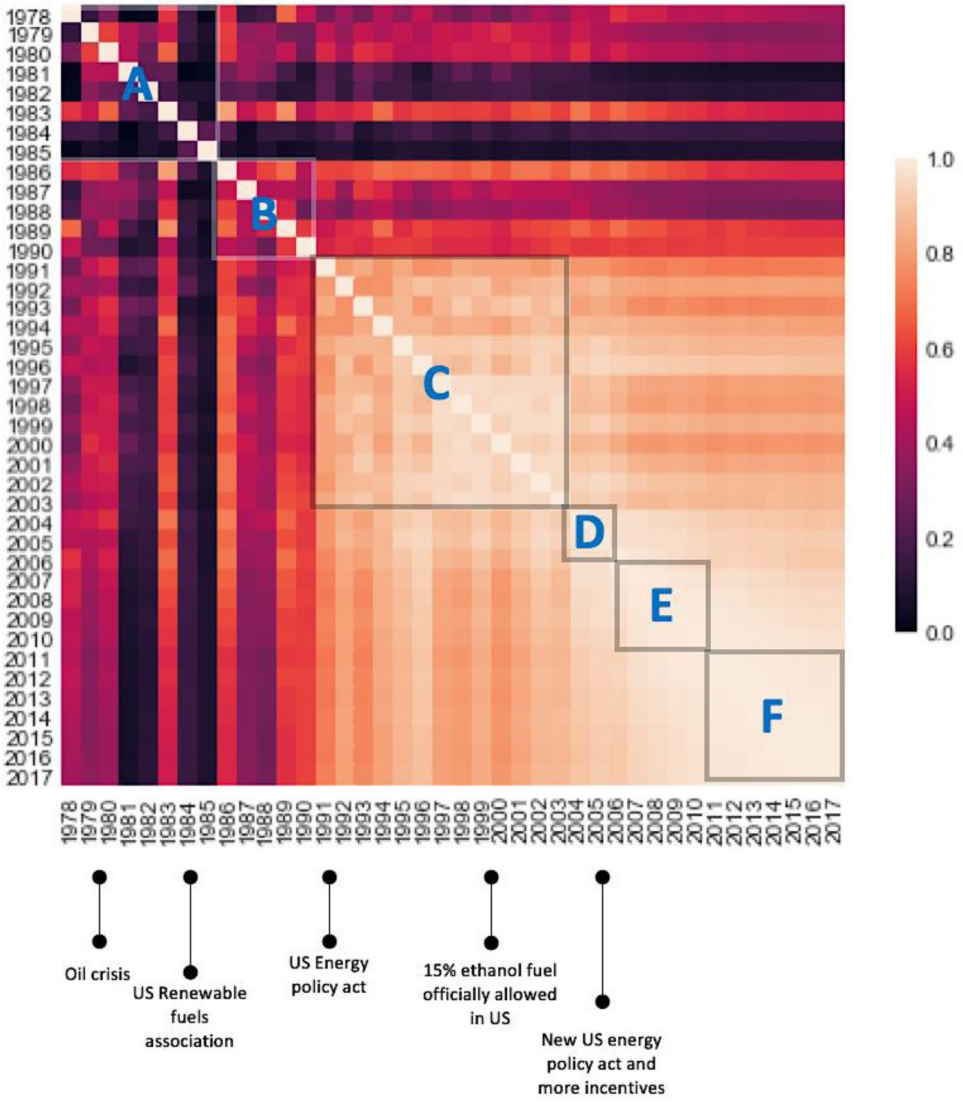

Fig. 11. Historical milestones in the development of modern biofuels.

Goldfarb (2005)). Therefore, although previous approaches offer some advantages in exploring technological change, their drawbacks may lead to a potentially biased and distorted view of technological dynamics and evolution (e.g. Funk and Owen-Smith (2017)). For example, an increase in R&D funding might result in a higher number of publications and patents without the expected technological diversification (Kook et al., 2017). In comparison, the network-based method proposed in this paper allows for investigation of the combinatorial possibilities on a macro-level and serves as a platform for complementary analyses of technological changes. This novel network-based representation and understanding of technological change also enables exploration of the temporal evolution of pre-existing configurations and subsequent additions or replacements.

