Technological Forecasting & Social Change 177 (2022) 121559

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

A text-embedding-based approach to measuring patent-to-patent technological similarity☆,☆☆

Daniel S. Hain [*,a], Roman Jurowetzki [a], Tobias Buchmann [b], Patrick Wolf [b]

[a] AI:Growth Lab, Aalborg University Business School, Denmark
[b] Centre for Solar Energy and Hydrogen Research Baden-Württemberg (ZSW), Germany

ARTICLE INFO

Keywords:
Technological similarity
Patent data
Natural-language processing
Technology network
Patent landscaping
Patent quality

ABSTRACT

This paper describes an efficiently scaleable approach to measuring technological similarity between patents by combining embedding techniques from natural language processing with nearest-neighbor approximation. Using this methodology, we are able to compute similarities between all existing patents, which in turn enables us to represent the whole patent universe as a technological network. We validate both technological signature and similarity in various ways and, using the case of electric vehicle technologies, demonstrate their usefulness in measuring knowledge flows, mapping technological change, and creating patent quality indicators. This paper contributes to the growing literature on text-based indicators for patent analysis. We provide thorough documentation of our methods, including all code, and indicators at https://github.com/AI-Growth-Lab/patent_p2p_similarity_w2v).

1. Introduction

Patent data has long been used as a measure of innovative activity and performance (Griliches, 1990; Pavitt, 1985; 1988; Schmookler, 1966), with recent decades seeing a sharp increase in the use of patent-based indicators by academics and policymakers alike. It has been deployed, among other uses, to assess the innovation performance of countries (Fu and Yang, 2009; Tong and Davidson, 1994), sectors (Pavitt, 1984), and firms (Ernst, 2001; Hagedoorn and Cloodt, 2003). Its wide availability and coverage across sectors and countries and over time makes patent data one of the go-to data sources for analysing pattern of technological innovation.

Technological similarities between patent pairs or larger patent portfolios represent a key enabling indicator for patent-based technology analysis. Patent-to-patent (p2p) technological similarity analysis enables the application of methods and techniques from network analysis to map and understand the structure of technology on various levels of aggregation (e.g. patent, firm, technology, geographical region). A large body of literature has leveraged p2p technological similarity indicators to assess knowledge spillovers (Jaffe et al., 1993) and visualize innovation opportunities (Breschi et al., 2003). P2p technological similarity also has many applications in technology analysis and forecasting (Hain et al., 2021), such as technology mapping and landscaping (e.g. Aharonson and Schilling, 2016; Alstott et al., 2017; Kogler et al., 2013), technology convergence prediction (e.g. San Kim and Sohn, 2020), disruptive technology detection (e.g. Zhou et al., 2020), and patent quality assessment (e.g. Arts et al., 2018; 2020).

The wide range of approaches to p2p technological similarity commonly use one or both of two key data sources: (i.) technology classifications (e.g., IPC, CPC, e.g., Aharonson and Schilling, 2016; Hain et al., 2018; Singh and Marx, 2013)., and (ii.) bibliographic data on the forward or backward citation of patents (e.g., Barirani et al., 2013; Huang et al., 2003). Both are subject to a number of limitations. While technology (sub)classes are well suited to categorizing technologies, they are usually too broad to identify the concrete technological content of a patent (Archibugi and Planta, 1996; Righi and Simcoe, 2019; Thompson and Fox-Kean, 2005), and their static nature limits their usefulness in analysing dynamic phenomena such as the emergence (Kay et al., 2014) or convergence (Preschitschek et al., 2013) of technology. Patent citations, meanwhile, are intended to reflect prior work

☆ This article belong to the special section on Advanced Tech Mining: Technical Emergence Indicators and Measurements.☆☆ All code necessary to recreate our workflow, indicator creation, and analysis is freely available at https://github.com/daniel-hain/patent_embedding_research. All data is also available for download and use in third-party analysis. Financial support for ZSW's research provided by BMBF Kopernikus ENavi (FKZ:03SFK4W0). – Workflow, Code, and Applications –

* Corresponding author.
E-mail address: dsh@business.aau.dk (D.S. Hain).

https://doi.org/10.1016/j.techfore.2022.121559
Received 6 March 2020; Received in revised form 31 January 2022; Accepted 3 February 2022
Available online 10 February 2022

rather than technological content, and citation practices exhibit strategic behavior (Criscuolo and Verspagen, 2008; Lampe, 2012), vary across individuals and jurisdictions (Lemley and Sampat, 2012; Picard and de la Potterie, 2013), and are subject to home bias and other forms of bias (Alcacer and Gittelman, 2006; Bacchiocchi and Montobbio, 2010; Griffith et al., 2011; Li, 2014).

Recently, a growing body of research has focused on applying natural language processing (NLP) techniques to the textual components of patent data to derive p2p technological similarity indicators. The textual description of a patent (i.e., title, abstract, claims, and full text) contains all necessary information to allow readers familiar with the relevant patent domains to comprehend the embodied technologies and functionalities. To distill this information in an automated manner, a variety of techniques have been applied, ranging from keyword extraction (Arts et al., 2018; 2020; Gerken and Moehrle, 2012; Kelly et al., 2018; Lee et al., 2009; Moeller and Moehrle, 2015; Yoon, 2008) to the linguistic analysis of subject-action-object (SAO) structure (Li et al., 2020; Sternitzke and Bergmann, 2009; Yufeng et al., 2016) or ontology (Soo et al., 2006; Taduri et al., 2011). While NLP techniques have already broadened our methodological toolbox for patent analysis, several challenges remain. Keyword-based approaches are relatively simple to implement and comprehend, yet their interpretation is complicated by the richness of domain-specific vocabulary, technical and legal jargon, synonyms (the same technology termed differently across domains), and homonyms (the same word referring to different technologies across domains) typical of patent text (Beall and Kafadar, 2008; Qi et al., 2020; Tseng et al., 2007). Methods analyzing patents' SAO structure or ontology express semantic information better (Yang et al., 2017), but in turn are more time-consuming to calibrate and interpret and require domain-expert knowledge.

The use of deep learning-based (LeCun et al., 2015) embedding techniques (Mikolov et al., 2013; Pennington et al., 2014) has led to a paradigm shift in NLP, achieving unprecedented performance in many language tasks such as text classification, translation, and semantic search. Such embeddings enable the creation of latent vector representations of textual data that, to a large extent, preserve the original context and meaning. The potential of embedding techniques has recently been further demonstrated by improvements in automated patent classification tasks (e.g. Bekamiri et al., 2021; Grawe et al., 2017; Kim et al., 2020; Lee and Hsiang, 2020; Li et al., 2018; Risch and Krestel, 2019). This suggests that embedding techniques can be used to improve on earlier attempts to create text-based p2p technological similarity measures.

In this paper, we contribute to the research on patent-based technology mapping by providing a framework leveraging embedding techniques based on textual data to i.) create technological signature vector for a given patent and ii.) derive measures of patent-to-patent (p2p) technological similarity. We create technological signature vectors for all patents in the PATSTAT database based on their abstract text. We then apply an approximate nearest neighbor search that allows us to process massive datasets and compute p2p similarity measures for the whole universe of patents. We evaluate the validity of a patent's technological signature and derived p2p similarity in multiple ways. We first evaluate the quality and usefulness of the derived technological signature for automated technology classification as well as for semantic search. To evaluate the p2p technological similarity measure, we replicate a set of stylized facts and compare them to recent non-embedding-based approaches. Lastly, for the case of electric vehicle (EV) technologies, we showcase potential research applications in technology mapping and the creation of patent quality indicators to identify technological cross-country knowledge flows.

Through this process, we address several key gaps in the ongoing effort to leverage textual data for patent-based technology mapping and forecasting. First, we advance research on the creation of p2p similarity measures based on an embeddings approach that is more efficient and less sensitive to domain-specific jargon than keyword-based approaches (eg. Arts et al., 2018; 2020; Kelly et al., 2021). Second, by utilizing an approximate nearest neighbor search, we enable large-scale cross--technology and worldwide applications for technology mapping and forecasting without the use of supercomputing infrastructure. Finally, we contribute to the cumulative nature of research related to p2p similarity, and open and reproducible science more broadly, by sharing well-documented code and workflow instructions, as well as all intermediate and final outcomes, such as the p2p similarity measures, word-embeddings, and technological signatures. These are available at https://github.com/AI-Growth-Lab/patent_p2p_similarity_w2v. This will facilitate reconstruction and adaptation of our data and methods and support future research on the use, creation, advancement, and evaluation of text-based p2p technological similarity indicators.

The remainder of this paper is structured as follows. In Section 2, we review the literature on patent-based technology analysis, focusing on approaches to measuring p2p technological similarity. We discuss and contrast approaches based on citations, technology classification, and text data. In Section 3, we discuss methodological considerations and describe our approach to creating text-based technological signatures of patents and deriving measures of p2p technological similarity. We apply and evaluate these techniques and the obtained results in Section 4 for all patents found in PATSTAT. In Section 5, we explore the results of our analysis for the case of EV patents and demonstrate potential research applications. Finally, Section 6 offers concluding observations and points towards promising avenues for future research.

2. Technological similarity: Literature review and state of the field

2.1. Technology class-based approaches to measuring technological similarity

The "International Patent Classification" (IPC) or "Cooperative Patent Classification" (CPC) are taxonomies aimed at capturing the entire universe of patented technology (McNamee, 2013), providing a complex hierarchy of categories to aggregate technological concepts on different levels. Technology classes have frequently been leveraged as a foundation for measuring patent similarity (Zhang et al., 2016); co-classification analyses in particular have been widely applied (Boyack and Klavans, 2008; Suh, 2017). Such analyses measure the similarity between technology fields by examining the co-occurrence of classification codes between different patents (Engelsman and van Raan, 1994). Co-classification approaches have the disadvantage that they usually only consider only direct overlap and do not take into account the potential similarity of assigned technology classes. More recent approaches therefore derive similarity measures based on the underlying similarity structure of the assigned technologies (e.g. Aharonson and Schilling, 2016).

While a large body of research utilises technology classifications for patent analysis, other research has critiqued this approach. To start with, technology classification systems are said to be too general to satisfy the specific needs of technological forecasting, research planning, technological positioning or strategy making (Archibugi and Planta, 1996). They can be seen as rather vague (Zhang et al., 2016) as researchers are limited to rigid predefined classes. Existing classification schemes do not sufficiently capture the technological characteristics of an invention on the class level (Benner and Waldfogel, 2008; Preschitschek et al., 2013; Thompson and Fox-Kean, 2005), while the more fine-grained group- and subgroup-level classes are less stable (WIPO, 2017). On these finer levels, technology classes also tend to display a substantial overlap, leading to very technologically similar patents being placed in distant classes (McNamee, 2013). A further challenge in relying on patent classification schemes is that, as technology changes, similar technology-oriented applications may draw from patents in different hierarchical categories (Kay et al., 2014). Classification systems define new technologies based on pre-existing technologies or their

combinations, leading to uncertainty regarding the accuracy of patent-class fit for new innovations. Previous research has also found broad heterogeneity across different patent classification schemes in terms of their weighting of technological functionality versus industry-specific applicability (Adams, 2001), concluding that, in practice, several classifications should be applied and considered to provide a more complete picture (Wolter, 2012). Even within a single classification scheme, between countries, important features may be lost in the process of classifying the technical ideas described in the patent using a common language (Meguro and Osabe, 2019). Finally, due to the reliance on human judgement, patents are sometimes also just poorly categorized by the respective authorities (Leydesdorff, 2008).

2.2. Citation-based approaches used to measure technological similarity

Patent citation represents another popular data source with which to measure p2p similarity. Within this methodological category, we can distinguish three main approaches: (i.) co-citation, (ii.) bibliographic coupling, and (iii.) direct or indirect citation.

Co-citation approaches measure p2p similarity using the number of forward citations shared by two patents Yan and Luo (2017), following the supposition that such co-citations signal overlap in patents' technologies or applications. Leveraging forward-citation data, co-citation-based measures are only available ex-post patent application, once sufficient citations have accumulated.

