

Truly standard-essential patents? A semantics-based analysis

Lorenz Brachtendorf¹ | Fabian Gaessler^{2,3}  | Dietmar Harhoff^{1,4}

¹Department of Innovation and Entrepreneurship Research, Max Planck Institute for Innovation and Competition, Munich, Germany

²Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain

³Barcelona School of Economics, Barcelona, Spain

⁴Centre for Economic Policy Research (CEPR), London, UK

Correspondence

Fabian Gaessler, Department of Economics and Business, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain
Email: fabian.gaessler@ip.mpg.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: CRC TRR 190 “Rationality and Competition”; European Patent Office, Grant/Award Number: Academic Research Programme 2018; Agencia Estatal de Investigación, Grant/Award Number: Barcelona School of Economics CEX2019-000915-S

Abstract

Standard-essential patents (SEPs) have become a key element of technical coordination via standard-setting organizations. Yet, in many cases, it remains unclear whether a declared SEP is truly standard-essential. To date, there is no automated procedure that allows for a scalable and objective assessment of SEP status. This paper introduces a semantics-based method for approximating the standard essentiality of patents. We provide details on the procedure that generates the measure of standard essentiality and present the results of several validation and robustness exercises. We illustrate the measure's usefulness in estimating the share of true SEPs in firm patent portfolios for several telecommunication standards.

1 | INTRODUCTION

In light of increasing demand for the interoperability and interconnectivity of information and communication technologies, standardization has become an important aspect of technological innovation. The successful development and adoption of standards depend on ex ante coordination among technology contributors and implementers—particularly if proprietary technologies are to be incorporated (Lerner & Tirole, 2015; Spulber, 2019). Standard-essential patents (SEPs) protect inventions that are part of technical standards. By definition, any firm implementing the standard requires a license for these SEPs to avoid infringement. However, due to the vast amount of potentially relevant patents and uncertain patent scope, the identification of SEPs poses a considerable challenge to potential implementers. To facilitate the adoption and diffusion of technology standards, standard-setting organizations (SSOs) typically demand from their members to timely disclose SEPs through declaration. This declaration of standard essentiality is based on the assessment of the respective patent holder and usually involves no further verification by the SSO or a third party.

Ideally, only those patents are declared to be standard-essential that protect a relevant technological contribution to the standard, that is, are truly standard-essential. However, there are several factors beyond technical merit that may influence whether a patent is declared standard-essential. Most notably, there are concerns that firms declare patents to

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Journal of Economics & Management Strategy* published by Wiley Periodicals LLC.

be standard-essential due to strategic reasons (Dewatripont & Legros, 2013).¹ Evidence from industry studies and policy reports strongly suggests that there is considerable overdeclaration (see Contreras, 2019, for an overview). Indeed, the claim of standard essentiality frequently fails to survive scrutiny if a declared SEP is disputed in court (Lemley & Simcoe, 2019). Uncertainty about the true standard essentiality of a patent may introduce legal and contractual frictions, as it creates considerable transaction costs during the standardization process and subsequent licensing negotiations. Given the importance of a fair and efficient framework for the development and adoption of technical standards, policy-makers have recently proposed to increase the transparency on the standard essentiality of patents (EC, 2017).

In this study, we introduce a method to approximate standard essentiality based on the semantic similarity between patents and technical standards. Over the last decade, text-based similarity measures have become a popular tool to assess the technological relatedness between patents (e.g., Arts et al., 2018; deGrazia et al., 2020; Gątkowski & Dietl, 2020; Hain et al., 2022; Natterer, 2016; Younge & Kuhn, 2016).² Extending this approach, we illustrate that natural language processing can also be used to calculate textual similarity between patents and standard documents. Through a set of internal and external validation exercises, we show that textual similarity indeed serves as a meaningful approximation of standard essentiality.

We test our semantics-based method on patents and standards at European Telecommunications Standards Institute (ETSI), a leading SSO in the information and communication technology (ICT) sector and responsible for widely used mobile telecommunication standards, such as Global System for Mobile Communications (GSM), Universal Mobile Telecommunications System (UMTS), and Long-Term Evolution (LTE). We calculate the semantic similarity of thousands of patent–standard pairs based on one proprietary and two open-source algorithms. In the first validation exercise, we show that the semantic similarity of SEP–standard pairs is systematically higher than the similarity of pairs where we replace either the SEP with a control patent from the same technology class or the referenced standard with a control standard from the same ETSI standards project. In the second validation exercise, we build up on the results by Bekkers et al. (2017) and show that only SEPs with a high similarity experience a positive “disclosure effect”; that is, they receive more patent forward citations after being declared as standard-essential.

As a third validation, we test the predictive power of semantic similarity to determine true standard essentiality among a set of SEPs declared to ETSI telecommunication standards. To this end, we leverage data on more than 2000 manual SEP assessments, as used in the US court case of *TCL v. Ericsson*. We employ a multivariate logit framework and show that semantic similarity is a statistically significant and important predictor of true essentiality. Notably, this result is robust across all three algorithms to calculate semantic similarity. In other words, the successful implementation of our method does not depend on one particular algorithm. Moreover, we find that the predictive power is lower when selecting the earliest instead of the latest publication in a patent family and when restricting patent text to claims.

In an empirical application, we illustrate our method’s usefulness to determine standard essentiality at the patent portfolio level. As recent legal disputes have exemplified, the calculation of licensing fees for standard technologies often involves not just single SEPs but entire SEP portfolios.³ Extending our predictions to all SEPs to standards for mobile telecommunication (GSM, UMTS, and LTE), we estimate the share of (presumably) true SEPs in firm patent portfolios, and illustrate the high accuracy of our approach at the portfolio level. We find strong firm-level differences in the estimated share of (presumably) true SEPs. These differences are statistically significant and economically substantial. Among all SEP portfolios, the highest-ranked firm has a share of (presumably) true SEPs that is about twice as large as the one for the lowest-ranked firm. In addition, we observe a decline in the share of (presumably) true SEPs over the three successive generations of mobile telecommunication standards. We discuss possible explanations for this trend.

By introducing a semantics-based method to approximate standard essentiality, this study makes several academic and practical contributions. We illustrate how natural language processing can be used to measure the essentiality of patents to specific technical standards. This automated method is not only simple to use, it is also scalable, objective, and replicable. So far, essentiality assessments are done manually, requiring a substantial amount of time and technical knowledge. Our semantics-based method, by contrast, can be applied to large sets of patents and standards. This more resource-efficient approach may be useful to practitioners, policy-makers, and scholars alike. In particular, our semantics-based method may help determine the present or historical population of overdeclared SEPs for a given firm, standard, SSO, or industry. Such insights should facilitate the assessment whether current SSO policies achieve their goal of mitigating patent-related frictions in the standard-setting and implementation process.

The paper is structured as follows: Section 2 surveys the prior literature and describes the complex interplay between patents and standards. Section 3 details the methodology of our semantics-based approach. Section 4 then introduces the data for the subsequent analyses. The validation results are reported in Sections 5 and 6. Section 7 is dedicated to the prediction of SEP portfolio shares, which is followed by a brief summary and concluding remarks.

2 | INSTITUTIONAL BACKGROUND

In this section, we briefly set out how SSOs handle SEPs, explain why SEPs can be of particular value to their holders, and list the reasons behind SEP over- and underdeclaration. Lastly, we elaborate on the practical difficulties of determining the true standard essentiality of patents.

2.1 | SSOs and SEPs

In many industries, in particular the ICT industry, technical standards are developed collaboratively in so-called SSOs. These SSOs provide a forum for their members (e.g., firms, research institutes, and universities) to coordinate the development of standards. In working groups, SSO members identify possible technological solutions and then seek to include the solution that is most conducive to the standard's wide-spread implementation. However, these technological solutions can be proprietary; that is, they are patented. If the standard incorporates these technological solutions, the respective patent holders can exclude others from implementing the standard in their products and processes. To avoid such scenario, SSOs typically set certain rules concerning intellectual property rights (Baron & Spulber, 2018; Chiao et al., 2007; Lemley, 2002).

In these rules, SSOs demand from their members to disclose patents essential to the standard under development (Bekkers & Updegrave, 2013). This disclosure helps the working groups to determine which of the possible technological solutions require implementers to gain a license if included in the standard. The SSO then seeks a commitment from the respective patent holder to liberal licensing terms once the standard is finalized. If this commitment is not achieved, the technological solution will not become part of the final standard.

The exact rules concerning the disclosure of SEPs can differ in numerous ways between SSOs. In the following, we will highlight key differences with relevance to our study and refer the interested reader to Bekkers and Updegrave (2013), Baron and Spulber (2018), and Baron et al. (2019) for a detailed comparison of SSO rules in the ICT industry.