The proposed combinatorial view can easily be turned into a multi-level analysis to study 'stable component configurations', 'dominant designs', 'eras of incremental change', or 'disruptive shocks' at the macro level. At the same time, this view allows the investigation of different development stages of component technologies that emerge on a lower levels of analysis (van den Oord and van Witteloostuijn, 2018). For example, using our study of the bioenergy R&D field, we provide evidence of periods of rapid combinatorial change (e.g. 1990–1991, 2003–2004) and periods of incremental combinatorial change (e.g. 2008–2009, 2012–1017). We can also observe the emergence of each of the four generations of biofuels before the notion of such generations was first introduced. Further analysis related to the identification of these various above-mentioned patterns is only one of the potential avenues for future research.

This method provides a comprehensive encapsulation of technological change on the level of the whole domain and a single technology (Funk and Owen-Smith, 2017; van den Oord, 2010). Moreover, it also offers a complementary perspective on the quantitative background of combination and recursiveness; two of the three principles proposed by Arthur (2009). Thereby, retrospective exploration of development patterns of mature and emerging technologies is made possible, in addition to an explanation from a systemic point of view of "how they came into being" (Arthur, 2009). Aligning with work done by Hekkert et al. (2007), the sequential representation of technology dynamics over time could lay the empirical and theoretical foundations for mapping the functions of individual innovation systems.

The network changes on the level of the technology domain identified through the analysis of year-to-year similarities permit introduction of the temporal dimension to the combinatorial view. In particular, insights obtained through the year-to-year analysis allow identification of patterns within different time periods and comparative analyses of given timeframes. As such, historical milestones and technological trends are indicated. Unlike analytic approaches that seek to identify specific patterns connected to technological emergence, which may be manifested in a potentially large number of individual patterns, the proposed single indicator resulting from the quantification method proposed here encompasses all identified changes from one period to the next in the form of the overall configurational structure of a given technological space. As such, it provides a complementary perspective to the studies done by Lee and Berente (2013). The method proposed here is more encompassing and adds more contextual information into the statistical analysis of a narrower technology domain. It also harvests insights into explicit links between individual technologies; links that are revealed through the analysis of network changes. Moreover, the method supports both research streams of technological change dynamics defined by Adegbesan and Ricart (2007), i.e. it includes both drivers and outcomes of technological change as represented by different reference points in the previous section.

From a methodological implication point of view, the proposed method offers three main benefits that are related to three different steps: The creation of a document corpus, the creation of a dictionary of terms, and the measurement of year-to-year changes in the matrix of term co-occurrences. Firstly, the method does not rely on data specific to only one data source and provides wider boundaries and demarcation of the technology domain. As such, it offers an exploration of alternative demarcations as indicated previously in (van den Oord and van Witteloostuijn, 2018). Aligned with previous studies such as Arts et al. (2018) and Arts and Veugelers (2018), this study employs a similar text-mining approach for the quantification of technological spaces. In that way, it addresses recent calls for extending data corpora used for this type of studies, going beyond the much-explored corpus of patents and publications (O'Keeffe and McCarthy, 2012). For example, scholars such as Abercrombie et al. (2012) and Li (2015) have highlighted the importance of drawing upon additional disparate sources of digital traces such as projects and industrial facilities to make study results more robust and reliable. In order to utilize and combine a wider and more representative range of data sources, such as scientific publications, patents, descriptions of R&D projects as well as biofuel facilities and projects, the method developed and applied here focused on the data that was common across all of these sources; the textual description and date. It is better to understand and characterize the influence of individual data sources, avenues for further research may include in-depth studies that compare the obtained results from the overall dataset with individual data sources, e.g. using technological similarity measures by Arts et al. (2018).

Secondly, the method includes the creation of a comprehensive dictionary following the hybrid method introduced in Section 3.2, aggregating multiple structured lists of terms. Furthermore, to account for the emergence of new terms over time, as new terms appear, the dictionary can be expanded without changing the measures calculated for previous years. This possibility to expand and update the dictionary gives more flexibility and adaptability with regard to the level of granularity of technologies being analyzed, and it allows for a bottom-up technology discretization strategy that does not require fixed top-down classifications. In addition, as an extension of the proposed method, a comprehensive dictionary can be created extracting terms directly from the whole text corpus instead of using a separate source for their identification. This extension can be performed using text-mining approaches such as bag-of-words, and techniques that seek to extract terms and/or create semantic classifications such as the Latent Dirichlet Allocation (LDA), Latent Semantic Analysis or Doc2Vec (Alghamdi and Alfalqi, 2015; Feldman and Sanger, 2007).