In contrast, bibliographic coupling approaches measure p2p similarity by the number of joint backward citations Yan and Luo (2017), assuming that overlap in two patents' references indicates both are built on and utilize similar technologies (Von Wartburg et al., 2005). Since a patent's backward citations are available at the time of application, derived measures can be used to create ex-ante indicators.

The third approach relies on using direct/indirect citation (paths) to measure p2p similarity. It involves calculating a compound similarity matrix based on a patent citation network, represented by a direct similarity matrix, and the resulting indirect similarity matrices (Wu et al., 2010). This approach has advantages compared to approaches based on co-citation and bibliographic coupling as it provides more complete information, allowing a more precise assessment of the technological distance between two given innovations (Rodriguez et al., 2015; Wu et al., 2010).

Generally, citation-based approaches to deriving similarity measures are subject to several shortcomings. Citation practices differ across patent authorities (Picard and de la Potterie, 2013) and even examiners (Lemley and Sampat, 2012). Further, applicants may withhold citations for strategic reasons (Criscuolo and Verspagen, 2008; Lampe, 2012) or may simply not provide useful citations (Cotropia et al., 2013), which instead have to be added by the examiner (Alcácer et al., 2009). It also cannot be taken for granted that examiners are willing or able to refer to all relevant prior work. Furthermore, both inventors (Griffith et al., 2011; Li, 2014) and patent examiners (Bacchiocchi and Montobbio, 2010) are subject to home bias, i.e., they are more likely to cite patents with higher geographical, social, or institutional proximity. Furthermore, despite innovation in the field of patent databases and search technology, prior work discovery and examination remains a challenging activity, and patent search reports have varying quality and information richness (Michel and Bettels, 2001) and still require substantial domain expertise to be used correctly. Consequently, the absence of citations is not a sufficient condition to establish the absence of similarity, and p2p similarity measures based on citation data are likely to result in an substantial number of false negatives.

2.3. Natural language-based approaches to measuring technological similarity

Recently, researchers have started leveraging text-based approaches (based on, e.g., patent titles, abstract,s keywords, or claims), attempting to describe the technologies embodied in a patent in a more nuanced way to measure p2p similarity and map technology landscapes and evolution. In this regard, different methodologies have been developed, including (i.) keyword-based approaches, (ii.) SAO structure analysis, (iii.) ontology-based analysis, and iv.) machine learning and deep learning approaches.

Keyword-based methods are based on keyword frequency and co-occurrence measures. This type of approach has often been used in the past due to its simplicity and straightforwardness (Kelly et al., 2018; Lee et al., 2009; Moeller and Moehrle, 2015; Yoon, 2008). Keyword data used to measure p2p similarity can take a number of forms, including raw (Arts et al., 2018), "term frequency-inverse document frequency" (TFIDF, Salton and Buckley, 1988), or weighted (Arts et al., 2020) keyword or multi-word (n-gram) (Gerken and Moehrle, 2012) keyword co-occurrence. However, a major shortcoming of keyword-based approaches is that they fail to reflect relationships among the related concepts described by different words. This is particularly true for documents that are rich in domain-specific vocabulary, technical and legal jargon, synonyms (the same technology termed differently across domains), and homonyms (the same word refers to different technologies across domains), as patents usually are (Beall and Kafadar, 2008; Qi et al., 2020; Tseng et al., 2007).

Other studies apply a subject - action - object (SAO) analysis to the structure of patent texts as their semantic representation, aiming to study the meaning and grammatical structure of patents (Yang et al., 2017). As a method for calculating patent similarities, the method has often been combined with additional models and indicators, such as vector or "Visual Syntax Method" (VSM) models (Yufeng et al., 2016), the Sørensen-Dice index (Li et al., 2020), or the Jaccard and Cosine index (Sternitzke and Bergmann, 2009). While such approaches are able to provide a deeper examination of the semantics in texts, a major drawback can be seen in the focus on only a small proportion of the available words, which introduces the possibility of missing relevant information.

Another methodology that has recently gained attention is the analysis of patent texts by their ontology. This approach is based on the construction of an ontology that describes the concepts and respective relations for a specific domain. Based on this domain, a semantic annotation on patent texts is performed. Examples include the analysis system proposed by Taduri et al. (2011) and Soo et al. (2006). While providing strong semantic modelling, the ontology-based approach is highly labor-intensive and context-sensitive, which makes it hard to apply to a broader scope of patents.

While machine learning (and later deep learning) based approaches for text analysis and classification have been around since the 1990s (Hayes and Weinstein, 1990; Newman, 1998), they have only recently found growing attention in patent analysis, mainly for automated patent technology classification. They are able to map complex relationships between unstructured texts and have yielded promising results when applied to patent text (Li, 2018). Tran and Kavuluru (2017) explored text data and machine learning-based classification in the context of the Cooperative Patent Classification (CPC) system. Such exercises build on earlier work in developing automated patent classification for IPC classes (Fall et al., 2003). More recently, the use of deep learning-based embedding techniques has led to a general paradigm shift in NLP and replaced a large set of keyword-based and linguistic approaches to language modelling. Embedding techniques utilize deep neural networks trained on large datasets to create high-dimensional vector representations of words or documents, preserving the original meanings and contexts. Embedding techniques have found applications in patent analysis, mainly for deep learning based automatic patent classification. For instance, Grawe et al. (2017) computed word embeddings in order to develop a patent text classifier. Li et al. (2018) developed a deep learning-based patent classification algorithm that uses convolutional neural networks and word vector embeddings. Chen et al. (2020) developed a method for extracting semantic information from patent texts using deep learning, and Lee and Hsiang (2020) and Bekamiri et al.

Table 1
General overview on methods for patent similarity analysis.

Method		Description	Characteristics		Example
			Advantages	Limitations	
Patent Class-based measures	Co-classification / Classification overlap	Measures distance between technology fields by examining the co-occurrence of the classification codes (IPC/CPC) between different patents	• Easy access to data from well-structured database • Relatively easy to apply	• Comparison is limited to predefined classes of patents • Classification systems are often too coarse	Engelsman and van Raan (1994); Yan and Luo (2017)
Citation-based measures	Direct / indirect	Direct or indirect citation (paths) between patents are used to derive similarity indicators	• Easy access to data from well-structured database • Relatively easy to apply	• Subject to strategic citation and other citation bias	Rodriguez et al. (2015); Wu et al. (2010)
	Co-citation	Measures the knowledge distance by calculating the amount of shared forward citations of two patents	• Easy access to data from well-structured database	• Not applicable for most recent patents as enough forward citations are needed • Not applicable if patent hasn't been cited	Mowery et al. (1998); Rothaermel and Boeker (2008)
	Bibliographic coupling	Measures the knowledge distance by calculating the amount of shared backward citations/references of two patents	• Easy access to data from well-structured database	• Patent references are often not fully comprehensive	Garfield (1966); Von Wartburg et al. (2005)
Text-based measures	SAO-based	Looks at the subject-verb-object structures of patent texts to extract expressions of meaning. The similarity of patents can than be assessed by the similarity of the SAO-structures	• Includes grammatical structure and meaning • Allows to take a deeper look at semantics in texts	• Still considers only a small proportion of the available words (possible miss of relevant information) • Often treats each identified relationship as equally important, which does not necessarily provide an accurate measure of patent similarity	Li et al. (2020); Wang et al. (2019); Yufeng et al. (2016); Sternitzke and Bergmann (2009)
	Keyword-based	Caompares patent texts by frequency and/or co-occurrence of specified keywords	• Established approach • Straightforward in application	• Solely relies on single word or few single words • Fails to reflect the relationship between concepts • Prone to false negatives when synonyms or technology specific jargon is used	Kelly et al. (2018); Moeller and Moehrle (2015); Yoon (2008)
	Ontology-based	Compares patents by representing the textual content (abstracts, claims, etc.) of a patent as an ontology	• Approach provides a strongly semantic modelling	• Highly labour-intensive • Highly context-sensitive • Barely usable for analyses with a very broad scope	Taduri et al. (2011); Soo et al. (2006);
	Embedding-based / vector-based	Transforms text into a numeric vector by using a vector space model. Texts can then be compared by the similarity of the vectors	• Higher accuracy • Able to map complex relationships of unstructured texts • Able to handle large amounts of unstructured Text	• Not straightforward • Often requires either extensive hardware use or a reduction of the number of patents	Whalen et al. (2020); Grawe et al. (2017); Younge and Kuhn (2016); Li et al. (2018)

(2021) applied current state-of-the-art transformer-based language models, achieving unprecedented performance in text-based patent classification.

The ability of embedding-based machine learning models to predict a patent's assigned technology class with high accuracy indicates that embedding techniques indeed preserve the technological content and context of a patent. While proven to perform well in the task of classifying patents with respect to their technology class, they do not provide similarity measures between patents or suggest a workflow for how to do so. In contrast, applications using embedding techniques to derive and evaluate p2p technological similarity measures are scarce, face computational challenges, and are generally not reproducible. Younge and Kuhn (2016) deployed massive distributed computing power to create similarity measures for all patents. Most recently, Whalen et al. (2020) developed a patent similarity dataset based on a vector space model that contains the similarity scores of US utility patents. Overall, computational bottlenecks have resulted in little progress so far, and the few existing approaches can either provide similarity measures only for subsets of patents or require massive computational power and are therefore not accessible for most researchers.

In summary, although they have only recently gained popularity, text-based approaches to patent analysis already have a significant and successful history. In the past, the most frequently used approaches were keyword-based, but more recently the particularities of patent text, such as the frequent use of synonyms and domain-specific technical jargon, have led to a gradual replacement of these approaches by embedding-based NLP methods. These have been applied successfully for patent classification tasks, but less so for measuring technological similarity, and then only on limited subsets of patents and/or by using computational power not accessible for the common researcher. The reasons for this are include the massive computational demands when creating large similarity matrices. Therefore, we face limitations in utilizing such similarity to map and analyze global technology development.

2.4. Summary of the approaches to measuring technological distance

Table 1 summarises the common approaches to measure technological similarity based on patent data. It delineates past work related to patent similarity measures and how modern approaches based on word vectors and text embeddings can improve the quality of these measures. In particular, while we draw on earlier research on text vectorization and embeddings, we are among the first to apply this approach to derive p2p technological similarity measures and provide a reproducible workflow for doing so. Additionally, we have developed a method that delivers relatively high accuracy combined with time efficiency and scalability, allowing its application to very large numbers of patent pairs without the need for supercomputer power.