First and foremost, SSOs apply different definitions of standard essentiality (Bekkers & Updegrave, 2013). In general, a patent is standard-essential if the implementation of the respective standard involves a technological solution that falls within the patent's scope of protection—the solution would infringe upon the patent. Most SSOs follow the notion of technical essentiality: a patent is standard-essential if there is no technically feasible way to implement the standard without the patented technological solution. However, some SSOs also consider a patent's commercial essentiality: A patent is standard-essential even if there are technically feasible ways as long as these are not commercially feasible (e.g., too costly). Relatedly, some SSOs also consider patents as standard-essential if they protect optional features of the standard, although implementation is possible without infringing these patents.⁴ Second, SSO rules also differ in terms of the information that patent holders need to disclose. Patent holders formally disclose through SEP declaration letters. Some SSOs mandate that their members declare each SEP specifically. Other SSOs allow their members to make blanket declarations, in which the members only need to declare that they hold SEPs to a given standard, but not how many or which ones exactly. Third, SSO rules can differ regarding the timing of disclosure. Most SSOs specify rules that demand a timely disclosure either upon an official call for patents, before the standard is finalized, or as soon as possible.⁵

2.2 | The value of SEPs

SSOs typically mandate that SEPs be licensed out on predefined terms, such as “fair, reasonable, and nondiscriminatory” (FRAND). This limits the royalties SEP holders can expect from a standard implementer. However, given that practically *all* standard implementers are potential licensees, SEPs can still be a source of considerable revenues. Furthermore, SEPs represent “bargaining chips” that improve their holders' position in cross-licensing negotiations (Shapiro, 2001), which in

turn increases the holders' freedom to operate for product market and innovation activities (cf. Bekkers et al., 2011; Gaessler et al., 2019).

In some cases, SEP holders commit to even less favorable licensing terms, such as royalty-free licenses or nonassertion. Despite foregoing any licensing revenues in this case, SEP holders may still benefit due to the mere fact of having their patented technologies included in a standard (Bekkers et al., 2017). These benefits lie in the SEP holders' familiarity with their own technology and its compatibility with other patented complements.⁶ Irrespective of any licensing revenues, SEPs may also be valuable to their holders as a signal of technological competence to stakeholders, leading, for instance, to more attractive financing deals.

Empirical evidence indeed suggests that SEPs are particularly valuable intellectual property rights. On average, SEPs attract relatively more citations from other patents (which indicates commercial value) over their life time, and especially after their declaration (Bekkers et al., 2017; Rysman & Simcoe, 2008). Moreover, SEP ownership correlates with a firm's financial performance (Hussinger & Schwiebacher, 2015; Pohlmann et al., 2016). The high value of SEPs can also be inferred from the various actions that firms undertake to increase the chance of holding some. First, firms promote their own patented technologies for standard inclusion through SSO membership and active participation in the standardization process (Bekkers et al., 2011; Kang & Motohashi, 2015; Leiponen, 2008). Second, firms also conduct what is commonly known as *just-in-time patenting* to increase the likelihood of holding SEPs (Kang & Bekkers, 2015). That is, firms intentionally file their patent applications late in the standard-setting process to tailor the patent claims to draft versions of the standard. Likewise, firms try to change the claims of already pending patents to ensure that a given standard technology falls within their scope of protection (Berger et al., 2012; Righi & Simcoe, 2020).

2.3 | SEP over- and underdeclaration

In an ideal world, a patent is declared as an SEP if and only if it is truly standard-essential. In reality, however, it has become clear that not all declared SEPs are truly essential and not all truly essential patents are declared as SEPs. In the following, we elaborate on the reasons for such over- as well as underdeclaration.

2.3.1 | Reasons for overdeclaration

The reasons behind overdeclaration concern the nonstatic nature of patents and standards, varying interpretations of patent scope and standard essentiality, and opportunistic behavior.

A declared SEP may lose its essentiality if the standard or the patent's scope of protection changes (Bekkers et al., 2017; Contreras, 2017a, Chap. 13). A patent may have been declared as an SEP based on its essentiality to a preliminary draft of the standard. However, the technology underlying the patent was ultimately excluded from the approved version of the standard. This kind of overdeclaration is particularly likely if patent holders declare their SEPs early in the standard development process and if the SSO disclosure rules also encourage the declaration of patents on technological solutions that *potentially* become part of the standard. Likewise, an SEP can be overdeclared because its scope of protection has changed since declaration.⁷ While these dynamics on standard and patent side can erode true standard essentiality, SEP declarations are rarely withdrawn or updated.

An SEP may be incorrectly declared in good faith (Contreras, 2017a, Chap. 13; Lemley & Simcoe, 2019). There can be considerable uncertainty regarding a patent's exact scope of protection and hence standard essentiality. As a result, the patent holder declares the patent as standard-essential, but an objective assessment may lead to a different conclusion. In a similar manner, there can be uncertainty regarding the exact definition of standard essentiality.

Likewise, an SEP may be incorrectly declared for opportunistic reasons. More specifically, patent holders may pursue overdeclaration to inflate their SEP portfolio as this improves their standing in licensing negotiations (Aoki & Arai, 2018; Bekkers & West, 2009; Dewatripont & Legros, 2013). These incentives likely vary by the practiced method of royalty calculations, the overall SEP landscape, and the patent holder's primary business model. The licensing revenues SEP holders can expect are often a direct function of the number of SEPs in their portfolio. This is particularly true for some royalty calculation practices, such as top-down approaches, where a fixed aggregate royalty rate for standard implementation is often divided between the licensors through simple SEP counting (Contreras, 2017a, Chap. 13).

From this follows that a firm's incentives to overdeclare likely increase with the degree of SEP ownership fragmentation and the overall number of SEPs for a given standard. As the gain from inflating the SEP portfolio primarily relates to licensing revenues, the incentives to overdeclare also vary by the patent holder's business model. Specialized technology contributors are argued to be more inclined to overdeclare SEPs as their business model heavily relies on licensing revenues (cf. Dewatripont & Legros, 2013; Lerner et al., 2016; Simcoe et al., 2009).

2.3.2 | Reasons for underdeclaration

In contrast, underdeclaration refers to the missing declaration of a truly essential patent. The failure to declare an SEP can be unintentional because the patent holder is unaware of the patent's relevance to a particular standard. Indeed, it can take a considerable amount of a patent holder's resources to keep track of all standards under development, to regularly assess which patents in its portfolio are (or may become) essential to these standards, and to generate the information necessary for a specific SEP declaration (Baron & Pohlmann, 2018; Bekkers et al., 2017; Lerner et al., 2016). Given that most SSO rules stay silent concerning the effort members should exert in their patent search (Bekkers & Updegrove, 2013), it is more than plausible that some patent holders miss out on declaring (all of) their truly essential patents.

However, patent holders may also refrain from declaring their truly essential patents to evade the restrictive licensing regimes that SSOs typically impose on declared SEPs. Here, patent holders deliberately keep their patents undisclosed until the SSO has approved the standard and implementers have made irreversible investments into standard implementation (Farrell et al., 2007). The patent holder, free from any royalty cap provisions, can then seek higher licensing fees from standard implementers and enforce the patent without restrictions (Larouche & Zingales, 2017; Lemley & Shapiro, 2006). As a result, keeping their truly essential patents undeclared can—in theory—be highly profitable for a patent holder.

2.3.3 | Prevalence

Most SSOs have intellectual property (IP) rules in place that induce rather over- than underdeclaration (Bekkers & Updegrove, 2013; Contreras, 2017a, Chap. 13). Above all, patent holders may face serious economic and legal consequences if they fail to declare SEPs in a timely manner. For one, courts may render these undeclared patent unenforceable. Moreover, the patent holder can become subject to antitrust liabilities when accused of a so-called patent ambush (Royall et al., 2008). Therefore, most patent holders prefer declaring a given patent as SEP and committing to a moderately attractive licensing regime over the risk that the respective patent becomes unenforceable.

In contrast, declaring standard-irrelevant patents as SEPs is usually without any consequence for the patent holder. In fact, SSOs rarely verify or challenge SEP declarations. For that reason, patent holders, unsure whether their patent is truly essential, may prefer to err on the side of overdeclaration. Importantly, only SSO members and other participants in the development of the standard need to adhere to these rules. Standard outsiders typically have no obligation to declare relevant patents as SEPs (Contreras, 2016), which makes the phenomenon of underdeclaration probably more likely among this group of patent holders.

Whether SEP over- and underdeclaration exist, and to what extent, is ultimately an empirical question, where the answer likely depends on the SSO under investigation (Bekkers et al., 2017). Available estimates from litigation,⁸ industry studies,⁹ and the academic literature¹⁰ suggest considerable overdeclaration rates, indicating that probably less than half of all declared SEPs are truly standard-essential. These estimates have to be interpreted with some caution as they are based on mostly small and nonrandom samples of scrutinized SEPs.¹¹ Nonetheless, at least at some SSOs, such as ETSI, SEP overdeclaration seems to be rampant, causing concern among policy-makers and practitioners (EC, 2017, 2019).

The prevalence of overdeclaration does not automatically imply that underdeclaration does not exist.¹² However, given that undeclared SEPs are hard to detect, the empirical evidence for underdeclaration is anecdotal and concerns legal disputes where patent holders sought to enforce undeclared patents against standard implementers. One often-cited dispute is the one of *FTC v. Rambus*. Rambus failed to disclose its relevant patents during the development of a standard at the Joint Electron Device Engineering Council, an SSO active in the microelectronics industry. Rambus' subsequent royalty claims against locked-in implementers were quickly followed by legal disputes and antitrust

concerns (Tansey et al., 2005). However, due to the lack of any systematic investigation, it remains open whether underdeclaration exists beyond few, but prominent cases.

2.4 | Determining true standard essentiality

A higher transparency concerning the standard essentiality of patents should reduce divergences from the norm; that is, the prevalence of over- and underdeclaration. The most promising way to increase transparency is through standard-essentiality assessment, which—for the time being—represents highly manual work (Aoki & Arai, 2018; Bekkers et al., 2022; EC, 2016).