Thirdly, while also allowing for a more in-depth exploration of patterns, the method provides a summary measure of year-to-year change based on the configurational similarity between the matrices of one year and another. This measure can be used for the representation of a combinatorial view that can be applied as an aggregate view of the trends within the given domain regarding specific technologies. Such a view is complementary to currently used indicators for tracking year-to-year changes of specific technology domains, which are often related to inputs (e.g. R&D investment and R&D personnel statistics) or outputs (e.g. scientific production in terms of overall numbers of scientific papers and citations) and not the process itself. In addition, although technological change is highly context specific (van der Vooren, 2014), the proposed indicator goes beyond single domain case-based studies and, as such, permits comparison of configurational changes between different technological areas.

From a point of view of the implications for industry- and policy practice, the emphasis in this study is on the temporal perspective of technological changes. Insights obtained by complementary studies, e.g. Arts et al. (2018) and Arts and Fleming (2019), focusing on the influence of motivation of individual entrepreneurs, add opportunities for additional analysis information and context for qualifying the descriptions of the gathered dataset. More specifically, the analytical approaches presented in Arts and Fleming (2019) and Fleming et al. (2007) support the distinction between the influence of social interaction mechanisms and individual characteristics on inventive output. This may support future exploration of multiple phenomena such as the localization of knowledge spill overs and the

influence of technological and spatial proximity. In a similar manner, information obtained from other complementary studies that utilize explicit links (references) between digital documents (Érdi et al., 2013; Uzzi et al., 2013) might expand understanding of the technological domain under analysis.

The present study has three main limitations. The first limitation is that the proposed indicator does not capture aspects such as impact, performance or cost within the analyzed technological domain. For that reason, it is most appropriate for early-stage R&D activities to identify and define overall technological configurations. The second limitation is associated with the interpretation of term co-occurrences that are stored in the adjacency matrices. They are valid only at the aggregate document corpus level, i.e. the findings should not be extended to individual documents. Finally, the third limitation is that the volume of document records affects the possibility of finding a given term within the analyzed corpus. Therefore, this type of analysis is more suitable for large document corpora composed of several thousand records.

In conclusion, the novel method for quantifying technological change proposed here and its application using a large-scale dataset of worldwide bioenergy R&D is intended for technological change scholars, large-scale systems design researchers, technology forecasters in industry, and R&D policy makers. The proposed indicator for technological change as a combinatorial process opens the following avenues for further impact pathways: 1) comparison of technological change curves between different countries and industries based on combinatorial technological change principles, 2) development of predictive models to anticipate the next period of radical or disruptive technological changes, 3) comparative studies of different data sources (e.g. patents and publications) better to study the similarities and differences in their evolutionary patterns, and finally 4) provision of analytical support to build recommendation engines that use the network structure of the combinatorial process to identify and predict technological combinations that have not happened yet but are statistically likely to occur.

Acknowledgment

Funding has been received from the European Union (EU) Horizon 2020 Framework Programme for Research and Innovation under Grant Agreement No. 770420 – EURITO.