3. An embedding approach to creating P2P technological similarity measures in patent data

3.1. General logic

Building on recent research in text-based patent analysis as well as methodological advances in the broader NLP field, we apply embedding techniques to patent text. Our aim in doing so is to capture each patent's

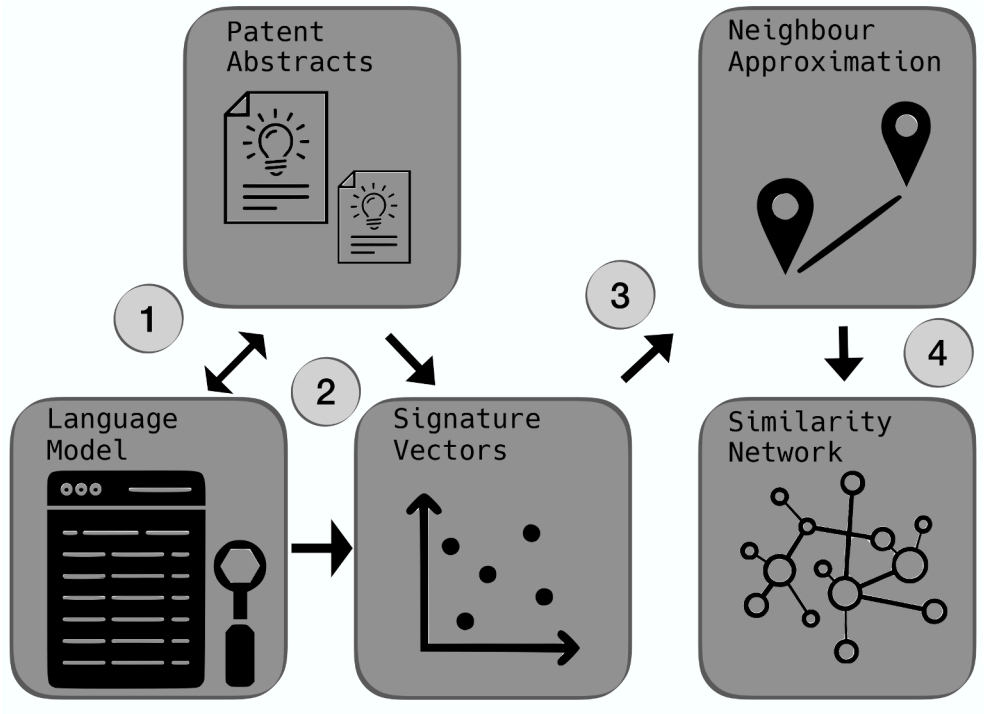

Fig. 1. Preprocessing pipeline.

technological features and content using a high-dimensional numeric vector, which can be interpreted as the patent's "technological signature". We argue that this technological signature not only represents a more appropriate and nuanced characterization of a patent's technological content than citation and IPC class based approaches, but is also a more suitable foundation for p2p similarity measures. In this section, we describe in detail the techniques, parameters, and general logic behind each step of the proposed indicator computation. Figure 1 illustrates the proposed techniques and workflow for creating a p2p technological similarity indicator based on textual data, which can be used for a variety of analyses and indicator creation.

3.2. From patent text to technological signature vector: Creating document embeddings

To provide the desired text-based technological signature, we leverage word and document embedding techniques, representing a methodological breakthrough that has revolutionized NLP research over the last decade. In word embedding models such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), word meanings are learned from the context that surrounds the term rather than merely the within-document co-occurrence. This principle has been famously summarized as "you shall know a word by the company it keeps" (Firth, 1957). Training such models on large datasets allows us to account for syntax and to extract higher-level meaning structures for terms. Over the last decade, embedding has become one of the most promising techniques within NLP, and it is increasingly applied for a broad range of applications such as semantic search and text classification. Recently, it has been demonstrated that embedding techniques are indeed able to infer and encode complex relationships from textual data, such as the relationships between chemical molecules (cf. Tshitoyan et al., 2019). Embedding techniques help us to overcome the limits of keyword-based approaches for patent data, such as synonyms, homonyms, jargon, and changes in meanings over time (Beall and Kafadar, 2008; Qi et al., 2020; Tseng et al., 2007).

Summing and averaging such word vectors has been proven to generate good document representations that are able to deal with some of the idiosyncrasies of natural language that simpler models are unable to account for. To calculate document embeddings, we first train a custom word embedding model, using the Word2Vec approach, on approximately 48 million English-language patent abstracts found in PATSTAT. We train this custom model instead of using generic pre-trained word embeddings due to the arguably specific language found in patent descriptions. In addition, we train a simple TF-IDF model on the whole corpus of patent abstracts. Abstract embeddings are obtained by taking the dot-product of multiplying the word-embedding matrix with the dense TF-IDF weighted bag-of-word representations of the abstracts. As a result, we obtain a 300-dimensional patent signature vector that can be used for further calculations. While TF-IDF vectors sometimes used in similar work are often sparse high-dimensional objects (1 dimension per term in the vocabulary or combinations of terms), and Arts et al. (2020) used 1,362,971 dimensions, our 300-dimensional vectors are in comparison relatively compact, allowing efficient computation and making them easy to share and reuse in different contexts. Below, we describe how we further increase the efficiency of our analytical approach by applying approximate nearest neighbor search and encapsulating the embedding within a "search object".[1]

3.3. From technological signature to p2p similarity: Approximate nearest neighbor search

After creating a technological signature vector for each patent, we attempt to measure p2p technological similarity in the universe of existing patents. For smaller datasets, this can be done with a standard k-nearest neighbor (KNN) search, whereby a similarity score (e.g. euclidean or cosine distance) is calculated for each pair of observation. However, for our population of approx. 48 million patents, this would

[1] Python's Gensim library is used for the training https://radimrehurek.com/gensim/. Bi-grams occurring over 500 times are aggregated into individual tokens before training. The Word2Vec model runs over 5 iterations, using a window of 8 words and 300 dimensions for the target vectors, and terms occurring less than 20 times are not considered.

not be possible with reasonable effort since it would require the calculation of a matrix of the size $n * (n-1)$.[2]

Approximate nearest neighbor computation is an active area of research in machine learning, and one of the common approaches to this problem is using k-d trees that partition the space to reduce the required number of distance calculations. A search of the nearest neighbors is then performed by traversing the resulting tree structure. Utilizing such an approach can reduce the complexity at least to $O[DNlog(N)]$. In our case, this leads to an efficiency increase by a factor of at least $1.12e^4$.[3] In the next step, we calculate the cosine similarity between the focal patent and all other patents found in the neighboring leaves of the search tree:

$$sim^{cosine}(x,y) = \frac{x^T y}{\| x \| \cdot \| y \|} \quad (1)$$

We discard patent pairs with a cosine similarity below the threshold of 0.65 in the relational evaluation (4.3) in order to create an appropriate level of sparsity to avoid the problem of storing and processing extremely large matrices and reduce the analysis to a space where similarity can be meaningfully measured.[4]

Overall, we argue that this approach, although combining several techniques, has its strength in being extremely scalable and efficient. In comparison to many of the other techniques proposed in the literature, our patent vector representations can be created on readily accessible hardware. The resulting approximate nearest neighbor search object is a self-contained file that includes all embeddings and fits within the storage space limitations of most modern notebook computers. Here, n similar patents are identified among the full sample of over 48 million patents within 60 ms (wall time) on average. For large computational tasks, like ours, where the aim is to construct a similarity network for all patents, the object can be preloaded into memory, bringing the calculation time down to under 0.5ms.

3.4. Data

3.4.1. Textual data

For text-based patent analysis, potentially usable text information likely to contain technological information comprises a patent's i.) abstract, ii.) claims, iii.) full text description, or a combination thereof. These text segments are drafted for different purposes and subject to different requirements; therefore, they display different properties when used for a text-based analysis. Previous research indicates that simply combining different textual sources tends to decrease rather than improve the information gained as compared to a single text source analysis (Cetintas and Si, 2012).

Intuitively, a patent's full-text description could be assumed to contain the largest and most nuanced set of information regarding its embodied functionality and technology. However, patent full-text descriptions are subject to greater scrutiny and regulations regarding their format than abstracts and claims, and thus diverge less in terms of length, style, and clarity. Due to their text heterogeneity and increased signal-to-noise ratio, previous research has predominantly favored the use of abstract or claims text (Noh et al., 2015). Among those, a considerable body of research has utilized patent claims text for patent classification (e.g. Lee and Hsiang, 2020) and to infer patent quality indicators such as novelty (e.g. Marco et al., 2019). As a data source for the creation of p2p technological similarity measures, an abstract arguably contains a broader overview of the embodied technologies. An abstract usually contains not only all of the inventions' features given in the patent claims, but also further information on the technical field and examples of possible uses. Further, claims tend to highlight legally protectable differences rather than similarities. Abstracts are therefore better suited when aiming to analyze technological similarity in general, while claims are, for example, well suited for specifically looking at patent infringement. Abstracts are said to "communicate the technical description in a concise and straightforward manner, avoiding unnecessary words that may increase noise in the extraction process" (Tshitoyan et al., 2019, p. 98). Lastly, abstracts, in comparison to claims, are widely available across databases and jurisdictions, easing attempts to carry out global technology mapping.

3.4.2. Patent data

The patent data we use for our study was retrieved from the EPO's Worldwide Patent Statistical Database (PATSTAT, Autumn 2018 edition), which covers bibliographic patent data from more than 100 patent offices of developed as well as developing countries over a period of several decades.

As our first step, we create technological-signature embeddings for all patents where English-language abstracts are available (approx. 48 million). However, for the calculation of the p2p technological similarity score, we only use a subset of all patent applications. Our first inclusion criterion is that we only include patent applications that have been granted. We also limit our sample to patent applications made between 1980 and 2017. We further follow De Rassenfosse et al. (2013) and only include priority filings, whereby we only consider one priority per extended (INPADOC) patent family. Here, we select the earliest priority filing per extended patent family that has already been granted and for which an English language abstract is available. This leads to a final dataset containing roughly 12 million patent applications.

4. Evaluation

In this section, we begin our empirical evaluation of the technological signatures and derived p2p similarity measures generated via the method described above.

4.1. Evaluation strategies

In NLP research generally, the evaluation of methods to assess the similarity of text pairs is a fairly standardized process. In short, such models are usually evaluated on how well they perform on a set of established pre-annotated benchmark "semantic textual similarity" (STS) datasets (e.g., Bowman et al., 2015, who provide 570k of labeled sentence pairs). While useful for benchmarking models for general language, this type of evaluation tends not to perform well for complex and domain-specific language (Chandrasekaran and Mago, 2020) such as the technical descriptions found in patent text. For patents, no STS dataset currently exists.[5] Guided by recent research on the

[2] An example of the data and computing intensity of such an approach is provided by Younge and Kuhn (2016), who produced a patent similarity matrix with 14 trillion entries, using thousands of distributed CPUs for months to do so.

[3] We utilize the efficient *annoy* (Approximate Nearest Neighbor Oh Yeah!, Bernhardsson, 2017). Documentation of the *annoy* package can be found at https://github.com/spotify/annoy. Our implementation constructs a forest of trees (100) using random projections.

[4] The chosen threshold of 0.65 is based on the comparison of patent pairs at a certain similarity threshold, where we decided, based on information from domain experts, that patent-pairs below this threshold still contain enough meaningful relatedness allowing an interpretation.

[5] The only exception is the non-public dataset of patent similarity used by Arts et al. (2018, 2020), who employed technology experts to label a total of 850 patent-pairs with their similarity. Such a number of pairs only hints at the promise of text-based patent indicators, particularly since they applied a rather outdated methodology. However, to provide a proper benchmark for future improvements of such measure and fine-tuning of models, a much larger evaluation dataset is needed. While we are actively engaged in developing such a community-curated patent STS dataset, these efforts go beyond the scope of the present paper

Table 2
Face validity evaluation of technological signature.

Method	Text data	Data Source	N patents	Target	Level (n class)	Precision	Recall	F1
DeepPatent (Li et al., 2018)	title, abstract	USTPO	2.000.147	IPC	subclass (637)	73	n.a.	n.a.
	title, abstract	EPO, WIPO	742.097	IPC	subclass (637)	45	75	55
PatBERT (Lee and Hsiang, 2020)	title, abstract	USTPO	1.950.247	IPC	subclass (637)	80	64	64
	claims	USTPO	1.950.247	CPC	subclass (635)	84	66	66
Our approach	title, abstract	EPO	1.000.000	IPC	subclass (637)	54	53	52

Note: n.a. indicates not reported metrics.

Table 3
Face validity evaluation of similarity.

Evaluation	shared	no shared	*t*-test
IPC class	0.032	0.009	***
IPC subclass	0.049	0.010	***
IPC group	0.071	0.010	***
IPC subgroup	0.108	0.011	***
inventor	0.039	0.006	***
assignee	0.026	0.004	***
citation	0.071	0.001	***
citation XY	0.084	0.001	***
citation examiner	0.112	0.002	***

Note: Two-sided *t*-test. H_1: True difference in means $\neq 0$.

non-STS-based evaluation of the text-based vector representation of patents (e.g., Whalen et al., 2020), we apply three complementary strategies: (i.) a "direct" one, in which we aim to evaluate the created technological signatures of patents; (ii.) a "relational" one, in which we evaluate the derived patent-to-patent similarity.; and, as detailed later in Section 5, an (iii.) "indirect" evaluation, in which we investigate the plausibility and usefulness of the obtained results and further derived indicators in a concrete technological setting.