Manual essentiality assessments are costly and hard to scale because they require highly qualified people, such as engineers, specialized patent attorneys, or examiners, who can construe patent claims and understand the technical descriptions in standard specifications (EC, 2014). In general, the costs per essentiality assessment are a function of multiple factors, such as the scrutiny of the assessment, the level of detail of the SEP declaration, and the availability of claim charts (Bekkers et al., 2022). At the lower bound, assessments cost around 600–1800 EUR per patent, where experts spend only a few hours per patent. More rigorous essentiality checks are conducted by patent pools to ensure the inclusion of only truly essential patents (Contreras, 2017a, Chap. 13; Quint, 2014). Here, the estimated costs range between 5000 and 15,000 EUR per patent. Even more sophisticated assessments, as they are conducted in the context of infringement disputes, can easily cost 20,000 EUR and more per patent.

Given these costs, manual approaches for determining the standard essentiality of all declared SEPs and the dark figure of potentially underdeclared SEPs are unfeasible. There are tens of thousands of declared SEPs at some SSOs and the same patent may need to be assessed more than once given the many-to-many relationship between patents and standards, and changes in patent scope over time. In short, a large-scale system that relies on manual assessments to determine the standard essentiality of patents would require an insurmountable amount of resources and time. This calls for more efficient and cost-effective methods for assessing essentiality.

3 | METHODOLOGY

In this section, we introduce a method to approximate standard essentiality based on the semantic similarity between patents and technical standards. We start with summarizing the current state of the literature on patent text-based similarity measures and on bibliographic approaches to link patents to standards. We then explain the challenges of measuring semantic similarity between patents and standards, and detail the mechanics of our approach. We conclude with a brief discussion on the suitability of semantic similarity to approximate standard essentiality.

3.1 | Prior literature

3.1.1 | Patent text-based similarity measures

Over the last decade, text-based similarity measures have become a popular tool to assess the technological relatedness between patents, complementing—if not replacing—earlier metadata-based measures that relied on the patents' technology classes and citation network (see Hain et al. (2022) and Liu and Shih (2011) for an overview).

As one of the first studies, Gerken and Moehrlé (2012) propose semantic analysis as a method to measure patent novelty and apply it to full-text information on US patents in the automotive industry. Natterer (2016) introduces a calibrated algorithm to measure semantic similarity that can be used to search for technologically closely related patents. His results set the foundation of the *octimine* algorithm, which we will use to calculate the semantic similarity between patents and standards. Younge and Kuhn (2016) use a computationally intensive vector space model and frequency weighting to calculate the full matrix of similarities between US patents based on full-text information. Arts et al. (2018) extract keywords from titles and abstracts of US patents to measure the technological relatedness between patents. Both deGrazia et al. (2020) and Kelly et al. (2021) borrow the term frequency-inverse document frequency (tf-idf) approach of Younge and Kuhn (2016).¹³ Gątkowski and Dietl (2020) calculate a latent semantic similarity model based on full-text information from US patents to identify patent thickets. Whalen et al. (2020) and Hain et al. (2022)

both train embedding models to measure technological similarity between patents. Whereas Whalen et al. (2020) use full-text information, Hain et al. (2022) focus on the abstracts of US patents.

Despite their different approaches to measure textual similarity, all these studies use a homogeneous corpus; that is, texts within the patent universe. One notable exception in this matter is the early study by Magerman et al. (2009). Here, the authors use a vector space model and latent semantic indexing to measure patent-paper similarities of a small set of academic inventors.

3.1.2 | SEP determination

To the best of our knowledge, approximating standard essentiality based on the semantic similarity of patents to standard documents has not yet been explored before.

However, there are two studies that leverage patent metadata to investigate patent–standard relationships. First, Baron and Pohlmann (2018) use the classification of declared SEPs to determine the most relevant technology classes for each standard. They suggest that among all undeclared patents, those in standard-related technology classes are more likely to be of standard relevance. Second, Cho et al. (2021) propose that the essentiality of a patent can be inferred from its position in the extended citation network of SEPs. The more central a patent in that network, the more likely its standard essentiality.

These two methods are interesting and have their merit, especially in contexts where patent text information is not available. However, both methods also come with some limitations. First, the method by Baron and Pohlmann (2018) uses a hierarchical technology classification scheme that remains even at its lowest level coarse. Consequently, there are considerable within- and between-class differences that are not taken into account. Second, the method by Cho et al. (2021) relies on citation networks that are relatively sparse compared with the complete set of patent-to-patent combinations. Finally, and perhaps most importantly, both methods require a critical mass of declared SEPs. This reduces their applicability to investigate recent standards and standards with few or no declared SEPs. In contrast to that, our semantics-based method can capture more nuanced differences between patents and does not depend on any other information than patent text.

3.2 | Measuring semantic similarity between patents and standards

Our semantics-based method to approximate standard essentiality requires an algorithm that can handle long texts, compare texts of different natures, and process large quantities of documents.¹⁴ In the following, we describe the mechanics of three algorithms—one proprietary and two open-source algorithms—that we will use to measure textual similarity (see Table 1 for an overview of the different algorithms).

3.2.1 | Main algorithm

We rely on a sophisticated and field-proven text-mining algorithm to measure the semantic similarity between patents and standards. This algorithm is part of a commercial tool—*octimine*—with the main use case of facilitating prior art search (Jürgens & Clarke, 2018).¹⁵

The *octimine* algorithm has been specifically developed to measure the semantic similarity between patents and texts from other corpora (e.g., product specifications, scientific publications, or Wikipedia articles). It incorporates various text preprocessing techniques and automatic language corrections. To give particular words and concepts in a given text more importance, the algorithm applies weighting methods (i.e., term frequency and entropy). In line with many other approaches, such as Gątkowski and Dietl (2020), every text input is positioned in a reduced vector space, where one dimension can represent multiple words with the same semantic meaning. The cosine distance between any two vectors expresses their semantic similarity with each other.

For our purpose, the main advantages of using this algorithm lie in its efficient handling of long text inputs and its immense scalability. Any input text (such as a standard document) is instantly compared with practically all patents published by the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO), the World Intellectual Property Office (WIPO), or the German Patent and Trademark Office (DPMA).¹⁶ However, using the algorithm also comes with some limitations. To reduce computational complexity, similarity is calculated at the patent

**TABLE 1** Overview of approaches to measure patent–standard text similarity

Approach	Main optimine	Alternatives	
		tf-idf	embeddings
Open source	No	Yes	Yes
Libraries	N/a	tm (R), NLTK (Python)	TensorFlow, PyTorch
Reference	Natterer (2016)	Salton and Buckley (1988)	Beltagy et al. (2019) and Devlin et al. (2019)
<i>Algorithm</i>			
Model	Vector space model	Vector space model	SciBERT
Preprocessing	Stop-word removal, stemming, term reduction, and so forth	Stop-word removal, stemming, term reduction	–
Representation	Latent semantic indexing	Bag-of-words	Document embeddings
Weighting	log-tf + entropy	tf-idf	SPECTER
Similarity metric	Cosine	Cosine	Cosine
<i>Patent corpus</i>			
Sample	All	SEP subsample	SEP subsample
Documents	Latest publication	Earliest/latest publication	Earliest/latest publication
Text input	Full text	Full text/full text without description/claims	Full text

Abbreviations: SciBERT, scientific bidirectional encoder representations from transformer; SEP, standard-essential patent; SPECTER, Scientific Paper Embeddings using Citation-informed Transformer; tf-idf, term frequency-inverse document frequency.

family level, using each patent family's most recent publication. Moreover, this algorithm is not flexible in terms of the patent text input. It always calculates similarity based on full-text information and does not allow the exclusion of certain parts, such as a patent's description or claims. Finally, the algorithm provides us only with similarity scores for the 3000 most similar patents to a given standard text.

3.2.2 | Open-source algorithms

We employ two additional algorithms to measure semantic similarity between patents and standards: *tf-idf* and *embeddings*. Both approaches rely on freely available libraries for natural language processing, but differ in the way how semantic similarity is measured. Employing altogether three distinct algorithms to measure semantic similarity, we can test the generalizability of our semantics-based method to approximate standard essentiality.

tf-idf: The first open-source algorithm, *tf-idf*, is part of the natural language processing framework “tm” in the programming language R. The algorithm first removes special characters, punctuation, numbers, and English stop words, which is followed by word stemming. The preprocessed data are then converted into a (sparse) document–term–matrix. Words are weighted by *tf-idf*, which takes into account a word's importance to the document and its frequency within the entire corpus. The *tf-idf* algorithm relies on a vector space model, where each dimension represents a distinct word. As with our main algorithm, the cosine distance between a patent and a standard text expresses their semantic similarity.

Given its straightforward implementation, scholars frequently use variations of the *tf-idf* algorithm to calculate patent text similarity (e.g., deGrazia et al., 2020; Kelly et al., 2021; Younge & Kuhn, 2016). As the algorithm calculates term weights based on the used set of text (in our case the set of ETSI standards and declared SEPs), specific technical terms and acronyms with high relevance receive more weight. However, the algorithm also comes with two major limitations. First, it cannot relate concepts that are described by different words (synonyms) or distinguish between concepts that are described by the same words (homonyms). Second, without dimensionality reduction, the vector space quickly expands with the volume of text input. As a consequence, the algorithm is computationally demanding and, thus, not easily scalable.¹⁷

embeddings: The second open-source algorithm, *embeddings*, explicitly considers the context of a given text. The algorithm is based on the bidirectional encoder representations from transformers (BERT) developed by Google Inc. (Devlin et al., 2019). While the original language model is trained on Wikipedia articles, we use SciBERT, which is trained on scientific publications (Beltagy et al., 2019), and should be more adequate for our purpose. The use of embeddings allows the representations of text while preserving its original meaning.¹⁸ The algorithm is comparable to those in Whalen et al. (2020) and Hain et al. (2022), but we additionally exploit document-level interdependencies with Scientific Paper Embeddings using Citation-informed Transformers (SPECTER) model weights. Although embeddings have proven to be very powerful in measuring semantic similarity, they may fail to account for certain peculiarities (in patents and standard documents) that do not exist in the training text corpus (i.e., scientific publications).