References

Abdi, H., 2007. RV coefficient and congruence coefficient. Encycl. Meas. Stat. 849–853.
Abercrombie, R.K., Udoeyop, A.W., Schlicher, B.G., 2012. A study of scientometric methods to identify emerging technologies via modeling of milestones. Scientometrics 91, 327–342. https://doi.org/10.1007/s11192-011-0614-4.
Adegbesan, J.A., Ricart, J.E., 2007. What do we really know about when technological innovation improves performance (and when it does not)? (January 2007). IESE Business School Working Paper No. 668. Available at SSRN: https://ssrn.com/abstract=982335 or https://doi.org/10.2139/ssrn.982335.
Alghamdi, R., Alfalqi, K., 2015. A survey of topic modeling in text mining. Int. J. Adv. Comput. Sci. Appl. 6, 147–153. https://doi.org/10.14569/IJACSA.2015.060121.
Alshamsi, A., Pinheiro, F.L., Hidalgo, C.A., 2018. Optimal diversification strategies in the networks of related products and of related research areas. Nat. Commun. 9. https://doi.org/10.1038/s41467-018-03740-9.
Aro, E.M., 2016. From first generation biofuels to advanced solar biofuels. Ambio 45, 24–31. https://doi.org/10.1007/s13280-015-0730-0.
Arthur, W.B., 2009. The Nature of technology: What it is and How it Evolves. Free Press.
Arthur, W.B., Polak, W., 2006. The evolution of technology within a simple computer model. Complexity 11, 23–31. https://doi.org/10.1002/cplx.20130.
Arts, S., Cassiman, B., Gomez, J.C., 2018. Text matching to measure patent similarity. Strateg. Manag. J. 39, 62–84. https://doi.org/10.1002/smj.2699.
Arts, S., Fleming, L., 2019. Paradise of novelty-or loss of human capital? exploring new fields and inventive output. Organ. Sci. 29, 1074–1092. https://doi.org/10.1287/orsc.2018.1216.
Arts, S., Veugelers, R., 2018. CEPR Discussion Paper No. DP1270. . https://doi.org/10.2139/ssrn.2877108.
Arts, S., Veugelers, R., 2015. Technology familiarity, recombinant novelty, and breakthrough invention. Ind. Corp. Chang. 24, 1215–1246. https://doi.org/10.1093/icc/dtu029.
Benson, C.L., Magee, C.L., 2013. A hybrid keyword and patent class methodology for selecting relevant sets of patents for a technological field. Scientometrics 96, 69–82. https://doi.org/10.1007/s11192-012-0930-3.
Buckland, M., Gey, F., 1994. The relationship between recall and precision. J. Am. Soc. Inf. Sci. 45, 12–19. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L.
Carnabuci, G., Bruggeman, J., 2009. Knowledge specialization, knowledge brokerage and the uneven growth of technology domains. Soc. Forces 88, 607–641. https://doi.org/10.1353/sof.0.0257.
Chang, P.-L., Wu, C.-C., Leu, H.-J., 2010. Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display. Scientometrics 82, 5–19. https://doi.org/10.1007/s11192-009-0033-y.
Chang, S.-B., Lai, K.-K., Chang, S.-M., 2009. Exploring technology diffusion and classification of business methods: using the patent citation network. Technol. Forecast. Soc. Change 76, 107–117. https://doi.org/10.1016/j.techfore.2008.03.014.
Cho, H., Choi, W., Lee, H., 2017. A method for named entity normalization in biomedical articles: application to diseases and plants. BMC Bioinform. 18, 1–12. https://doi.org/10.1186/s12859-017-1857-8.
Choe, H., Lee, D.H., Seo, I.W., Kim, H.D., 2013. Patent citation network analysis for the domain of organic photovoltaic cells: country, institution, and technology field. Renew. Sustain. Energy Rev. 26, 492–505. https://doi.org/10.1016/j.rser.2013.05.037.
Choi, J., Hwang, Y.S., 2014. Patent keyword network analysis for improving technology development efficiency. Technol. Forecast. Soc. Change 83, 170–182. https://doi.org/10.1016/j.techfore.2013.07.004.
Chuck, C.J., 2016. Biofuels for aviation: feedstocks. Technology and Implementation. Elsevier. https://doi.org/10.1016/C2014-0-03505-8.
Biofuel Digest, 2018. DIgestData – The advanced bioeconomy data portal [WWW document]. URLhttp://biofuelsdigest.com/digestdata/.
Clarivate, 2018a. Web of science [WWW document]. http://www.webofknowledge.com.
Clarivate, 2018b. Derwent innovations index [WWW document]. URLhttp://www.webofknowledge.com.
COMET Centre, 2018. Bioenergy 2020+ [WWW document]. URL https://www.bioenergy2020.eu/en/home.
Cook, H.V., Jensen, L.J., 2019. A guide to dictionary-based text mining. In: Methods in Molecular Biology, pp. 73–89. https://doi.org/10.1007/978-1-4939-9089-4_5.
Curci, Y., Mongeau Ospina, C.A., 2016. Investigating biofuels through network analysis. Energy Policy 97, 60–72. https://doi.org/10.1016/j.enpol.2016.07.001.
Dale, R., Moisl, H., Somers, H.L., 2000. Handbook of Natural Language Processing. Marcel Dekker, New York.
Dernis, H., Squicciarini, M., de Pinho, R., 2015. Detecting the emergence of technologies and the evolution and co-development trajectories in science (DETECTS): a 'burst' analysis-based approach. J. Technol. Transf. https://doi.org/10.1007/s10961-015-9449-0.
Dolfsma, W., Leydesdorff, L., 2009. Lock-in and break-out from technological trajectories: modeling and policy implications. Technol. Forecast. Soc. Change 76, 932–941. https://doi.org/10.1016/j.techfore.2009.02.004.
Duguet, E., MacGarvie, M., 2005. How well do patent citations measure flows of technology? Evidence from French innovation surveys. Econ. Innov. New Technol 14, 375–393. https://doi.org/10.1080/1043859042000307347.
Engelsman, E.C., van Raan, A.F.J., 1994. A patent-based cartography of technology. Res. Policy 23, 1–26. https://doi.org/10.1016/0048-7333(94)90024-8.
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.
European Commission, 2018. CORDIS EU dataset [WWW Document]. URLhttps://cordis.europa.eu/about.
European Technology and Innovation Platform Bioenergy, 2018. ETIP Bioenergy [WWW Document]. URLhttp://www.etipbioenergy.eu/.
Everett, M.G., Borgatti, S.P., 2013. The dual-projection approach for two-mode networks. Soc. Networks 35, 204–210. https://doi.org/10.1016/j.socnet.2012.05.004.
Feldman, R., Sanger, J., 2007. The Text Mining handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York, NY.
Ferreira, A.G., Lião, L.M., Monteiro, M.R., 2013. Biofuels. EMagRes. John Wiley & Sons, Chichester, UK, pp. 529–540. Ltd. https://doi.org/10.1002/9780470034590.emrstm1351.
Fleming, L., Mingo, S., Chen, D., 2007. Collaborative brokerage, generative creativity, and creative success. Adm. Sci. Q. 52, 443–475. https://doi.org/10.2189/asqu.52.3.443.
Fleming, L., Sorenson, O., 2001. Technology as a complex adaptive system: evidence from patent data. Res. Policy 30, 1019–1039. https://doi.org/10.1016/S0048-7333(00)00135-9.
Fleming, L., Waguespack, D.M., 2007. Brokerage, boundary spanning, and leadership in open innovation communities. Organ. Sci. 18, 165–180. https://doi.org/10.1287/orsc.1060.0242.
Funk, R.J., Owen-Smith, J., 2017. A dynamic network measure of technological change. Manage. Sci. 63, 791–817. https://doi.org/10.1287/mnsc.2015.2366.
Geels, F.W., 2005. Processes and patterns in transitions and system innovations: refining