4.2. Direct evaluation of technological signatures

The "direct" evaluation assumes that a model can be estimated which links each technological signature to the observable attributes of the corresponding patent. In contrast to a pure classification exercise, we are not interested *per se* in maximising the predictive performance of the technological signatures for classification tasks, but rather in their usefulness in measuring technological distance.[6] However, patent technology classifications contain strong signals regarding the technology embodied in a patent; therefore, the created technological signature should enable a better-than-random prediction of the associated IPC classes. This is helpful to assess whether the created vector representation contains meaningful information regarding the technology described in the patent.

In the first step, we examine whether the produced vectors can function as inputs for automated IPC symbol classification on the subclass level for the first-mentioned sub-class. This is a multi-class prediction problem with 637 outcome classes in our sample. Using the constructed embeddings to derive indicators requires them to be reliable and nuanced representations of the underlying patents. In order to capture the interactions and non-linearity between the technological signature and IPC assignments without explicitly modelling them, we deploy an artificial neural network (ANN) with 3 hidden layers, which takes as input the 300-dimensional technological signature vector of a patent and predicts as output the corresponding IPC assignment. Patents can have multiple IPC assignments, making this exercise a multi-class and multi-label problem of predicting all assigned classes. Due to the increased complexity in modelling and evaluation alike, we follow previous research (e.g., Lee and Hsiang, 2020; Li et al., 2018) and, for the first evaluation, only predict the first mentioned rather than all IPC assignments.[7] We trained the ANN on 9,471,069 observations and evaluated it on 10.0000 out-of-sample observations that had not been used to fit the prediction model (Table 2).

The classifier achieved a weighted precision of 54%, a weighted recall of 53%, and an F1 score of 52%, meaning that it was able to detect the right sub-class *out of 637 possible answers* for over half the patents in the test set. As a robustness test, we ran a "placebo type" model by shifting all vectors by one observation (relative to the classes). Training and predicting with that setup rendered accuracy and recall values of 0 for nearly all classes. Overall, we concluded that the information created technological signature vectors enables the reasonable retrieval of assigned IPC classes.

More qualitative evaluations can be carried out by providing technological signals and assessing the extent to which the patents embodying these technologies can be retrieved. To do so, we created a simple application which allows a user to enter a free text search string. The application vectorizes the input query using the same language model that our technological signature vectors are based on, and returns the patents with the most similar technological signature for visual inspection. We granted a set of domain experts access to this application for testing and evaluation over the period of several weeks. All domain experts reported satisfactory performance with respect to the ability of the application to provide them with patents embodying the technologies described in their search queries. Below, we report the results of an example search query.

4.3. Relational evaluation of p2p technological similarity

In the next step, we evaluate the comparability of the generated technological patent signatures, and consequently the quality of the calculated p2p technological similarity measure. "Relational" approaches evaluate the similarity that can be derived between two elements (in this case, the p2p technological similarity). Conceptually, the validity of the former represents a necessary but not sufficient condition for the latter. Due to the lack of a ground-truth benchmark dataset of annotated p2p similarity, we cannot directly validate how accurate our created measures are. However, we can investigate the correlation between our generated p2p similarity and the existing observable measures commonly used to approximate technological similarity. While this exercise as such cannot provide evidence of the advantage of embedding-based p2p similarity measures over other text-based or traditional approaches *per se*, it can serve as a first "sense check" of the face validity and plausibility of our results.

Initially, we compare different samples of patent parts that could intuitively be expected to display on average a higher (or lower) similarity. We rely on the assumption that two patents of the same patent

[6] For embedding-based exercises explicitly aimed at automated patent classification, consider for instance Li et al. (2018), Lee and Hsiang (2020), Kim et al. (2020), Grawe et al. (2017), Risch and Krestel (2019)s.

[7] Many patent authorities (e.g. USTPO) by law require a patent to be assigned to a main class, which must be mentioned first. However, for other authorities without such a legal concept, the order of classes is not binding.

Table 4
Comparison similarity calculation vs Arts et al. (2020).

	Class level	Subclass level	Group level
Arts et al. (2020)	0.3855106	0.2806797	0.1307425
Our approach	0.3639752	0.2442472	0.1007457

Table 5
Comparison similarity calculation.

Original	Most similar
"The invention relates to a lateral guidance control structure with one or more control variables for generating a steering input of a power steering system of a motor vehicle. A device for controlling lateral guidance of a vehicle is described."	"The invention relates to a device for operating a servo steering system which has at least one electric motor for generating a supporting motor steering torque for pivoting at least one steerable wheel of a vehicle."
"The invention relates to a method and a device for detecting a lane change in the context of a vehicle speed control system, such as ACC (Automatic Cruise Control) or a system for distance or collision warning in vehicles, in which a lane change mode is activated as a function of a lane change probability."	"Purpose an automatic driving system for detouring danger area in the automatic driving of a vehicle is provided to control progressive direction of the automatic traveling vehicle by using the reference direction information, 'constitution an automatic driving system for detouring danger area in the automatic driving of a vehicle comprises a road lane recognition part a car speed detecting part a train compartment distance detection part a gps signal receiving part a wireless communication unit a memory unit a driving controller and an automatic driving control unit" "The invention discloses an intelligent lane changing assisting system for an intelligent vehicle and a control method thereof which belongs to the field of automobile active safety."
"The invention relates to a method for environment detection of a vehicle in current driving situations, in which objects are detected and tracked from the environment and the detection and tracking of these objects is performed with an adjustment of the environment sensors executed as a function of the vehicle state."	"Monitoring device for a motor vehicle has a sensor device for detecting obstacles in front of or behind the vehicle and an evaluation unit for checking for the presence of an obstacle within a predefined distance from the vehicle within a monitoring surface." "The device has a vehicle camera provided as an environment sensor and an evaluation unit for determination and output of data based on lights of a vehicle to be automatically controlled."
"The invention relates to a control system for a vehicle with actuators."	"Problem to be solved to provide a control system for controlling a main electronically controlled vehicle system and further controlling at least one additional auxiliary vehicle system." "A vehicular electronic control apparatus includes a vehicle control means and a unit control means."
"The invention relates to a driving style evaluation device. A driving behavior representation parameter estimation unit provides an estimated value of a driving behavior representation parameter representing the driving behavior of the driver of a vehicle."	"The invention relates to a method for operating at least one motor vehicle said method involving the steps of providing s a data record characterizing a vehicle environment a driving behavior and the movement of the motor vehicle ascertaining s at least." "The present invention relates to an arrangement and a method for estimating the speed of a vehicle." "The method includes but is not limited to the steps of evaluating a driver s driving style."

family, developed by the same inventor(s), assigned to the same assignee (s), or citing each other are similar to a certain degree. We further assume that technological similarity should be more pronounced within technological domains, as approximated by technological classifications such as technological fields or IPC or CPC categories. Following the same argument, technological trajectories and the specialization of inventors and assignees should lead to a higher similarity of patents filed by the same inventor or assignee as compared to others. Finally, backward citations refer to relevant prior work; therefore, we assume that a pair in which one patent cites the other should on average display a higher technological similarity than a pair where this is not the case. In all cases, we retrieve all patent pairs where the respective condition is true (i.e. same IPC class, same inventor/assignee, one patent cites the other, etc.), and match it with a random sample of patent-pairs of equal size where this condition is not true. Table 3 reports the results.

On average, patents within the same IPC class display a significantly higher similarity than patents from different classes. Patents sharing an IPC class display an increased magnitude of similarity by a factor of roughly 3, which increases when repeating the same exercise on the subclass (5), group (7) and subgroup (/>9) levels.[8] In conclusion, patents that share at least one IPC subgroup classification are, according to our similarity indicator, almost ten times more similar than patents that do not. Repeating this procedure on the inventor and applicant levels leads to similar results. Likewise, patents filed by the same inventor or assignee are more similar by a factor of roughly 6. All mean differences are significant at the 1% level.

Patent pairs which are connected by a backward citation show, on average, a similarity score 50 times higher than those which are not. However, the average similarity of cited patents is still low, at approx. 7%. Similar results with slightly higher average similarity and higher correlation are obtained only when we limit ourselves to X and Y tag citations (meaning, they are crucial to evaluating the patent's novelty) and citations added by the examiner.[9] While overall reassuring, the outcome is highly skewed, with around 70% of patents citing each other do not display meaningful similarity. Likewise, there are many patent pairs with high similarity scores that do not cite each other, supporting previous findings regarding the bias associated with citation data (e.g. Alcacer and Gittelman, 2006; Bacchiocchi and Montobbio, 2010; Lampe, 2012; Picard and de la Potterie, 2013), and the conclusion that citations may only be a limited indicator of technological similarity.

Finally, we compare the performance of our similarity calculation with Arts et al. (2020). We construct and perform an approximate nearest neighbor search with a sample of over 6 million USPTO patents for which the authors published their similarity measures in a data repository. We then compare the overlaps in IPC assignments for the 10 most similar patents, determining how many of those identified as similar share IPC assignments on the class, subclass, and group levels. Table 4 below presents the share of patents with at least one overlap on each level. Our results are, on all levels, a few percent points lower than those reported by Arts et al. (2020), indicating that both approaches are capturing very similar features. Given the nature of the embeddings (i.e., our approach is tuned to capture synonyms), we suggest – following McNamee (2013) – that in contrast to IPC class based approaches they can detect similarity for patent-pairs with non-overlapping IPC assignment but which are used for similar application.

To exemplify the results of our semantic search queries, we select random patents from the EV field and search for the most similar patents in our patent database. We find that similarity is not always reflected in an overlap of IPC classes, but is often better mirrored in the similarity of technical descriptions. Moreover, a high degree of similarity can sometimes also be found even though different vocabulary is used. Table 5 shows examples of technical descriptions from selected patent abstracts and extractions of descriptions from most similar patents as found by the search query. As illustrated by the results, we obtain

[8] Similar results are obtained when using the CPC classification scheme instead. Sharing multiple classes further increases the similarity score.

[9] Likewise, the Pearson correlation coefficient between shared backward citations and the similarity of a patent pair is low, at 0.05, but statistically significant at the 1% level.

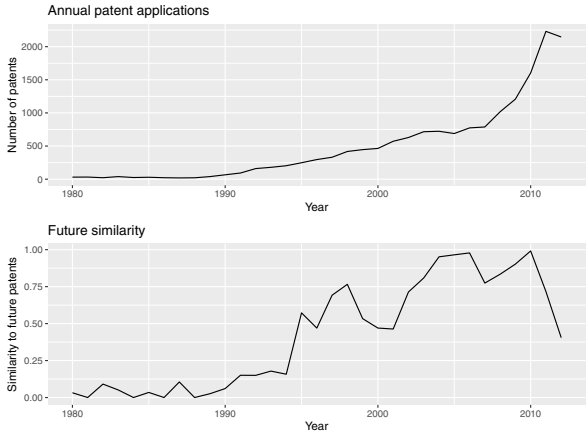

Fig. 2. Overall number and similarity of EV patents.

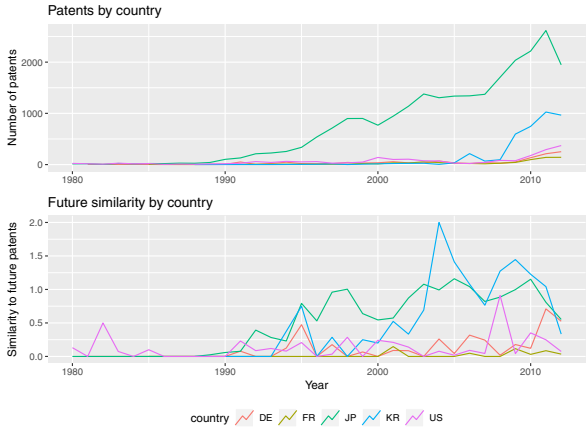

Fig. 3. Novelty & impact on the country level.

matches that resemble the search string in terms of meaning without the need for exact keyword matches. This indicates the capability of the presented method in identifying technologically related patents based on text queries of arbitrary length, and thus its usefulness for applications such as semantic search and patent retrieval.