Due to computational demands, we need to limit the calculation of textual similarity based on these two algorithms to a particular subsample of SEPs and ETSI standard texts. This prevents us from using all three algorithms in all parts of our empirical analysis. At the same time, we leverage the fact that the open-source algorithms give us more flexibility in specifying the patent text input. First, we measure similarity based on two instead of one document per patent family (the earliest in addition to the latest publication) as text input. Second, we also measure similarity based on different parts of the patent text.

3.3 | Semantic similarity and standard essentiality

For illustration purposes, we provide an example of a patent–standard pair with evidently high text similarity. On standard document side, we select the technical specification *ETSI TS 126 192 V8.0.0 (2009-01)*, which describes technologies related to speech coding and comfort noise aspects within the UMTS and LTE telecommunication standards projects. According to our main algorithm, the most similar patent for this specification is the granted US patent with publication number 6,662,155 (“Method and system for comfort noise generation in speech communication”). The patent was declared as essential to the respective standard on June 18, 2009. In Figure 1, we exemplarily contrast parts of the standard document with an excerpt of the patent’s description. Similar and identical words are highlighted to illustrate the semantic similarity of both.

In line with the previous literature on the textual similarity between patents, we interpret the semantic similarity between patents and standards as a measure of their technological relatedness. We consider this a valid extension for the following reasons. First, both patent and standard documents are highly technical texts and can be reasonably compared with each other as illustrated above. Second, standard documents are utilized by patent examiners, patent attorneys, and inventors alike, which underlines their role as informative descriptions of technological solutions.¹⁹

The technological relatedness (and, by implication, the semantic similarity) between a patent and a standard is a necessary but not a sufficient condition for standard essentiality. As Baron and Pohlmann (2018) note, SEPs are only a

Patent publication: US 6,662,155 B2 (2003-12-09) "The background noise can be classified as stationary or non-stationary based on the spectral distances ΔD_i from each of the spectral parameter (LSF or ISF) vectors $f(i)$ to the other spectral parameter vectors $f(j)$, $i = 0, \dots, l_{dtx} - 1$, $j = 0, \dots, l_{dtx} - 1$, $i \neq j$ within the CN averaging period (I_{dtx})."	Standard specification: ETSI TS 126 192 V8.0.0 (2009-01) "The encoder first determines how stationary background noise is. Dithering is employed for non-stationary background noise. The information about whether to use dithering or not is transmitted to the decode using a binary information (CN_{dith}-flag). The binary value for the CN_{dith}-flag is found by using the spectral distance ΔS_i of the spectral parameter vector $f(i)$ to the spectral parameter vector $f(j)$ of all the other frames $j = 0, \dots, l_{dtx} - 1$, $j \neq i$ within the CN averaging period (I_{dtx})."
--	---

FIGURE 1 Textual similarity between patents and standards. The quoted text of the patent publication is part of the patent description. Supporting Information Appendix Figure A1 provides a further example.

subset of all patents that are technologically related to standards. Some patented technological solutions may be useful, but not necessary (i.e., essential) for implementing the standard. Other patented technologies may not be part of the standard at all, but directly build up on solutions described in the standard. Hence, technological relatedness alone is not a sufficient condition for standard essentiality. That being said, if a patent has a low technological relatedness to the standard, it is unlikely that the patented technological solution is part of the standard.

As elaborated in Section 2.1, SSOs vary in their definitions of standard essentiality. This begs the question of which kind of standard essentiality is probably best approximated by semantic similarity. In general, semantic similarity should reflect standard essentiality best if no information beyond the focal patent and the standard is necessary to determine essentiality. However, technical essentiality requires that no alternative technology exists that would allow standard implementation without infringing the focal patent. Whether such alternative technologies exist cannot be inferred from the semantic similarity of the focal patent–standard pair alone. Given that commercial essentiality is more lenient concerning this point (it allows the existence of an alternative technology as long as it is not commercially feasible), semantic similarity may thus reflect standard essentiality according to this definition (marginally) better.²⁰

4 | DATA AND DESCRIPTIVES

In the remainder of the study, we seek to validate our semantics-based approach to approximate standard essentiality. To this end, we focus on one particular SSO: ETSI. In the following, we first introduce the data on ETSI standards and declared SEPs, and subsequently describe the sample.

4.1 | Data

4.1.1 | Standard documents and SEP declarations

The reasons for selecting ETSI to validate our measure are twofold. First, ETSI is one of the most important SSOs in the ICT sector and is responsible for widely used telecommunication standards. ETSI further stands out as the SSO with the by far largest number of declared SEPs (Baron & Pohlmann, 2018). Second, ETSI administers databases for SEP declarations and for standards, which can be freely accessed via its website.²¹

We retrieve data on declared SEPs from ETSI's IPR database. The data are based on declaration letters submitted during the standardization process. As ETSI requires “specific” declarations, the letters specify the publication numbers of all declared SEPs. On the standard side, the declaration letters vary in their level of detail. Whereas some declarations only refer to the overall standards project, many specify the technical specification (TS) and the particular version. We then merge the SEPs to the EPO Worldwide Patent Statistical Database (PATSTAT) and aggregate them at the patent family level. We count nearly 18,000 SEP families.

We process standard documents from ETSI's standards database. At the time of our data collection, the database stored more than 40,000 documents of ETSI standards in portable document format. About 37,000 of these are directly machine-readable. Most documents refer to European standards (EN) and TSs for the different generations of mobile telecommunication standards: GSM, UMTS, and LTE. These documents cover all releases and all versions of the approved standards, depicting the evolution of standards over time. The length of these documents varies substantially. The average number of pages per document is 129 (median: 44), with some documents comprising thousands of pages. We therefore leverage the structured format of standard documents and compartmentalize each document into chapters according to its table of contents.

4.1.2 | Patents

On the patent side, we use two types of data: (1) text information to measure similarity with standard texts, and (2) bibliographic information to characterize the patents. Our main algorithm makes use of patent text information for practically all patents published by the DPMA, EPO, USPTO, or WIPO. This full-text information of each patent

document encompasses the title, abstract, claims, and description of a patent document. Our alternative algorithms also make use of full-text information, but only for a preselected set of patents.

For each patent, we add bibliographic information from PATSTAT (autumn 2017 version).²² In particular, we retrieve information on each patent's family size, patent claims, technology class, inventor team size, and applicants. Moreover, we compute forward and backward citation measures at the patent family level, and calculate the length of the first claim as a proxy for patent scope (cf. Kuhn & Thompson, 2019).

For a particular subset of declared SEPs, we have information on their true essentiality from an external source. This information is based on manual assessments conducted in the context of a legal dispute and is publicly available (for more details, see Section 6.1).

4.1.3 | Similarity data

We construct a data set of pairwise similarity scores between patents and standard texts. We use the similarity scores between patents and standards at the standard chapter level. We do so because only parts of a standard document are relevant for a given technology, and each document may refer to multiple technologies (Contreras, 2019). Indeed, the use of shorter texts (i.e., chapters) on the side of standard documents increases precision.²³ To reduce the computational burden and optimize memory usage, we focus on those standards that have been referenced at least once in SEP declarations. This leaves us with 4796 standard texts at the document and 62,482 standard texts at the chapter level.

Our main measure of semantic similarity is the *similarity score*. The similarity score is an absolute value calculated by the respective algorithm with a theoretical range between 0 and 1.²⁴ Similarity scores of 0 mean that the two input texts have nothing in common whereas scores of 1 imply that they are next to identical.

With our main algorithm, we retrieve pairwise similarity scores for the 3000 most similar patents (at the patent family level) for each standard text. These 3000 patents encompass declared SEPs as well as non-SEPs. Given that our main algorithm compares each standard text to all patents, we can rank patents by their similarity score to a given standard text.²⁵ The total number of calculated similarity scores is about 14 million between patents and standards at *document level*, and about 187 million at *chapter level*.²⁶ Unless stated otherwise, we maintain the highest similarity score of each patent. With our two open-source algorithms, we calculate pairwise similarity scores only for a subset of patents: SEPs with information on their true essentiality as judged by manual assessment.

4.2 | Sample description

We distinguish between three sets of patent–standard pairs in the following analyses. The first set concerns all patents that belong to the 3000 most similar ones. This set includes SEPs as well as non-SEPs. The second set is a subset of the first and concerns all SEPs. Likewise, the third set is a subset of the second and concerns assessed SEPs. These are SEPs for which we have information on their true essentiality.