the co-evolutionary multi-level perspective. Technol. Forecast. Soc. Change 72, 681–696. https://doi.org/10.1016/j.techfore.2004.08.014.
Genscape, 2018. Genscape biofuels[WWW Document]. URL https://apps.genscape.com/Biofuels/.
Goldfarb, B., 2005. Diffusion of general-purpose technologies: understanding patterns in the electrification of US manufacturing 1880-1930. Ind. Corp. Chang. 14, 745–773. https://doi.org/10.1093/icc/dth068.
Guan, J., Liu, N., 2016. Exploitative and exploratory innovations in knowledge network and collaboration network: a patent analysis in the technological field of nano-energy. Res. Policy 45, 97–112. https://doi.org/10.1016/j.respol.2015.08.002.
Bioenergy Research: Advances and Applications. In: Gupta, V.K., Tuohy, M.G., Kubicek, C.P., Saddler, J., Xu, F. (Eds.), Bioenergy Research: Advances and Applications. Elsevier. https://doi.org/10.1016/C2012-0-00025-7.
Guthrie, S., Wamae, W., Diepeveen, S., Wooding, S., Grant, J., Europe, R., 2013. Measuring research: a guide to research evaluation frameworks and tools. RAND Monographs. https://doi.org/10.2307/6005.
Hausmann, R., Hidalgo, C.A., 2011. The network structure of economic output. J. Econ. Growth 16, 309–342. https://doi.org/10.1007/s10887-011-9071-4.
Hekkert, M.P., Suurs, R.A.A., Negro, S.O., Kuhlmann, S., Smits, R.E.H.M., 2007. Functions of innovation systems: a new approach for analysing technological change. Technol. Forecast. Soc. Change 74, 413–432. https://doi.org/10.1016/j.techfore.2006.03.002.
Henderson, R.M., Clark, K.B., 1990. Architectural innovation: the reconfiguration of existing product technologies and the failure of established firms. Adm. Sci. Q. 35, 9. https://doi.org/10.2307/2393549.
Idaho National Laboratory, 2018. Bioenergy feedstock library[WWW Document]. URL https://bioenergylibrary.inl.gov/Home/.
Järvenpää, H.M., Mäkinen, S.J., Seppänen, M., 2011. Patent and publishing activity sequence over a technology's life cycle. Technol. Forecast. Soc. Change 78, 283–293. https://doi.org/10.1016/J.TECHFORE.2010.06.020.
Johnson, S.C., 1967. Hierarchical clustering schemes. Psychometrika 32, 241–254. https://doi.org/10.1007/BF02289588.
Jorgenson, D.W., 2001. Information technology and the U.S. Economy. Am. Econ. Rev. 91, 1–32. https://doi.org/10.1257/aer.91.1.1.
Josse, J., Pagès, J., Husson, F., 2008. Testing the significance of the RV coefficient. Comput. Stat. Data Anal. 53, 82–91. https://doi.org/10.1016/j.csda.2008.06.012.
Joung, J., Kim, K., 2017. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. Technol. Forecast. Soc. Change 114, 281–292. https://doi.org/10.1016/j.techfore.2016.08.020.
Kajikawa, Y., Takeda, Y., 2008. Structure of research on biomass and bio-fuels: a citation-based approach. Technol. Forecast. Soc. Change 75, 1349–1359. https://doi.org/10.1016/j.techfore.2008.04.007.
Kajikawa, Y., Yoshikawa, J., Takeda, Y., Matsushima, K., 2008. Tracking emerging technologies in energy research: toward a roadmap for sustainable energy. Technol. Forecast. Soc. Change 75, 771–782. https://doi.org/10.1016/j.techfore.2007.05.005.
Kim, J., Magee, C.L., 2017. Dynamic patterns of knowledge flows across technological domains: empirical results and link prediction. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2990729.
Kook, S.H., Kim, K.H., Lee, C., 2017. Dynamic technological diversification and its impact on firms' performance an empirical analysis of Korean IT firms. Sustain. 9. https://doi.org/10.3390/su9071239.
Kostoff, R.N., Toothman, D.R., Eberhart, H.J., Humenik, J.A., 2001. Text mining using database tomography and bibliometrics: a review. Technol. Forecast. Soc. Change 68, 223–253. https://doi.org/10.1016/S0040-1625(01)00133-0.