5. Case study and research applications: Electric vehicles

In particular, we demonstrate the use of a p2p similarity indicator and illustrate the results obtained for two popular research applications: (i.) creating patent quality indicators and (ii.) mapping cross-country knowledge flows. Acknowledging the vast body of research on both topics, we do not claim that these stylized applications advance both lines of research as such. Rather, they aim to provide examples of where and how p2p similarity measures can be used and offer insights about their outcomes within a well-defined technology case.

5.1. Context and data

EV technology is on the cusp of advancing from a niche field into the mass market. In doing so, it will foster a technological regime shift, leaving the internal combustion engine (ICE) behind. Apart from being more environmentally friendly, electric vehicles have several additional advantages: "Electric motors are low-maintenance, versatile and exceptionally quiet (p. 4 Deffke, 2013).

We identify EV technologies using IPC codes on the subclass level,[10] focusing on electric propulsion, a key technology of battery electric vehicles (BEVs). To identify EV-related patents, we follow Pilkington and Dyerson (2006) and select patents found in the IPC class B60L 11/00 and its sub-classes, as these are a "likely home for EV patents" (Pilkington and Dyerson, 2006, p. 85).[11] We include in our sample priority patents within these IPC classes granted between 1980 and 2012, resulting in 22,285 patents.

5.2. Research application 1: Patent quality indicators

It has long been recognized that the technological as well as economic significance of patents varies broadly (Basberg, 1987), and as a result, a large body of literature has explored the rich information contained in patent data to construct patent quality measures.[12] Such measures are traditionally constructed based on (i.) the number or composition of assigned IPC classes (e.g. Lerner, 1994), (ii.) the number and pattern of backward citations (e.g. Harhoff et al., 2003a; Lanjouw and Schankerman, 2001; Schoenmakers and Duysters, 2010) or (iii.) forward citations (e.g. Ahuja and Lampert, 2001; Hall et al., 2005; Harhoff et al., 2003b), or (iii.) the IPC class composition of the citing or cited patents (e.g. Shane, 2001; Trajtenberg et al., 1997; Uzzi et al., 2013), (Trajtenberg et al., 1997). Many suggested measures of different aspects of patent quality are explicitly or implicitly based on their similarity to other patents, particularly those published at earlier or later points in time. Patents with a high similarity to earlier patents are assumed to build on existing knowledge, technologies, and applications, whereas low similarity to earlier work indicates *novelty* (Arts and Veugelers, 2015; Uzzi et al., 2013). Patents with a high similarity to patents published after the focal one indicate the *promisingness* of the embodied technology,as it can be seen to be frequently applied in later innovations. Recently, first attempts at creating text-based patent quality measures leveraging p2p similarity have been made, primarily leveraging simple (Arts et al., 2018) or TFIDF-weighted (Arts et al., 2020; Kelly et al., 2021) co-occurrence. Indeed, text- and similarity-based measures of patent quality appear to correlate well with a large array of ex-post quality measures, such as patent value (Kelly et al., 2021) and the association with prestigious technology awards (Arts et al., 2020).

However, traditional as well as text-based ex-ante quality measures are found to vary substantially with respect to the different post-grant outcomes associated with patent quality, and display significant variation within the same measure across technologies within outcomes (Higham et al., 2021). In short, patent quality is an ongoing field of study, and while text- and similarity-based indicators appear promising, there exists no consensus on how to construct them and based on which particular data source, particularly since they tend to be sensitive to the outcome of interest as well as variations between technologies. Without claiming to provide a superior approach, we suggest that our embedding-based p2p similarity measure can be used to complement and augment existing approaches, since embeddings are less sensitive to the domain-specific technical jargon of particular technology fields. In this section, we thus provide a simple approach to leveraging embedding-based p2p similarity to construct two popular measures of patent quality, namely technological *novelty* (a lack of similarity to

[10] Whereas group and subgroup labels allow even more nuanced identification, they are also less stable over time due to more frequent revision, addition, and reclassification (WIPO, 2017).

[11] A list of all used IPC classes and their description is given in Table A.1 in the appendix. Figure A.1 , A.2 and A.3 provide an additional visualisation of the technological relationships between these IPC classes.

[12] For a somewhat recent and exhaustive review of patent quality measures, consider Squicciarini et al. (2013), and for a more critical reflection on them, see Higham et al. (2021).

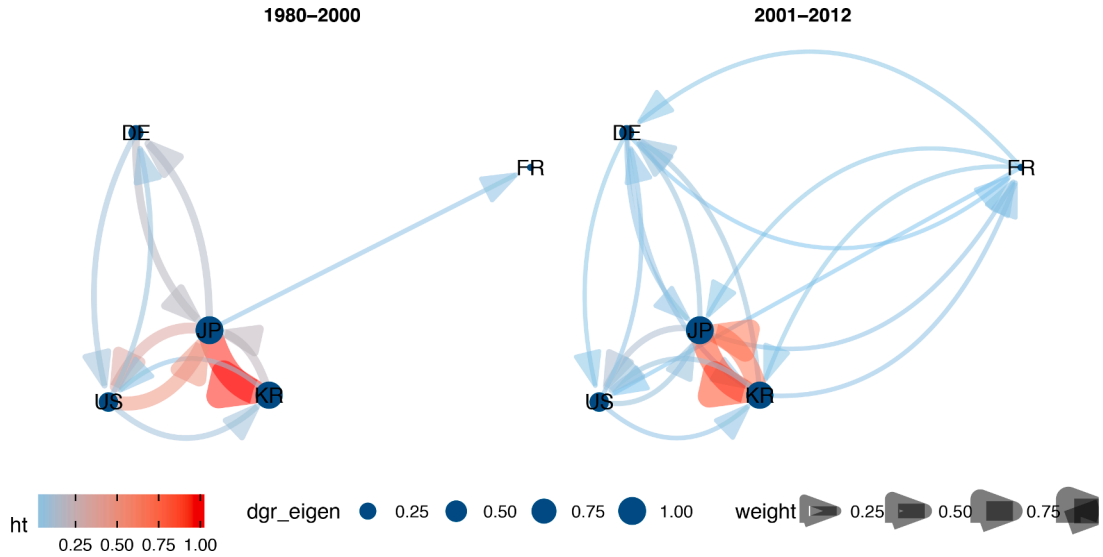

Fig. 4. Knowledge flows between countries. *Note:* Directed network based on $sim_{i,j}^{future}$ on the country level. Node size reflects the in-degree eigenvector centrality edge color and weight represents edge weight.

earlier applications) and *impact* (the similarity to later applications), which can be used as points of departure for future indicator development.

To do so, we first sum all the similarity relationships a patent displays to the universe of other patents, resulting in indicator sim_i. For every patent i, $J_i[1:m]$ will contain patents j with earlier or later application dates. This difference is measured in years with the parameter $\Delta t_{j,i} = t_j - t_i$. With that information, we can construct a temporal similarity index on the patent level, which captures its similarity to other patent applications filed earlier (sim_i^{past}) or later (sim_i^{future}).

$$sim_i = \sum_{j=1}^{m} \frac{s_{i,j}}{m}$$

$$sim_i^{past} = \sum_{j=1}^{m} \frac{(-\tau > \Delta t_{j,i} \geq -\lambda)\ s_{i,j}}{m}$$

$$sim_i^{future} = \sum_{j=1}^{m} \frac{(\tau < \Delta t_{j,i} \leq \lambda)\ s_{i,j}}{m}$$

The resulting indicators represent i's share of similar patents with an application date in the past (sim_i^{past}) or future (sim_i^{future}), weighted by their similarity $s_{i,j}$. $(-\tau > \Delta t_{j,i} \geq -\lambda)$ is a logical condition, leading to the inclusion of a multiplier for $s_{i,j}$ of one or zero, depending on whether the condition is fulfilled or not. To offset the fact that there may be a delay between the patent application and its official publication of up to 12 months (Squicciarini et al., 2013), we introduce a parameter τ that represents the minimum $\Delta t_{j,i}$ for j to be considered in the temporal similarity indicators (we here set $\tau = 1$). In addition, the parameter λ restricts the maximum $\Delta t_{j,i}$ for patent j to be included in the calculation of the temporal similarity indicators. To make this parameter consistent and comparable with the traditionally used 5-year forward citation count as a patent quality indicator (eg. Harhoff et al., 2003a; Squicciarini et al., 2013), we set $\lambda = 5$.

Using the case of EV technologies, we now illustrate and discuss the obtained results.[13] For a first overview of the sector and technology, Fig. 2 displays the development of the number of EV patent applications as well as their average sim_n^{future} over time.

While we see marginal activity in patent applications in the 1980s, there is steady growth beginning in the 1990s, with a sharp increase in the mid-2000s. sim_n^{future}, however, follows a different trajectory. Until the mid-1990s, almost no patent showed similarity to future patents, indicating the generally low patenting activity but also the non-cumulative and fragmented nature of technology development in this period. However, in the mid-1990s, we witness a sudden peak of sim_n^{future}, followed by further peaks in the mid-2000s and early 2010s, hinting at an emerging technology life-cycle. Here, the first main peak around the year 1997 coincides with the development of the Toyota Prius, which became the first mass-produced hybrid-electric vehicle and a forerunner in the field of (hybrid) EV technology. The later peaks fall within a period of growing patenting activity in the energy storage field in general, but with a steadily rising focus on lithium-ion technologies in particular beginning in 2005 (Dinger et al., 2010). This technology played an important role in EV development over the next decade, which explains the high future similarity during this time. This figure also illustrates the forward-looking nature of sim_n^{future}, where new technological trends and developments are traceable before the corresponding technology starts to gain popularity. Consequently, we suggest that on different levels of aggregation, sim_n^{future} can be interpreted as an indicator of the "impact" of a certain technology.

We can also provide an overview of EV patenting and our similarity-based indicators on the country level.[14] 3 illustrates the technological development of the five countries accounting for the highest number of EV patent applications, namely Japan, South Korea, the United States, Germany, and France.

Based on these trends, Japan can clearly be identified as the leading country in the field of core EV patents, showing a sharp increase in output since the 1990s as the general forerunner in EV technologies. This is in accordance with the development of Japan's vehicle industry, which was the first to introduce vehicles with alternative powertrains

[13] Further examples of similar applications can be found in Hain et al. (2020), where we utilize p2p to measure technological catching-up efforts on the country level.

[14] PATSTAT data is known to have difficulty capturing correct and complete inventor addresses (approx. 30% of patents cannot be clearly assigned to any geographical location), a problem which is particularly amplified in Asian countries. Therefore, for this research, we leverage recent efforts by De Rassenfosse et al. (2019) to provide more comprehensive geo-information for PATSTAT data, covering more than 90% of global patenting activity. Since most patents have multiple inventors listed, we assign each geolocation a fractionalised number representing the share of inventors of a particular patent at that particular location. We choose the inventor information instead of the more commonly used applicant information to assign patents to countries in order to capture the location of inventive activity rather than the location of the intellectual property rights ownership (Squicciarini et al., 2013).

and was also strongly supported by governmental programs at an early stage of this transition (Åhman, 2006). This first position remained unchallenged for the whole period considered. However, in the mid-2000s, Korean EV research started to take off and subsequently increased its patent output rapidly. This upswing is clearly in line with the 2004 founding of the Pangyo Techno Valley (PTV), a large research cluster accumulating eight of the top ten Korean tech companies and more than 1300 IT-companies, as well as the introduction of the Korean "Innocity" policy in 2007 to establish new innovation cities (Lee et al., 2017). The United States, Germany, and France, in the meantime, showed only negligible activity and only became somewhat significant around 2010. This possibly results from the comparatively late introduction of EV innovation policies in the United States in 2009 (Gu and Shao, 2014) and the PPP Green Car Initiative of the European Commission starting in 2008. Overall, patenting in EV technologies appears to be rather concentrated, with Japan accounting for 41% of all patents filed, and the leading five countries together accounting for 89% of all patents. In terms of the development of the sim_n^{future} indicator, we see a slightly different picture compared to the total patent count. The development of Japan's sim_n^{future} roughly follows its number of patent applications and shows a somewhat stable trend of high future similarity. The huge impact of the Japanese patent count can also be seen in the resemblance of the Japanese course to the overall similarity in Fig. 3. However, we also spot several peaks for countries with simultaneously minimal patenting activity, yet the development of promising technologies. Particularly noticeable is the peak of South Korea in the mid-2000s, when the average sim_n^{future} of Korean patents overtakes Japan's lead.[15] Overall, the high average future similarity of Korean patents in the following years with a rising patent count suggests highly innovative and future-driven patenting behavior.