In Table 2, we report summary statistics for the *similarity score*. We find notable differences in the average similarity between the three sets of patents. The mean *similarity score* in the large set of all patents is 0.18. In contrast, the mean *similarity score* of all SEPs is 0.33, which is about two standard deviations higher than in the superset. For assessed SEPs, we find a mean similarity score that is statistically indistinguishable from the superset of all SEPs. This is not surprising as these assessed SEPs supposedly constitute a random sample (see Section 6 for more information). For completeness, we also report summary statistics for two additional similarity score variables as calculated by our open-source algorithms: *similarity score (tf-idf)* and *similarity score (embeddings)*. Given their differences in mean and standard deviation, these scores cannot be directly compared with each other.

In Figure 2a, we split the set of all patents into SEPs and non-SEPs and plot the two similarity score distributions. The two distributions visibly differ from each other, with the mass of SEPs having considerably higher similarity scores than non-SEPs. That said, there are non-SEPs with high similarity scores and SEPs with low ones. In Figure 2b, we present the distribution of SEPs by their *similarity rank*. About one-third of all declared SEPs belong to the 20 most similar patents for their respective standard text.

We distill three stylized facts from these descriptive statistics. First, SEPs show an overall higher similarity to standards than non-SEPs. We corroborate this finding more systematically in Section 5.1. Second, there is

TABLE 2 Summary statistics: Similarity scores for different sets of patents

	Mean	Std. Dev.	Minimum	Median	Maximum
<i>All patents</i> ($N = 1,772,240$)					
Similarity score	0.180	0.073	0.000	0.166	1.000
<i>All SEPs</i> ($N = 17,823$)					
Similarity score	0.327	0.120	0.048	0.316	0.782
<i>Assessed SEPs</i> ($N = 2287$)					
Similarity score	0.314	0.114	0.048	0.300	0.758
Similarity score (tf-idf)	0.259	0.138	0.018	0.228	0.895
Similarity score (embeddings)	0.626	0.088	0.399	0.622	0.865

Notes: Summary statistics for *similarity score* for all patents that belong to the 3000 most similar ones to ETSI standards, the subset of all SEPs, and the subset of assessed SEPs. Patent–standard similarity is calculated at the standard chapter level. Similarity scores have a theoretical range between 0 and 1.

Abbreviations: ETSI, European Telecommunications Standards Institute; SEP, standard-essential patent; Std. Dev., standard deviation.

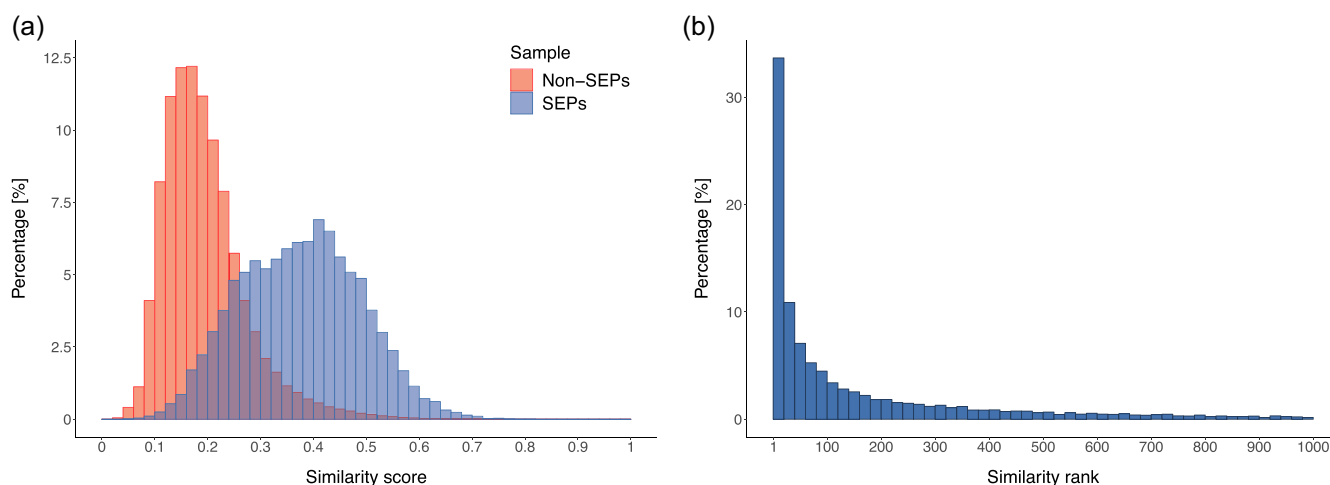


FIGURE 2 Distribution of SEPs and non-SEPs by similarity. (a) Similarity score distributions and (b) rank distribution of SEPs. The left-hand graph shows the similarity score distributions of the two subsets of all patents: non-SEPs (red bars) and SEPs (blue bars). The right-hand graph shows the *similarity rank* distribution for SEPs. Abbreviation: SEP, standard-essential patent.

considerable variation among SEPs with some SEPs having fairly low similarity scores. Taking these scores at face value, this finding may indicate SEP overdeclaration; that is, these particular SEPs may not be truly essential. We investigate this point in Sections 5.2 and 6. Third, there are also several non-SEPs that stand out due to their high similarity to standards.

We briefly compare patent characteristics between non-SEPs and SEPs with high similarity to better understand the reason why these non-SEPs have not been declared as standard-essential (see Supporting Information Appendix Table B1).²⁷ We find that non-SEPs with a high similarity are less valuable than SEPs as they have on average a smaller patent family, fewer references, and fewer forward citations. This may suggest that their patent holders did not see any (commercial) relevance or incentive in declaring them as standard-essential. Moreover, we find that these non-SEPs are filed either significantly earlier or later than the standard publication (see also Supporting Information Appendix Figure A3). This in turn may suggest that for these patents declaration was either unnecessary because the patent was no longer alive, or impossible because the patent was not yet filed. In sum, given that these non-SEPs have a lower value and a larger time difference from the standard publication, we remain doubtful whether they are the result of strategic underdeclaration.²⁸

5 | INTERNAL VALIDATION

In this section, we present the results of two empirical exercises as evidence for the validity of our semantics-based approach to approximate standard essentiality.

5.1 | Comparison of declared SEPs with control groups

In the first validation exercise, we leverage the fact that SEP declarations at ETSI usually include a reference to the relevant standard document. We use these predefined *SEP declaration pairs* to conduct a systematic comparison of their similarity with the similarity of control pairs. These control pairs are modified versions of the *SEP declaration pairs*, where we replace either the SEP with a control patent or the referenced standard with a control standard. The rationale for this comparison is the following: If similarity has any explanatory power in determining standard essentiality, these *SEP declaration pairs* will have a higher similarity than the respective control pairs.²⁹

For the first group of control pairs, we replace the SEP with an undeclared patent from the same technology class and cohort. To this end, we select patent with the same Cooperative Patent Classification-4 code, the same patent priority year, and at least one US or EP publication in the patent family. Control patents are then randomly chosen from this preselected group of patents. For the second group of control pairs, we hold the SEP fixed and replace the referenced standard document with another randomly chosen standard document from the same ETSI standards project³⁰ and the same publication year.

We count about 15,000 *SEP declaration pairs* with a calculated similarity score in our data. The corresponding set of control patents is randomly chosen from the patent universe, which implies that they are not necessarily within the set of the 3000 most similar patents for which we have a calculated similarity score.³¹ In this case, we conservatively assign the lowest observed similarity score for the given standard to the control patent, which likely overstates their similarity.³²

In Figure 3, we present the distribution of similarity scores for each group of pairs as box plots. On the left-hand side, the pairs differ by the patent (SEP vs. control patent). The mean difference in similarity scores between the two groups is 0.097. On the right-hand side, the pairs differ by the standards the patents are compared with. Here, the mean difference in similarity scores is 0.147. All differences are statistically significant with *t*-values greater than 100. Overall, the results of our control group comparison strongly suggest that the similarity of patent–standard pairs is higher for those pairs that—by declaration—the patent holder deemed standard-essential.

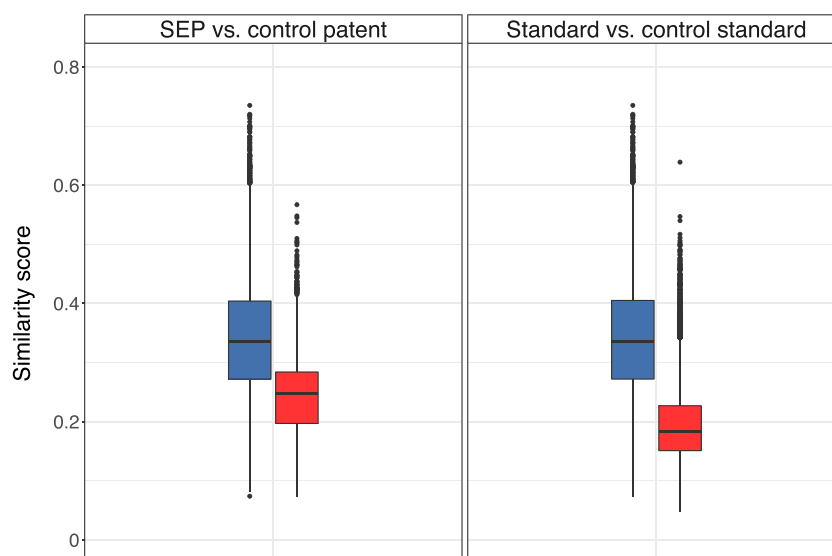


FIGURE 3 Comparison of SEP–standard pairs with control groups. The box plots on the left-hand side illustrate the difference in the distribution of similarity scores of SEPs (blue) and control patents (red) compared with the same standard. The box plots on the right-hand side illustrate the difference in the distribution of similarity scores of SEPs compared with referenced standards (blue) and control standards (red). Further statistics are reported in Supporting Information Appendix Table B2. SEP, standard-essential patent.