Lee, C., Jeon, J., Park, Y., 2011. Monitoring trends of technological changes based on the dynamic patent lattice: a modified formal concept analysis approach. Technol. Forecast. Soc. Change 78, 690–702. https://doi.org/10.1016/j.techfore.2010.11.010.
Lee, J., Berente, N., 2013. The era of incremental change in the technology innovation life cycle: an analysis of the automotive emission control industry. Res. Policy 42, 1469–1481. https://doi.org/10.1016/j.respol.2013.05.004.
Lee, W.S., Han, E.J., Sohn, S.Y., 2015. Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. Technol. Forecast. Soc. Change 100, 317–329. https://doi.org/10.1016/j.techfore.2015.07.022.
Li, M., 2015. A novel three-dimension perspective to explore technology evolution. Scientometrics 105, 1679–1697. https://doi.org/10.1007/s11192-015-1591-9.
Liu, W., Gu, M., Hu, G., Li, C., Liao, H., Tang, L., Shapira, P., 2014. Profile of developments in biomass-based bioenergy research: a 20-year perspective. Scientometrics 99, 507–521. https://doi.org/10.1007/s11192-013-1152-z.
Moed, H.F., De Bruin, R.E., Van Leeuwen, T.N., 1995. New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. Scientometrics 33, 381–422. https://doi.org/10.1007/BF02017338.
Moore, G.E., 1998. Cramming more components onto integrated circuits. Proc. IEEE 86, 82–85. https://doi.org/10.1109/JPROC.1998.658762.
Moro, A., Boelman, E., Joanny, G., Garcia, J.L., 2018. A bibliometric-based technique to identify emerging photovoltaic technologies in a comparative assessment with expert review. Renew. Energy 123, 407–416. https://doi.org/10.1016/j.renene.2018.02.016.
Nadeau, D., 2007. A survey of named entity recognition and classification. Linguist. Investig. 3–26. https://doi.org/10.1075/li.30.1.03nad.
National Research Council, 2014. Capturing Change in Science, Technology, and Innovation. National Academies Press, Washington, D.C.. https://doi.org/10.17226/18606.
Noh, H., Jo, Y., Lee, S., 2015. Keyword selection and processing strategy for applying text mining to patent analysis. Expert Syst. Appl. 42, 4348–4360. https://doi.org/10.1016/j.eswa.2015.01.050.
Nosella, A., Petroni, G., Salandra, R., 2008. Technological change and technology monitoring process: evidence from four Italian case studies. J. Eng. Technol. Manag. 25, 321–337. https://doi.org/10.1016/J.JENGTECMAN.2008.10.001.
NREL, 2018. Biofuels Atlas [WWW Document]. URLhttps://maps.nrel.gov/biofuels-atlas/.
O'Keeffe, A., McCarthy, M., 2012. The Routledge handbook of Corpus linguistics. Routledge. Milton Park Abingdon Oxon, New York.
Biofuels. In: Pandey, A., Larroche, C., Gnansounou, E. (Eds.), Biofuels. Elsevier. https://doi.org/10.1016/C2010-0-65927-X.
Parayil, G., 1993. Models of technological change: a critical review of current knowledge. Hist. Technol. 10, 105–126. https://doi.org/10.1080/07341519308581840.
Park, H., Magee, C.L., 2017. Tracing technological development trajectories: a genetic knowledge persistence-based main path approach. PLoS ONE 12, 1–18. https://doi.org/10.1371/journal.pone.0170895.
Phillips, F., Linstone, H., 2016. Key ideas from a 25-year collaboration at technological forecasting & social change. Technol. Forecast. Soc. Change 105, 158–166. https://doi.org/10.1016/j.techfore.2016.01.007.
Popper, R., 2008. How are foresight methods selected? Foresight 10, 62–89. https://doi.org/10.1108/14636680810918586.
Ramsay, J.O., ten Berge, J., Styan, G.P.H., 1984. Matrix correlation. Psychometrika 49, 403–423. https://doi.org/10.1007/BF02306029.
REEEP, 2018. reegle.info [WWW Document]. URLhttps://www.reeep.org/reegleinfo.