5.3. Research application 2: Mapping knowledge flows

The introduced temporal p2p similarity indicator naturally lends itself to a direct network analysis on different levels of aggregation. As an example, we create a directed network between the top-patenting countries based on aggregated $sim_{i,j}^{future}$. Since the similarity of patent applications in country i with patent applications in country j at a later point in time can be interpreted as a function of knowledge spillover, the resulting network illustrates technology-related knowledge flows between countries (Fig. 4).

We see that during the formative period of EV technologies until 2000, strong knowledge flows existed particularly from Japan to South Korea, but there were also bidirectional ones between the US and Japan. The network underscores that Japan can be seen as the central player during this period, building the knowledge base for the future development of the other top 4 patenting countries. However, it also becomes apparent that some Japanese developments in turn are based on US developments in the early 1980s, mainly a patent introducing $LiCoO_2$ as a new cathode material for lithium batteries (Godshall et al., 1982). The strongest knowledge flow for the first period is observable from Japan to Korea, accompanying the dissemination of knowledge in other technological fields. Former studies showed that for several high technology areas like flat panel displays (Hu, 2008; Jang et al., 2009) and mobile telephones (Lee and Jin, 2012), the knowledge source/patent citation often follows the order of industry entry lead by Japan, followed by the US and Korea (Han and Niosi, 2018). Further reasons for the strong connection might also be seen in the high resource-based dependence on Japan; for example, LG Chem, the largest Korean EV battery producer, is heavily reliant on Japanese materials.

Along with the overall higher connectivity of the knowledge flow network post-2000, the characteristics of the knowledge flow within the network also changed. With the apparent increase in interconnectedness between the countries, we now see strong bidirectional connections between Japan and South Korea, indicating mutually reinforcing knowledge flows. Conversely, knowledge flows between the US and Japan now mainly originate from Japan.

6. Conclusion

In this paper, we propose an efficient and scalable approach to creating vector representations of a patent's technological signature based on the textual information found in its abstract, utilizing embedding techniques from NLP. We leverage these technological signatures to derive p2p technological similarity measures. We suggest and demonstrate the use of approximate nearest neighbor matching to create similarity measures for large datasets, allowing us to represent the whole universe of patents as a similarity network and thereby opening the possibility for a large range of applications and analyses. We evaluate the properties of our embedding-based p2p similarity indicator in various ways, illustrate the obtained results, and suggest potential research applications using the case of electric vehicle technologies.

While the results so far demonstrate the usefulness of a semantic indicator of p2p technological similarity and offer strong initial insight into the possible applications, the full potential remains somewhat unexplored. We therefore offer some suggestions to improve the accuracy of the technological signature and the derived p2p similarity measures, to validate its outcomes, and apply it to a range of suitable problems.

First, the present approach utilizes patent abstracts as a data source, relying on them being rather standardised summaries of the technologies described by the patents. This allows us to apply existing NLP approaches without violating the underlying assumptions of the computational models used about the comparable length of texts across a corpus or the non-ambiguity of text fragments. Future research should, however – following current developments in language processing technologies – explore inputs beyond abstracts, utilizing patent claims and eventually full texts. To do so, an increased understanding is required of which sources of textual data contain extractable information on the technology, application, novelty, or legal protection, and how different data sources can be combined to provide more holistic representations of a patent's technology.

With respect to the validity, information content, and use of patent technological signatures, several avenues for future work exist. Based on our evaluation, we are confident that the vectors represent the underlying patents' technological features since they enable the prediction of a patent's technology class. Generally, the validation and verification of the proposed measure of technological similarity between patents is limited to the reproduction of stylized facts and the comparison with existing measures. While first attempts to utilize domain expert knowledge to validate and optimize technological similarity metrics have been made (Arts et al., 2018), the creation of a large-scale expert annotated dataset could create an objective benchmark, facilitating technology forecasting research. Guidance here can be drawn from the large pre-annotated "semantic textual similarity" (STS) datasets frequently used in NLP research.

Finally, we suggest the broader application of embedding techniques to map and forecast technological change as a promising avenue for future research. The here presented methods do not require a specific data structure or property such bibliographic references between documents or the presence of a predefined classification. Therefore, it can be deployed to map similarity within and across a broad host of data sources which include textual descriptions of technologies, such as

[15] However, this peak is mainly caused by the big differences in patent count among countries at this time and the graph being based on country averages, as the most promising patent from Japan still ranks three times higher in future similarity than the best Korean one.

academic publications (e.g. Franceschini et al., 2016), research grant descriptions (e.g. Freyman et al., 2016), product manuals (e.g. Jeong et al., 2021), and various types of web data such as technology blogs (e.g. Jurowetzki and Hain, 2014) or firm websites (e.g. Kinne and Axenbeck, 2020).

Declaration of Competing Interest

The authors declare no conflict of interest.

Appendix

Figure A.3

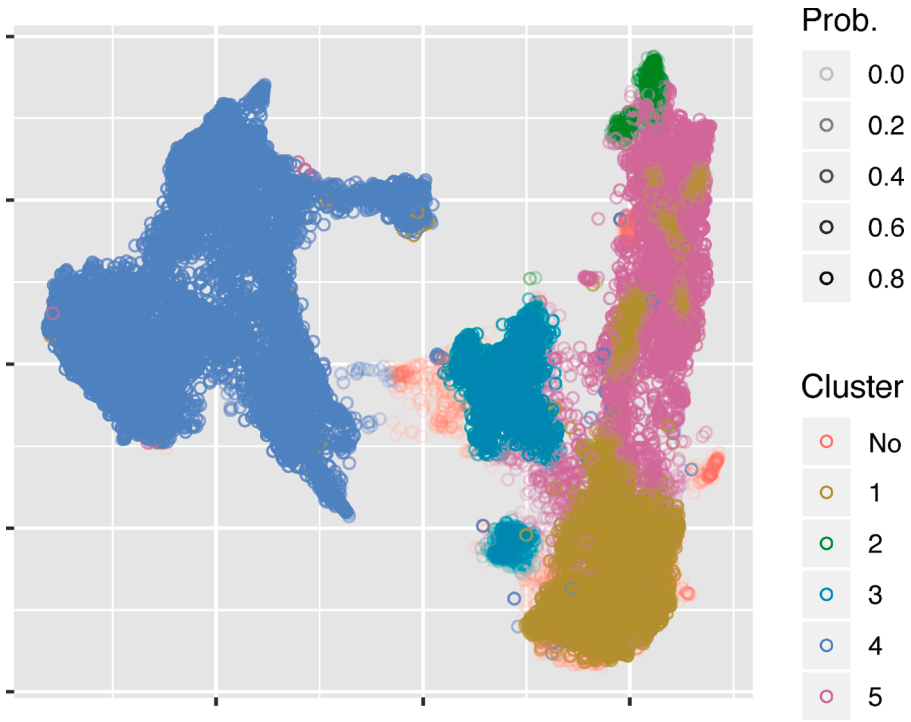

Fig. A.1. UMAP projection of patent vectors. *Note:* UMAP dimensionality reduction (McInnes et al., 2018) of EV patent signatures in 2-dimensional space. Colors indicate the outcome of a density-based clustering (HDBSCAN).

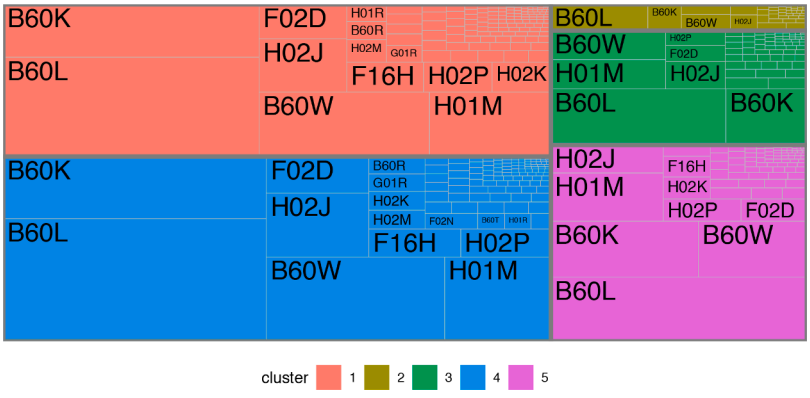

Fig. A.2. IPC class composition of technology clusters. *Note:* Illustration of IPC composition of identified clusters in EV patents.

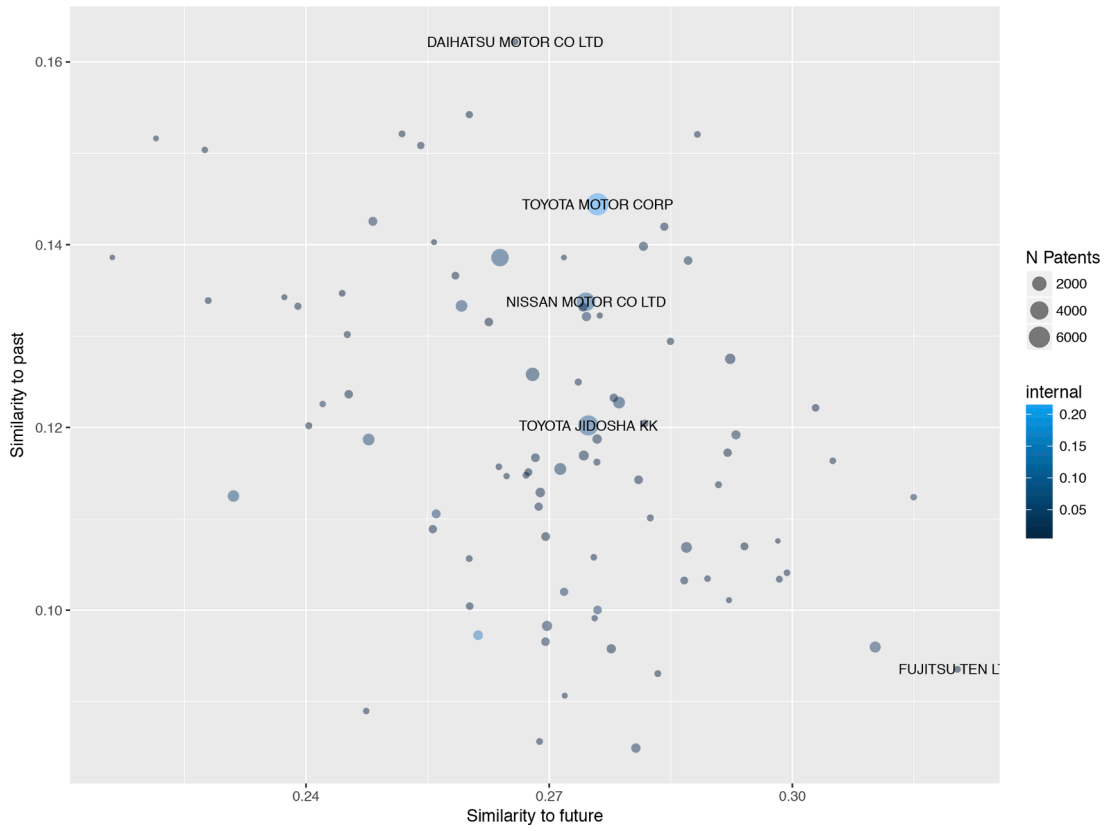

Fig. A.3. Novelty & impact on the firm level.