5.2 | Replicating the ETSI “disclosure effect”

In the second validation exercise, we build up on the study of Bekkers et al. (2017) and re-estimate the “disclosure effect” of SEP declarations on patent forward citations. Bekkers et al. (2017) propose that the disclosure of SEPs should lead to an increase in patent forward citations, reflecting the gain in economic value after the incorporation of the patented technology into a standard. While they find this to be true for SEPs declared at various other SSOs, the estimated effect is *negative* for SEPs declared at ETSI.³³ Consequently, ETSI may have a high share of declared SEPs that have never become part of a standard and therefore are not truly essential.

Thanks to our similarity measure, we can identify declared SEPs that are particularly similar to their associated standards relative to other patents. On the basis of the assumption that the similarity is particularly high for patents that have become part of the standard, we expect a positive disclosure effect for such a selection of declared SEPs.

We link our similarity information to the authors' data set on declared SEPs, which is publicly available as the “Disclosed Standard Essential Patents (dSEP) Database.” We identify 1183 SEPs declared at ETSI that are among the most similar patents for their associated standards.³⁴ We borrow the empirical design by Bekkers et al. (2017) and use a difference-in-differences approach, in which technologically similar patents with the same citation trend before the SEP declaration serve as control patents. We use a Poisson regression model to estimate the following equation:

$$cites_{it} = f(PostDisclosure_{it} \beta_i + \alpha_i + \gamma_{ay}). \quad (1)$$

The dependent variable $cites_{it}$ is at patent-year level and counts the number of forward citations received by subsequent patents. The independent variable of main interest $PostDisclosure_{it}$ is a binary variable that equals 1 for all SEPs in each year t after the declaration to ETSI and equals 0 otherwise.

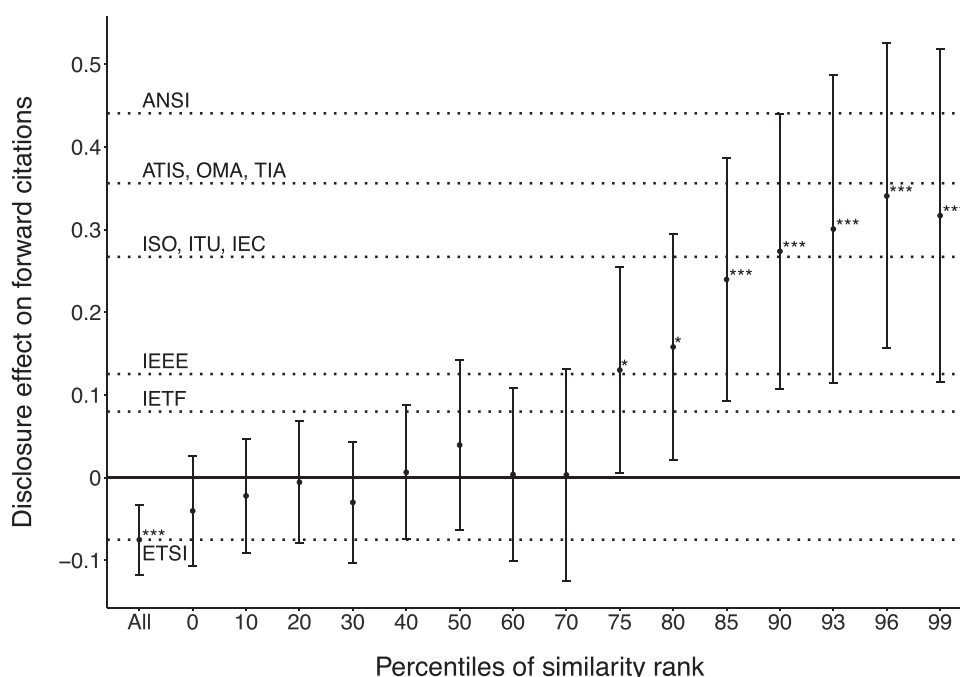


FIGURE 4 Positive disclosure effects of the highest-ranked SEPs at ETSI. This figure shows the estimates of the disclosure effect on forward citations based on the full sample (*All*) in Bekkers et al. (2017), the sample we successfully matched to our data (0) and several subsamples of that, which are defined by the respective percentile of the similarity rank distribution. Poisson estimates and 90% confidence intervals are shown. Each point corresponds to a separate estimate of the coefficient β as shown in Equation (1). Standard errors are clustered on the patent level. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dotted horizontal lines denote the effect sizes at other SSOs, as estimated by Bekkers et al. (2017). ANSI, American National Standards Institute; ATIS, Alliance for Telecommunications Industry Solutions; ETSI, European Telecommunications Standards Institute; IEC, International Electrotechnical Commission; IEEE, Institute of Electrical and Electronics Engineers; IETF, Internet Engineering Task Force; ISO, International Organization for Standardization; ITU, International Telecommunication Union; OMA, Open Mobile Alliance; SEP, standard-essential patent; SSO, standard-setting organization; TIA, Telecommunications Industry Association.

In Figure 4, we present the estimates for the effect of disclosure on forward citations. Following the econometric approach by Bekkers et al. (2017), we successfully replicate their negative disclosure effect for their full sample of SEPs declared at ETSI. However, when focusing on SEP subsamples in higher percentiles of similarity, as measured by their *similarity rank*, the estimated effect increases substantially. At percentiles of 75 and above, the effect sizes become comparable to those that Bekkers et al. (2017) estimated for disclosures at other SSOs. We consider this strong evidence for the validity of our similarity measure to delineate SEP subsamples that are truly essential.

6 | EXTERNAL VALIDATION: PREDICTING TRUE STANDARD ESSENTIALITY

As external validation of our approach, we use the information on the outcome of manual SEP assessments and test the predictive power of semantic similarity to determine true standard essentiality.

6.1 | Data origin and overview

We use the information on the outcome of SEP assessments conducted by an IP consulting firm in the context of a major SEP licensing dispute, *TCL Communication Technology Holdings, Ltd. v. Telefonaktiebolaget LM Ericsson (TCL v. Ericsson* in the following), before the District Court for the Central District of California.³⁵ The case concerned the calculation of royalty fees for SEPs, but also addressed the question of how many SEPs declared at ETSI are truly essential for GSM, UMTS, and LTE standards. The plaintiff (TCL) recruited the IP consulting firm to assess the essentiality of a representative subsample of declared SEPs.

Engineers manually assessed these SEPs by comparing them to the respective standard specifications and spent on average 45 min on each SEP. This process drew some criticism (see Mallinson, 2017, for details). In particular, it was argued that this amount of time would not allow a reliable assessment. In turn, a subsample of SEPs was cross-checked by an independent expert, who found an error rate of less than 10% with a balanced number of false positives and false negatives. The evaluations were ultimately confirmed and accepted in court. We therefore believe that these SEPs assessments provide a reasonably meaningful indicator of true standard essentiality for our purpose.

We match the essentiality assessment information to about 2300 SEPs (family level) from our data set. The assessment outcome is a binary variable with 1 indicating that the SEP is truly essential and 0 indicating that the SEP is nonessential for the relevant standard generation. The distribution of outcomes varies across the three standard generations: about 36% of SEP families are found to be truly essential for LTE, 40% for UMTS, and 39% for GSM standards. These shares are comparable to those found in other industry studies.

In Figure 5, we plot the distribution of similarity scores by assessment outcome for SEPs declared to LTE standards.³⁶ The cumulative frequency of truly essential SEPs strictly dominates the cumulative frequency of nonessential SEPs by similarity score. In other words, truly essential SEPs show a higher similarity.

6.2 | Main results

We use multivariate logistic regressions to predict true standard essentiality. We regress the outcome of the SEP assessment (*SEP truly essential* (*d*)) on semantic similarity (*similarity score*) using different sets of control variables, such as patent characteristics, year fixed effects, technology class fixed effects, and patent holder fixed effects.³⁷

In Table 3, we report the regression results for the sample of SEPs declared to LTE standards. We find positive and statistically significant coefficients for *similarity score* in all five specifications. These coefficients represent marginal effects. In the specification without any controls (column 1), the effect size for a one standard deviation increase in the similarity score ($=0.11$) is about 9 percentage points (pp). The size of the coefficient decreases with the addition of controls in columns (2) and (3), but remains statistically highly significant. The robustness to the inclusion of a large set of patent characteristics, year, and technology fixed effects alleviates the concern that the correlation of the similarity score with standard essentiality merely reflects different wording styles over time or technologies. In column (4),