Robert, P., Escoufier, Y., 1976. A unifying tool for linear multivariate statistical methods: the RV- Coefficient. Appl. Stat. 25, 257. https://doi.org/10.2307/2347233.
Schumpeter, J.A., 1934. The theory of economic development : an inquiry into profits, capital, credit, interest, and the business cycle. Transaction Publishers, Piscataway.
Smilde, A.K., Kiers, H.A.L., Bijlsma, S., Rubingh, C.M., Van Erk, M.J., 2009. Matrix correlations for high-dimensional data: the modified RV-coefficient. Bioinformatics 25, 401–405. https://doi.org/10.1093/bioinformatics/btn634.
Solé, R., Amor, D.R., Valverde, S., 2016. On singularities and black holes in combination-driven models of technological innovation networks. PLoS ONE 11, 1–13. https://doi.org/10.1371/journal.pone.0146180.
Sternitzke, C., Bartkowski, A., Schramm, R., 2008. Visualizing patent statistics by means of social network analysis tools. World Pat. Inf. 30, 115–131. https://doi.org/10.1016/j.wpi.2007.08.003.
Strumsky, D., Lobo, J., van der Leeuw, S., 2012. Using patent technology codes to study technological change. Econ. Innov. New Technol. 21, 267–286. https://doi.org/10.1080/10438599.2011.578709.
Suominen, A., 2013. Analysis of technological progression by quantitative measures: a comparison of two technologies. Technol. Anal. Strateg. Manag. 25, 687–706. https://doi.org/10.1080/09537325.2013.802930.
Suominen, A., Seppänen, M., 2014. Bibliometric data and actual development in technology life cycles: flaws in assumptions. Foresight 16, 37–53. https://doi.org/10.1108/FS-03-2013-0007.
Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., Pietronero, L., 2012. A new metrics for countries' fitness and products' complexity. Sci. Rep. 2, 1–4. https://doi.org/10.1038/srep00723.
Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical combinations and scientific impact. Science 342, 468–472. (80-.). https://doi.org/10.1126/science.1240474.
van den Oord, A., van Witteloostuijn, A., 2018. A multi-level model of emerging technology: an empirical study of the evolution of biotechnology from 1976 to 2003. PLoS ONE 13, e0197024. https://doi.org/10.1371/journal.pone.0197024.
van den Oord, J., 2010. The ecology of technology: the co-evolution of technology and organization. Eindhoven: Technische Universiteit Eindhoven. 308 p. https://doi.org/10.6100/IR658253.
van der Vooren, A., 2014. Accelerating Technological Change: Towards a more Sustainable Transport System. Utrecht University.
Venugopalan, S., Rai, V., 2015. Topic based classification and pattern identification in patents. Technol. Forecast. Soc. Change 94, 236–250. https://doi.org/10.1016/j.techfore.2014.10.006.
Wachsmuth, H., 2015. Text Analysis Pipelines. In: Text Analysis Pipelines. Lecture Notes in Computer Science, vol 9383. Springer, Cham.
Yayavaram, S., Ahuja, G., 2008. Decomposability in knowledge structures and its impact on the usefulness of inventions and knowledge-base malleability. Adm. Sci. Q. 53, 333–362. https://doi.org/10.2189/asqu.53.2.333.
Yoon, B., Park, Y., 2004. A text-mining-based patent network: analytical tool for high-technology trend. J. High Technol. Manag. Res. 15, 37–50. https://doi.org/10.1016/j.hitech.2003.09.003.
Yoon, J., Choi, S., Kim, K., 2011. Invention property-function network analysis of patents: a case of silicon-based thin film solar cells. Scientometrics 86, 687–703. https://doi.org/10.1007/s11192-010-0303-8.
Youn, H., Strumsky, D., Bettencourt, L.M.A., Lobo, J., 2015. Invention as a combinatorial process: evidence from US patents. J. R. Soc. Interface 12 20150272–20150272. https://doi.org/10.1098/rsif.2015.0272.