Table A.1
List of used IPC-classes.

IPC class	Level	Description
B60L 11/00	Subgroup	Electric propulsion with power supplied within the vehicle
B60L 11/02	Subgroup	Using engine-driven generators
B60L 11/04	Subgroup	Using dc generators and motors
B60L 11/06	Subgroup	Using ac generators and dc motors
B60L 11/08	Subgroup	Using ac generators and motors
B60L 11/10	Subgroup	Using dc generators and ac motors
B60L 11/12	Subgroup	With additional electric power supply, e.g. accumulator
B60L 11/14	Subgroup	With provision for direct mechanical propulsion
B60L 11/16	Subgroup	Using power stored mechanically, e.g. in flywheel
B60L 11/18	Subgroup	Using power supplied from primary cells, secondary cells, or fuel cells

References

Adams, S., 2001. Comparing the IPC and the US classification systems for the patent searcher. World Patent Inf. 23 (1), 15–23.

Aharonson, B.S., Schilling, M.A., 2016. Mapping the technological landscape: measuring technology distance, technological footprints, and technology evolution. Res. Policy 45 (1), 81–96.

Åhman, M., 2006. Government policy and the development of electric vehicles in Japan. Energy Policy 34 (4), 433–443.

Ahuja, G., Lampert, C., 2001. Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions. Strateg. Manage. J. 22 (6–7), 521–543.

Alcacer, J., Gittelman, M., 2006. Patent citations as a measure of knowledge flows: the influence of examiner citations. Rev. Econ. Stat. 88 (4), 774–779.

Alcácer, J., Gittelman, M., Sampat, B., 2009. Applicant and examiner citations in us patents: an overview and analysis. Res. policy 38 (2), 415–427.

Alstott, J., Triulzi, G., Yan, B., Luo, J., 2017. Mapping technology space by normalizing patent networks. Scientometrics 110 (1), 443–479.

Archibugi, D., Planta, M., 1996. Measuring technological change through patents and innovation surveys. Technovation 16 (9), 451–519.

Arts, S., Cassiman, B., Gomez, J.C., 2018. Text matching to measure patent similarity. Strateg. Manage. J. 39 (1), 62–84.

Arts, S., Hou, J., Gomez, J.C., 2020. Natural language processing to identify the creation and impact of new technologies in patent text: code, data, and new measures. Res. Policy 104144.

Arts, S., Veugelers, R., 2015. Technology familiarity, recombinant novelty, and breakthrough invention. Ind. Corp. Change 24 (6), 1215–1246.

Bacchiocchi, E., Montobbio, F., 2010. International knowledge diffusion and home-bias effect: do USPTO and EPO patent citations tell the same story? Scand. J. Econ. 112 (3), 441–470.

Barirani, A., Agard, B., Beaudry, C., 2013. Discovering and assessing fields of expertise in nanomedicine: a patent co-citation network perspective. Scientometrics 94 (3), 1111–1136.

Basberg, B.L., 1987. Patents and the measurement of technological change: a survey of the literature. Res. Policy 16 (2–4), 131–141.

Beall, J., Kafadar, K., 2008. Measuring the extent of the synonym problem in full-text searching. Evid. Based Libr. Inf. Pract. 3 (4), 18–33.

Bekamiri, H., Hain, D. S., Jurowetzki, R., 2021. PatentSBERTA: a deep NLP based hybrid model for patent distance and classification using augmented SBERT. arXiv preprint arXiv:2103.11933.

Benner, M., Waldfogel, J., 2008. Close to you? Bias and precision in patent-based measures of technological proximity. Res. Policy 37 (9), 1556–1567.

Bernhardsson, E., 2017. Annoy: Approximate nearest neighbors in C++/Python optimized for memory usage and loading/saving to disk. https://github.com/spotify/annoy.

Bowman, S. R., Angeli, G., Potts, C., Manning, C. D., 2015. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.

Boyack, K.W., Klavans, R., 2008. Measuring science–technology interaction using rare inventor–author names. J. Inf. 2 (3), 173–182.

Breschi, S., Lissoni, F., Malerba, F., 2003. Knowledge-relatedness in firm technological diversification. Res. Policy 32 (1), 69–87.

Cetintas, S., Si, L., 2012. Effective query generation and postprocessing strategies for prior art patent search. J. Am. Soc. Inf.Sci. Technol. 63 (3), 512–527.

Chandrasekaran, D., Mago, V., 2020. Domain specific complex sentence (DCSC) semantic similarity dataset. arXiv preprint arXiv:2010.12637.

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., Yang, G., 2020. A deep learning based method for extracting semantic information from patent documents. Scientometrics 1–24.
Cotropia, C.A., Lemley, M.A., Sampat, B., 2013. Do applicant patent citations matter? Res. Policy 42 (4), 844–854.
Criscuolo, P., Verspagen, B., 2008. Does it matter where patent citations come from? Inventor vs. examiner citations in european patents. Res. Policy 37 (10), 1892–1908.
De Rassenfosse, G., Dernis, H., Guellec, D., Picci, L., de la Potterie, B.v.P., 2013. The worldwide count of priority patents: a new indicator of inventive activity. Res. Policy 42 (3), 720–737.
De Rassenfosse, G., Kozak, J., Seliger, F., 2019. Geocoding of worldwide patent data. Nat. Sci. Data 6 (1), 1–15.
Deffke, U., 2013. Electric mobility - rethinking the car. Federal Ministry of Education and Research (BMBF), Department for Electronic Systems and Electric Mobility. Web Page. http://www.bmbf.de/pub/electric_mobility_rethinking_the_car.pdf.
Dinger, A., Martin, R., Mosquet, X., Rabl, M., Rizoulis, D., Russo, M., Sticher, G., 2010. Batteries for electric cars: challenges, opportunities, and the outlook to 2020. Boston Consult. Group 7, 2017.
Engelsman, E.C., van Raan, A.F., 1994. A patent-based cartography of technology. Res. Policy 23 (1), 1–26.
Ernst, H., 2001. Patent applications and subsequent changes of performance: evidence from time-series cross-section analyses on the firm level. Res. Policy 30 (1), 143–157.
Fall, C.J., Törcsvári, A., Benzineb, K., Karetka, G., 2003. Automated categorization in the international patent classification. Acm Sigir Forum. ACM, pp. 10–25.
Firth, J.R., 1957. A synopsis of linguistic theory, 1930–1955. Studies in Linguistic Analysis.
Franceschini, S., Faria, L.G., Jurowetzki, R., 2016. Unveiling scientific communities about sustainability and innovation. a bibliometric journey around sustainable terms. J. Clean. Prod. 127, 72–83.
Freyman, C.A., Byrnes, J.J., Alexander, J., 2016. Machine-learning-based classification of research grant award records. Res. Eval. 25 (4), 442–450.
Fu, X., Yang, Q.G., 2009. Exploring the cross-country gap in patenting: astochastic frontier approach. Res. Policy 38 (7), 1203–1213.
Garfield, E., 1966. Patent citation indexing and the notions of novelty, similarity, and relevance. J. Chem. Doc. 6 (2), 63–65.
Gerken, J.M., Moehrle, M.G., 2012. A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. Scientometrics 91 (3), 645–670.
Godshall, N., Huggins, R., Raistrick, I., 1982. Ternary Compound Electrode for Lithium Cells. Technical Report.
Grawe, M.F., Martins, C.A., Bonfante, A.G., 2017. Automated patent classification using word embedding. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 408–411.
Griffith, R., Lee, S., Van Reenen, J., 2011. Is distance dying at last? Falling home bias in fixed-effects models of patent citations. Quant. Econ. 2 (2), 211–249.
Griliches, Z., 1990. Patent statistics as economic indicators: a survey. J. Econ. Lit. 28 (4), 1661–1707.
Gu, L., Shao, Y., 2014. The analysis of innovation policies for new energy vehicle technology. Stud. Sociol. Sci. 5 (3), 133–137.
Hagedoorn, J., Cloodt, M., 2003. Measuring innovative performance: is there an advantage in using multiple indicators? Res. Policy 32 (8), 1365–1379.
Hain, D., Buchmann, T., Kudic, M., Müller, M., 2018. Endogenous dynamics of innovation networks in the german automotive industry: analysing structural network evolution using a stochastic actor-oriented approach. Int. J. Comput. Econ. Econom. 8 (3–4), 325–344.
Hain, D.S., Jurowetzki, R., Konda, P., Oehler, L., 2020. From catching up to industrial leadership: towards an integrated market-technology perspective. An application of semantic patent-to-patent similarity in the wind and EV sector. Ind. Corp. Change 29 (5), 1233–1255.
Hain, D.S., Jurowetzki, R., Zhou, Y., Lee, S., 2021. Introduction to the special issue: machine learning and AI for science, technology, and (eco-)system mapping and forecasting. Scientometrics.(forthcoming)
Hall, B.H., Jaffe, A., Trajtenberg, M., 2005. Market value and patent citations. RAND J. Econ. 36 (1), 16–38.
Han, X., Niosi, J., 2018. The Revolution in Energy Technology: Innovation and the Economics of the Solar Photovoltaic Industry. Edward Elgar Publishing.
Harhoff, D., Scherer, F.M., Vopel, K., 2003. Citations, family size, opposition and the value of patent rights. Res. Policy 32 (8), 1343–1363.
Harhoff, D., Scherer, F.M., Vopel, K., 2003. Citations, family size, opposition and the value of patent rights. Res. Policy 32 (8), 1343–1363.
Hayes, P.J., Weinstein, S.P., 1990. CONSTRUE/TIS: a system for content-based indexing of a database of news stories. IAAI, vol. 90, pp. 49–64.
Higham, K., De Rassenfosse, G., Jaffe, A.B., 2021. Patent quality: towards a systematic framework for analysis and measurement. Res. Policy 50 (4), 104215.
Hu, M.-C., 2008. Knowledge flows and innovation capability: the patenting trajectory of Taiwan's thin film transistor-liquid crystal display industry. Technol. Forecast. Social Change 75 (9), 1423–1438.
Huang, M.-H., Chiang, L.-Y., Chen, D.-Z., 2003. Constructing a patent citation map using bibliographic coupling: a study of Taiwan's high-tech companies. Scientometrics 58 (3), 489–506.
Jaffe, A.B., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. Q. J. Econ. 108 (3), 577–598.
Jang, S.-L., Lo, S., Chang, W.H., 2009. How do latecomers catch up with forerunners? Analysis of patents and patent citations in the field of flat panel display technologies. Scientometrics 79 (3), 563–591.
Jeong, Y., Jang, H., Yoon, B., 2021. Developing a risk-adaptive technology roadmap using a Bayesian network and topic modeling under deep uncertainty. Scientometrics 126 (5), 3697–3722.
Jurowetzki, R., Hain, D.S., 2014. Mapping the (r-) evolution of technological fields–a semantic network approach. International Conference on Social Informatics. Springer, pp. 359–383.
Kay, L., Newman, N., Youtie, J., Porter, A.L., Rafols, I., 2014. Patent overlay mapping: visualizing technological distance. J. Assoc. Inf. Sci. Technol. 65 (12), 2432–2443.
Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 2018. Measuring Technological Innovation over the Long Run. Technical Report. National Bureau of Economic Research.
Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 2021. Measuring technological innovation over the long run. Am. Econ. Rev. 3 (3), 303–320.
Kim, J., Yoon, J., Park, E., Choi, S., 2020. Patent document clustering with deep embeddings. Scientometrics 1–15.
Kinne, J., Axenbeck, J., 2020. Web mining for innovation ecosystem mapping: a framework and a large-scale pilot study. Scientometrics 125 (3), 2011–2041.
Kogler, D.F., Rigby, D.L., Tucker, I., 2013. Mapping knowledge space and technological relatedness in us cities. Eur. Plann. Stud. 21 (9), 1374–1391.
Lampe, R., 2012. Strategic citation. Rev. Econ. Stat. 94 (1), 320–333.
Lanjouw, J.O., Schankerman, M., 2001. Characteristics of patent litigation: a window on competition. RAND J. Econ. 129–151.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Lee, J.-S., Hsiang, J., 2020. Patent classification by fine-tuning bert language model. World Patent Inf. 61, 101965.
Lee, K., Jin, J., 2012. From learning knowledge outside to creating knowledge within: Korea's mobile phone industry compared with those of Japan, Taiwan and China. Korean Science and Technology in an International Perspective. Springer, pp. 197–218.
Lee, S., Yoon, B., Park, Y., 2009. An approach to discovering new technology opportunities: keyword-based patent map approach. Technovation 29 (6–7), 481–497.
Lee, S.Y., Noh, M., Seul, J.Y., 2017. Government-led regional innovation: a case of 'Pangyo' IT cluster of South Korea. Eur. Plann. Stud. 25 (5), 848–866.
Lemley, M.A., Sampat, B., 2012. Examiner characteristics and patent office outcomes. Rev. Econ. Stat. 94 (3), 817–827.
Lerner, J., 1994. The importance of patent scope: an empirical analysis. RAND J. Econ. 319–333.
Leydesdorff, L., 2008. Patent classifications as indicators of intellectual organization. J. Am. Soc. Inf.Sci. Technol. 59 (10), 1582–1597.
Li, G., 2018. A literature review on patent texts analysis techniques. Int. J. Knowl. Lang. Process. 9 (3), 1–15.
Li, S., Hu, J., Cui, Y., Hu, J., 2018. DeepPatent: patent classification with convolutional neural networks and word embedding. Scientometrics 117 (2), 721–744.
Li, X., Wang, C., Zhang, X., Sun, W., 2020. Generic SAO similarity measure via extended Sørensen-Dice index. IEEE Access 8, 66538–66552.
Li, Y.A., 2014. Borders and distance in knowledge spillovers: dying over time or dying with age?–Evidence from patent citations. Eur. Econ. Rev. 71, 152–172.
Marco, A.C., Sarnoff, J.D., Charles, A., 2019. Patent claims and patent scope. Res. Policy 48 (9), 103790.
McInnes, L., Healy, J., Melville, J., 2018. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.
McNamee, R.C., 2013. Can't see the forest for the leaves: similarity and distance measures for hierarchical taxonomies with a patent classification example. Res. Policy 42 (4), 855–873.
Meguro, K., Osabe, Y., 2019. Lost in patent classification. World Patent Inf. 57, 70–76.
Michel, J., Bettels, B., 2001. Patent citation analysis. A closer look at the basic input data from patent search reports. Scientometrics 51 (1), 185–201.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, pp. 3111–3119.
Moeller, A., Moehrle, M.G., 2015. Completing keyword patent search with semantic patent search: introducing a semiautomatic iterative method for patent near search based on semantic similarities. Scientometrics 102 (1), 77–96.
Mowery, D.C., Oxley, J.E., Silverman, B.S., 1998. Technological overlap and interfirm cooperation: implications for the resource-based view of the firm. Res. Policy 27 (5), 507–523.
Newman, M.A., 1998. Method for Syntactic and Semantic Analysis of Patent Text and Drawings. US Patent 5,774,833.
Noh, H., Jo, Y., Lee, S., 2015. Keyword selection and processing strategy for applying text mining to patent analysis. Expert Syst. Appl. 42 (9), 4348–4360.
Pavitt, K., 1984. Sectoral patterns of technical change: towards a taxonomy and a theory. Res. Policy 13 (6), 343–373.
Pavitt, K., 1985. Patent statistics as indicators of innovative activities: possibilities and problems. Scientometrics 7 (1), 77–99.
Pavitt, K., 1988. Uses and abuses of patent statistics. Handbook of Quantitative Studies of Science and Technology. Elsevier, pp. 509–536.
Pennington, J., Socher, R., Manning, C.D., 2014. GloVe: global vectors for word representation. Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.http://www.aclweb.org/anthology/D14-1162
Picard, P.M., de la Potterie, B.v.P., 2013. Patent office governance and patent examination quality. J. Public Econ. 104, 14–25.
Pilkington, A., Dyerson, R., 2006. Innovation in disruptive regulatory environments: apatent study of electric vehicle technology development. Eur. J. Innov. Manage. 9 (1), 79–91.