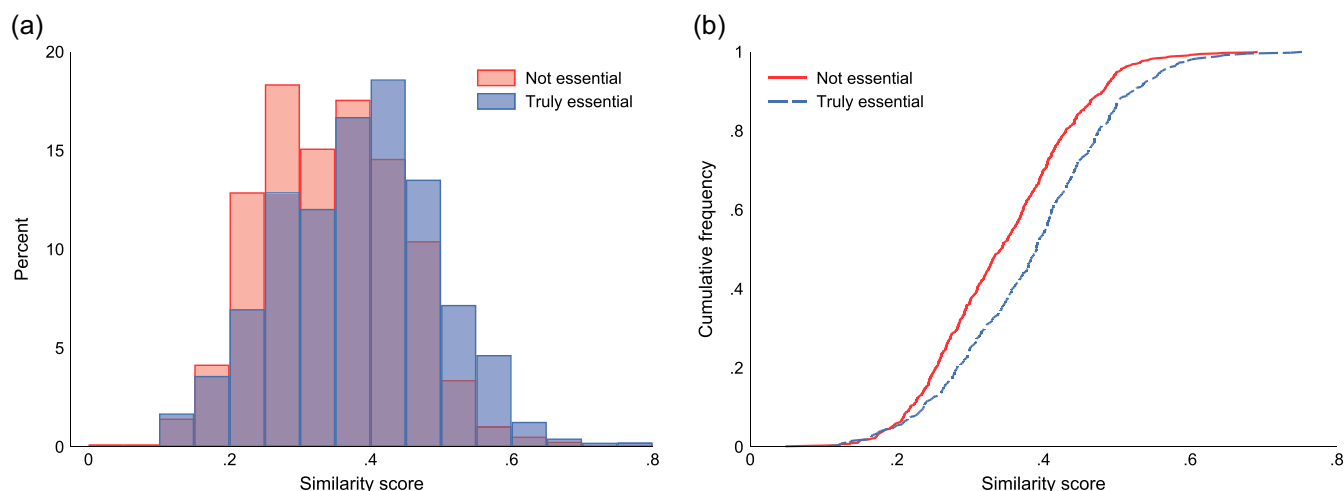


FIGURE 5 Distribution of SEP similarity scores by essentiality assessment (LTE standards). (a) Relative frequencies and (b) cumulative frequencies. The left-hand graph shows the similarity score distributions of the two subsets of assessed SEPs declared to ETSI LTE standards: not essential SEPs (red bars) and truly essential SEPs (blue bars). The right-hand graph shows cumulative frequencies for both subsets. ETSI, European Telecommunications Standards Institute; LTE, Long-Term Evolution; SEP, standard-essential patent.

TABLE 3 Logistic regressions: LTE standard essentiality

	DV: SEP truly essential (d)				
	(1)	(2)	(3)	(4)	(5)
Similarity score	0.8258*** (0.1319)	0.6993*** (0.1382)	0.5441*** (0.1649)	0.3454** (0.1746)	0.5115*** (0.1527)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority year FE	No	No	Yes	Yes	Lasso
Earliest declared year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Firm FE	No	No	No	Yes	No
Pseudo- R^2	0.024	0.053	0.124	0.156	0.120
AUC	0.606	0.667	0.730	0.755	0.726
Observations	1241	1241	1241	1241	1241

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for LTE standards as judged by SEP assessment. Marginal effects of one unit change are reported. The sample size is fixed in all specifications to ease the comparison of coefficients from different models. Standard errors in parentheses. Information on the control variables can be found in Supporting Information Appendix Table B1.

Abbreviations: AUC, Area under ROC curve; CPC, Cooperative Patent Classification; DV, dependent variable; FE, fixed effects; LTE, Long-Term Evolution; ROC, receiver operating characteristic; SEP, standard-essential patent.

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

we further include patent holder fixed effects. The significant correlation of the similarity score in this specification suggests that similarity has explanatory power for true essentiality even *within* SEP portfolios.

To avoid overfitting due to the inclusion of an excessively large number of independent variables, we also employ a penalized (Lasso) regression approach that tosses out independent variables with insufficient explanatory power. The estimated coefficient in the Lasso regression model in column (5) is highly comparable to the one in column (3) without patent holder fixed effects (which we deliberately exclude from the Lasso regression). This result further underlines the explanatory power of the similarity score. As a matter of fact, the similarity score is selected as one of the very first (i.e., most relevant) independent variables in the Lasso specification.

We can corroborate the positive relationship between the *similarity score* and true standard essentiality for SEPs declared to GSM and UMTS standards (see Supporting Information Appendix Tables B6 and B7). Although the number of assessed SEPs for these two standard generations is considerably smaller, we find statistically significant correlations in practically all specifications. The effect sizes appear to be even larger for SEPs declared to these standard generations. A one standard deviation increase in the similarity score corresponds to a 14.8 pp increase in true essentiality for UMTS and 15.3 pp for GSM.

6.3 | Alternative similarity measures

So far, we measured semantic similarity on basis of our main algorithm. In the following, we will extend our result (semantic similarity predicts true standard essentiality) to similarity measures based on our alternative open-source algorithms.

In Table 4, we report the regression results for the sample of SEPs declared to LTE standards with similarity measured by *similarity score (tf-idf)* in columns (1)–(5) and *similarity score (embeddings)* in columns (6)–(10).³⁸ We

TABLE 4 Logistic regressions: LTE standard essentiality (alternative algorithms)

	DV: SEP truly essential (d)				
	(1)	(2)	(3)	(4)	(5)
Similarity score (tf-idf)	0.7046*** (0.1043)	0.6624*** (0.1091)	0.5838*** (0.1245)	0.4492*** (0.1308)	0.6266*** (0.1158)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority year FE	No	No	Yes	Yes	Lasso
Earliest declared year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Firm FE	No	No	No	Yes	No
Pseudo- R^2	0.028	0.060	0.131	0.161	0.120
AUC	0.612	0.665	0.735	0.760	0.727
Observations	1241	1241	1241	1241	1241
	(6)	(7)	(8)	(9)	(10)
Similarity score (embeddings)	0.7983*** (0.2509)	0.6651*** (0.2555)	0.5347** (0.2782)	0.4365 (0.2961)	0.5502** (0.2670)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority year FE	No	No	Yes	Yes	Lasso
Earliest declared year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Firm FE	No	No	No	Yes	No
Pseudo- R^2	0.006	0.042	0.120	0.155	0.112
AUC	0.556	0.641	0.726	0.754	0.720
Observations	1241	1241	1241	1241	1241

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for LTE standards as judged by SEP assessment. *Similarity score (tf-idf)* and *Similarity score (embeddings)* are based on two alternative open-source algorithms: tf-idf and embeddings. For more information on the algorithms, see Section 3.2. Marginal effects of one unit change are reported. AUC, Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Information on the control variables can be found in Supporting Information Appendix Table B3.

Abbreviations: AUC, Area under ROC curve; CPC, Cooperative Patent Classification; DV, dependent variable; FE, fixed effects; LTE, Long-Term Evolution; ROC, receiver operating characteristic; SEP, standard-essential patent; tf-idf, term frequency-inverse document frequency.

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

find positive and statistically significant coefficients for both similarity measures. For *similarity score (tf-idf)*, the increase in true standard essentiality by one standard deviation ($= 0.14$) is comparable to that of our main algorithm in the univariate model and even higher when considering the Lasso models. For *similarity score (embeddings)*, the coefficients appear very similar to those of our other two algorithms. However, the effect sizes are considerably smaller considering the score's dispersion: a one standard deviation increase in the similarity score ($= 0.06$) increases the likelihood of true standard essentiality by about 5 pp. Combined with the lower precision in the estimates, the *similarity score (embeddings)* appears to be inferior to the other two similarity measures in predicting standard essentiality.

In sum, we can confirm the generalizability of our semantics-based method in terms of *how* semantic similarity is measured. That said, the explanatory power varies between the algorithms. The tf-idf algorithm performs surprisingly well given that it cannot capture latent semantic similarity (e.g., synonyms). Presumably, this weakness is outweighed by the fact that the algorithm is able to weight words according to their corpus-specific relevance. The embeddings approach, while technically more capable to represent semantic meaning, probably suffers from not being calibrated to patent and standard text. A modified version that is pretrained on patent and standard documents (instead of scientific articles) may perform substantially better.

6.4 | Robustness

In the following, we check the robustness of our results by changing the underlying patent text used for calculating the similarity score of the patent (family).

In Supporting Information Appendix Table B10, we report the coefficients of similarity (*similarity score [tf-idf]*) based on different parts of the patent publication of assessed SEPs declared to LTE standards.³⁹ We find that the similarity score shows a statistically significant positive correlation with true standard essentiality across all variants of patent text (full text, full text without description, and claims text only). The explanatory power of the similarity score, however, is lower for claims text when compared with the two more comprehensive text variants.⁴⁰ This result is consistent with intuition: claims are typically written in a highly abstract and generic language, which should complicate its direct comparison to standard texts. In line with this, interviewed patent attorneys confirmed that a patent's description is frequently considered in manual essentiality checks. At the same time, the inclusion of the description may not be representative of the patented invention, as patent holders can deliberately add information to the description that goes beyond the boundaries of the invention itself. This trade-off between precision and potential bias in the selection of patent text deserves more attention in future work.

In Supporting Information Appendix Table B10, we further report the results for the similarity score based on different text variants from the *earliest* publication in the SEP family. We find that the positive relationship with true standard essentiality is throughout weaker for the earliest publication compared with the latest one. Given that standard essentiality was typically not assessed based on these publications but later ones within the SEP family, this result is hardly surprising. Nonetheless, it is interesting to see that even claim text variation *within* a patent family, which likely reflects changes in patent scope, has a bearing on the prediction of standard essentiality.

6.5 | Predictive performance

We briefly discuss the performance of our logistic regressions in predicting true standard essentiality. In Figure 6, we present the precision-recall curve and the composition of SEPs by essentiality status within bins of the predicted probability of true essentiality (based on the model in Table 3, column 5). We can infer that our classifier performs moderately well across different probability thresholds, and is very accurate in classifying SEPs with a particularly low (or high) probability of true essentiality.