Pedro Parraguez received the M.Sc. degree in Innovation and Technology Management from the University of Bath, U.K., in 2010, and the Ph.D. degree in Engineering Systems from DTU - Technical University of Denmark, Denmark, where he continued as a post-doctoral researcher until 2018. He is currently the co-founder and CEO of Dataverz, a technology-based startup that develops decision-support systems to tackle societal challenges. Pedro's research and applied work are focused on complex socio-technical systems, with emphasis on network science and data-driven analyses. His work includes the study and development of decision-making support for industrial clusters, complex organisations, and large engineering projects.

Stanko Škec received the M.Sc. degree in Mechanical Engineering from the University of Zagreb, Croatia, in 2010, and the Ph.D. degree at the same institution in 2015. He is currently an Assistant Professor at the University of Zagreb, Chair of Design and Product Development. He is also appointed as Visiting Assistant Professor at Engineering Systems Group, DTU - Technical University of Denmark, Lyngby, Denmark. His primary field of research and scientific focus has been a multidisciplinary field of the product-service systems design and development. His work includes product lifecycle management, knowledge management, innovation management and product development processes management and monitoring.

Duarte Oliveira e Carmo is a Digital Consultant at GN Audio. Previously, he studied in the University of Lisbon, the Beijing Institute of Technology, and the Technical University of Denmark (DTU) where he obtained his master's degree in engineering management. His thesis, “Measuring the uniqueness of technological capabilities: A data-driven network exploration”, focused on studying and analysing a complex research dataset of more than 10.000 patents and scientific publications. His interests lie at the intersection of Management, Innovation, Data Science, and Open Source software.

Anja M. Maier received the M.A. degree in political science, communication science, and philosophy from the University of Muenster, Muenster, Germany, in 2002 and the Ph.D. degree in engineering design from the University of Cambridge, Cambridge, U.K., in 2007. She is Professor of Engineering Systems Design with DTU - Technical University of Denmark, Lyngby, Denmark. She was also a Consultant in the manufacturing and software industries. Professor Maier's research focuses on engineering systems design, with a particular emphasis on complexity and human behaviour. This includes design in networks, design communication, and design cognition. Dr Maier serves on the Editorial Board of the Journal of Engineering Design, as Associate Editor of the journal Design Science, and as Advisory Board Member of the worldwide Design Society. Professor Maier is a fellow of the Cambridge Philosophical Society, a member of INCOSE, of the National Academy of Science and Engineering (acatech), Germany, and of the Danish Academy of Technical Sciences (ATV).