Preschitschek, N., Niemann, H., Leker, J., Moehrle, M.G., 2013. Anticipating industry convergence: semantic analyses vs IPC co-classification analyses of patents. Foresight.
Qi, J., Lei, L., Zheng, K., Wang, X., 2020. Patent analytic citation-based VSM: challenges and applications. IEEE Access 8, 17464–17476. https://doi.org/10.1109/ACCESS.2020.2967817.
Righi, C., Simcoe, T., 2019. Patent examiner specialization. Res. Policy 48 (1), 137–148.
Risch, J., Krestel, R., 2019. Domain-specific word embeddings for patent classification. Data Technol. Appl.
Rodriguez, A., Kim, B., Turkoz, M., Lee, J.-M., Coh, B.-Y., Jeong, M.K., 2015. New multi-stage similarity measure for calculation of pairwise patent similarity in a patent citation network. Scientometrics 103 (2), 565–581.
Rothaermel, F.T., Boeker, W., 2008. Old technology meets new technology: complementarities, similarities, and alliance formation. Strateg. Manage. J. 29 (1), 47–77.
Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24 (5), 513–523.
San Kim, T., Sohn, S.Y., 2020. Machine-learning-based deep semantic analysis approach for forecasting new technology convergence. Technol. Forecasti. Social Change 157, 120095.
Schmookler, J., 1966. Invention and Economic Growth. Harvard Univ. Press, Cambridge, MA.
Schoenmakers, W., Duysters, G., 2010. The technological origins of radical inventions. Res. Policy 39 (8), 1051–1059.
Shane, S., 2001. Technological opportunities and new firm creation. Manage. Sci. 47 (2), 205–220.
Singh, J., Marx, M., 2013. Geographic constraints on knowledge spillovers: political borders vs. spatial proximity. Manage. Sci. 59 (9), 2056–2078.
Soo, V.-W., Lin, S.-Y., Yang, S.-Y., Lin, S.-N., Cheng, S.-L., 2006. A cooperative multi-agent platform for invention based on patent document analysis and ontology. Expert Syst. Appl. 31 (4), 766–775.
Squicciarini, M., Dernis, H., C, C., 2013. Measuring Patent Quality: Indicators of Technological and Economic Value. Technical Report, Working Papers, 2013/03. OECD Science, Technology and Industry, OECD Publishing.
Sternitzke, C., Bergmann, I., 2009. Similarity measures for document mapping: acomparative study on the level of an individual scientist. Scientometrics 78 (1), 113–130.
Suh, Y., 2017. Exploring convergence fields of safety technology using arm-based patent co-classification analysis. J. Korean Soc. Saf. 32 (5), 88–95.
Taduri, S., Lau, G.T., Law, K.H., Kesan, J.P., 2011. Retrieval of patent documents from heterogeneous sources using ontologies and similarity analysis. 2011 IEEE Fifth International Conference on Semantic Computing. IEEE, pp. 538–545.
Thompson, P., Fox-Kean, M., 2005. Patent citations and the geography of knowledge spillovers: a reassessment. Am. Econ. Rev. 95 (1), 450–460.
Tong, X., Davidson, F., 1994. Measuring national technological performance with patent claims data. Res. Policy 23 (2), 133–141.
Trajtenberg, M., Henderson, R., Jaffe, A., 1997. University versus corporate patents: a window on the basicness of invention. Econ. Innov. New Technol. 5 (1), 19–50.
Tran, T., Kavuluru, R., 2017. Supervised approaches to assign cooperative patent classification (CPC) codes to patents. International Conference on Mining Intelligence and Knowledge Exploration. Springer, pp. 22–34.
Tseng, Y.-H., Lin, C.-J., Lin, Y.-I., 2007. Text mining techniques for patent analysis. Inf. Process. Manage. 43 (5), 1216–1247.
Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G., Jain, A., 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature 571 (7763), 95–98.
Uzzi, B., Mukherjee, S., Stringer, M., Jones, B., 2013. Atypical combinations and scientific impact. Science 342 (6157), 468–472.
Von Wartburg, I., Teichert, T., Rost, K., 2005. Inventive progress measured by multi-stage patent citation analysis. Res. Policy 34 (10), 1591–1607.
Wang, X., Ren, H., Chen, Y., Liu, Y., Qiao, Y., Huang, Y., 2019. Measuring patent similarity with sao semantic analysis. Scientometrics 121 (1), 1–23.
Whalen, R., Lungeanu, A., DeChurch, L., Contractor, N., 2020. Patent similarity data and innovation metrics. J. Empir. Legal Stud. 17 (3), 615–639.
WIPO, 2017. Guide to the International Patent Classification.
Wolter, B., 2012. It takes all kinds to make a world–some thoughts on the use of classification in patent searching. World Patent Inf. 34 (1), 8–18.
Wu, H.-C., Chen, H.-Y., Lee, K.-Y., Liu, Y.-C., 2010. A method for assessing patent similarity using direct and indirect citation links. 2010 IEEE International Conference on Industrial Engineering and Engineering Management. IEEE, pp. 149–152.
Yan, B., Luo, J., 2017. Measuring technological distance for patent mapping. J. Assoc. Inf. Sci. Technol. 68 (2), 423–437.
Yang, C., Zhu, D., Wang, X., Zhang, Y., Zhang, G., Lu, J., 2017. Requirement-oriented core technological components' identification based on sao analysis. Scientometrics 112 (3), 1229–1248.
Yoon, B., 2008. On the development of a technology intelligence tool for identifying technology opportunity. Expert Syst. Appl. 35 (1–2), 124–135.
Younge, K.A., Kuhn, J.M., 2016. Patent-to-Patent Similarity: A Vector Space Model. Available at SSRN 2709238.
Yufeng, D., Duo, J., Lixue, J., Guiping, Z., 2016. Patent similarity measure based on sao structure. J. Chin. Inf. Process. 30 (1), 30–35.
Zhang, Y., Shang, L., Huang, L., Porter, A.L., Zhang, G., Lu, J., Zhu, D., 2016. A hybrid similarity measure method for patent portfolio analysis. J. Inf. 10 (4), 1108–1130.
Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., Zhang, L., 2020. Forecasting emerging technologies using data augmentation and deep learning. Scientometrics 123 (1), 1–29.

Daniel Stefan Hain, PhD Daniel Hain is an Associate Professor in Data Science, Applied Econometrics, and Innovation Studies. After an initial career in industrial engineering, he dedicated his PhD thesis (Economics) and subsequent work the economics of technological change, particularly the application and development of computational methods to map, visualize, and analyze its causes and consequences. These methods range from traditional econometric modeling to supervised machine learning and predictive modeling, network and complex system modeling, bibliometrics, and agent-based simulation. Daniel is Co-Founder and main instructor at the Social Data Science program (2018) at AAU and involved in a range of initiatives to promote the application of machine learning and artificial intelligence methods in social science.

Roman Jurowetzki, Assistant Professor, PhD Roman Jurowetzki is an Assistant Professor in Innovation Studies and Data Science. He commenced working with mapping of complex technological systems using big and unstructured data as a PhD student in 2012, exploring the development of the Danish Smart Grid system. Over the years, Roman has developed an expertise in combining natural language processing with other computational and statistical methods for application in mainly innovation studies projects. He has extensive project management and policy engagement experience through his work at GLOBELICS, a global network for innovation and development, in the years 2012–2016. Since 2017, he is affiliated with the Sino-Danish-Center (SDC), Beijing where he teaches and engages in collaborative research on Machine Learning and Artificial Intelligence technologies and their application in the Chinese context. Roman is Co- Founder and main instructor at the Social Data Science programme (2018) at AAU.