To further assess predictive performance, we employ fivefold cross-validation with three repeats for all of our predictions. To this end, we split the sample into a test and training data set. A total of 70% of the data are used for training and 30% to test our model.⁴¹ These test and training data sets are also used in the SEP portfolio estimations in Section 7. The weighted precision and recall scores from the univariate model (see Table 3, column 1) are 62.6% and 65.1%. For the lasso model (Table 3, column 5), precision and recall scores slightly increase to 63.5% and 65.6%, respectively. This suggests that the inclusion of control variables does not improve the predictions substantially.

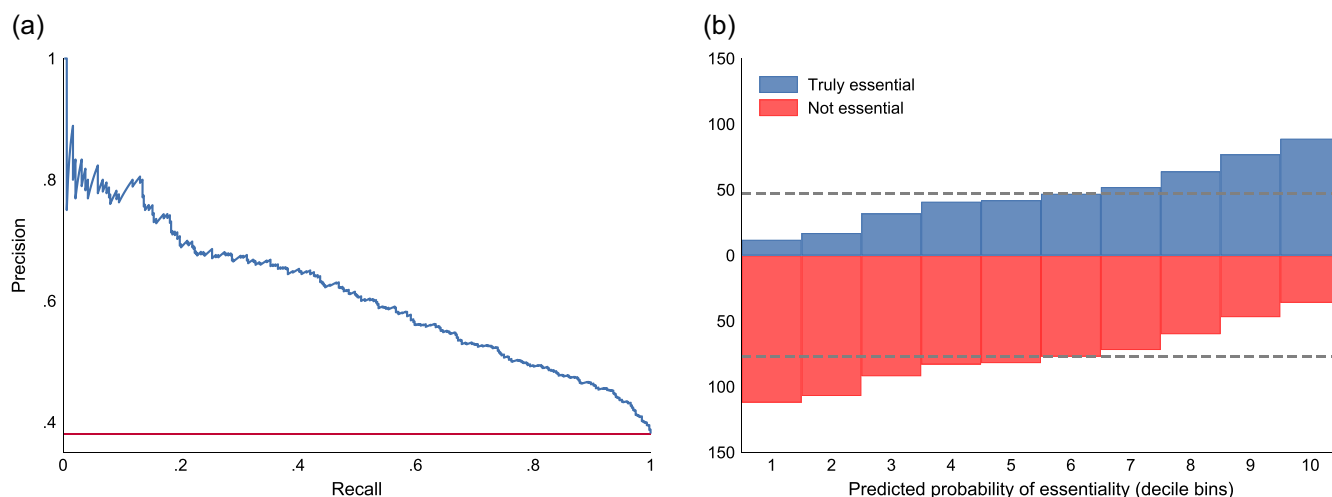


FIGURE 6 Predictive performance (LTE standards). (a) Precision-recall curve and (b) essentiality status by predicted essentiality. The left-hand graph illustrates the precision-recall curve. The right-hand graph visualizes the composition of SEPs by essentiality status within bins of the predicted probability of true essentiality. Visualization adopted from Baron and Pohlmann (2021). Predictions based on the specification in Table 3, column 5. LTE, Long-Term Evolution; SEP, standard-essential patent.

Likewise, alternative machine learning classifiers, such as Random Forest, Naïve Bayes Classifier, Support Vector Machine, and Extreme Gradient Boosting, do not further improve the predictions (see Supporting Information Appendix Table B14).⁴²

7 | PREDICTING SEP PORTFOLIO SHARES

We use the results from Section 6 to predict patent owner portfolio shares of declared SEPs that—according to our model and data—are truly standard-essential. As discussed above, our predictor's accuracy is only moderate at the level of individual patents. However, the prediction errors partially cancel out at the portfolio level as long as we have obtained consistent estimators at the micro level.

7.1 | Estimation

On the basis of the presented logistic regression results, we compute the predicted probabilities of standard-essentiality for a given SEP to estimate the share of truly essential SEPs $\hat{P}_F$ at the firm portfolio level with the following equation:

$$\hat{P}_F = \frac{1}{n} \sum_{i=1}^n \hat{p}_i = \frac{1}{n} \sum_{i=1}^n \frac{e^{\hat{\beta}_0 + \sum_{j=1}^K \hat{\beta}_j X_{ij}}}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^K \hat{\beta}_j X_{ij}}}, \quad (2)$$

where n is the number of patents for a given firm F and X_{ij} represents the explanatory variables used in the logistic (Lasso) regression (see Table 3, column 5).

We draw random portfolios from the test data set on LTE SEPs to determine the error of our prediction, on an aggregated level, as a function of the number of SEPs in the portfolio.⁴³

First, we compute the predicted probabilities for the test sample based on the logistic regression results from the training data set. We then use random sampling with 100 repetitions without replacement to determine the difference between actual and predicted essentiality ratios for varying numbers of portfolio sizes. Figure 7 plots these differences in predicted and actual shares of true SEPs against the size of the SEP portfolio. For portfolio sizes of 50 (200) patents, the error is approximately 5.5 pp (2.8 pp). Many firms have even larger SEP portfolios for a given standard. In such cases, the errors converge towards 0 in a strictly decreasing function.

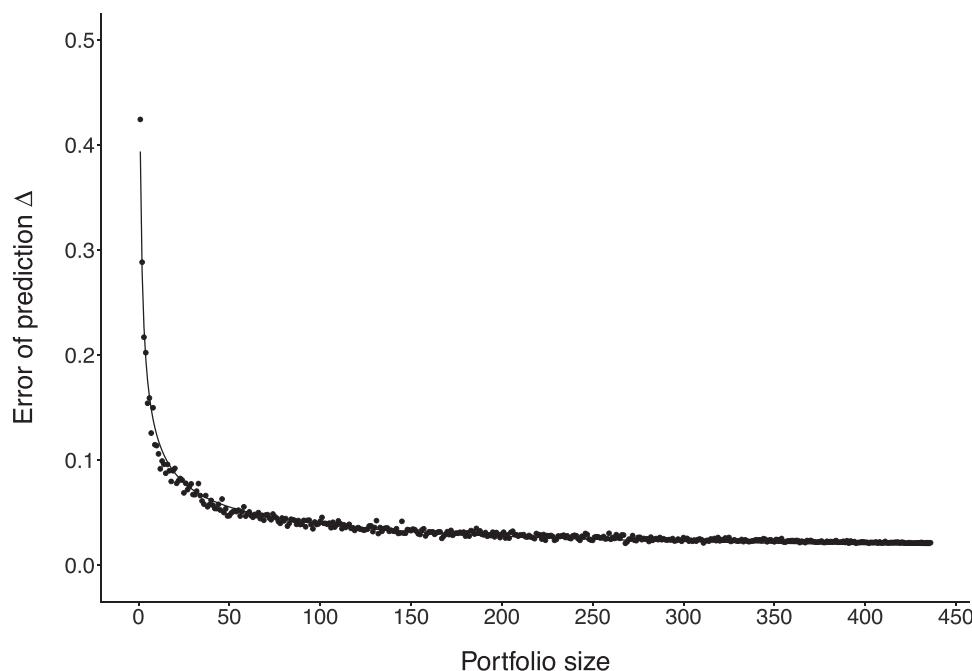


FIGURE 7 Error of prediction as a function of portfolio size (LTE). The error of prediction Δ is plotted as a function of portfolio sizes where portfolios are randomly drawn from the test sample. Additionally, a nonlinear least squares fit is shown for the test sample of LTE patents. The fitted function is a power law function. LTE, Long-Term Evolution.

7.2 | Results

In Figure 8, we present out-of-sample predictions for firm SEP portfolios separately for all three standard generations. In Figure 8a, the overall share of presumably true SEPs for LTE standards is 31.3%.⁴⁴ At the portfolio level, the share of presumably true SEPs varies substantially from less than 20% to more than 45%. The highest-ranked firm thus has a share of presumably true SEPs that is more than twice as large as the one for the lowest-ranked firm. Notably, there seems little correlation between the share of presumably true SEPs and portfolio size. In Figure 8b,c, we present similar estimations for patents declared to UMTS and GSM standards. Interestingly, the average shares of essential patents are larger for these older generations of mobile telecommunication standards (33.9% for UMTS and 35.7% for GSM).

We leave the question as to what causes this trend for future work. However, one explanation may be the changing composition of companies contributing technological solutions to standards. First, the number of firms holding a portfolio of at least 20 declared SEPs has increased sharply. Second, the set of patent holders has become more diverse in terms of their business models. Specifically, the number of upstream technology contributors with few, if any, activities in the product market has increased.⁴⁵

8 | SUMMARY AND CONCLUDING REMARKS

In this paper, we propose an automated method to approximate standard essentiality based on the semantic similarity between patents and technical standards. We provide evidence for the validity of our method through several empirical exercises. Most notably, we find that semantic similarity is a strong predictor of true standard essentiality among declared SEPs. Moreover, we can show that this result is robust across different algorithmic approaches to measure semantic similarity.

We illustrate the usefulness of our method in a first application: the estimation of the share of true SEPs in SEP portfolios. We find statistically and economically substantial differences between firms. The highest-ranked firm has a share of presumably true SEPs that is about twice as large as the one for the lowest-ranked firm. We further observe a general decline in the average share of presumably true SEPs over the three generations of mobile telecommunication standards. This pattern may be due to changes in the number and composition of technology contributors with

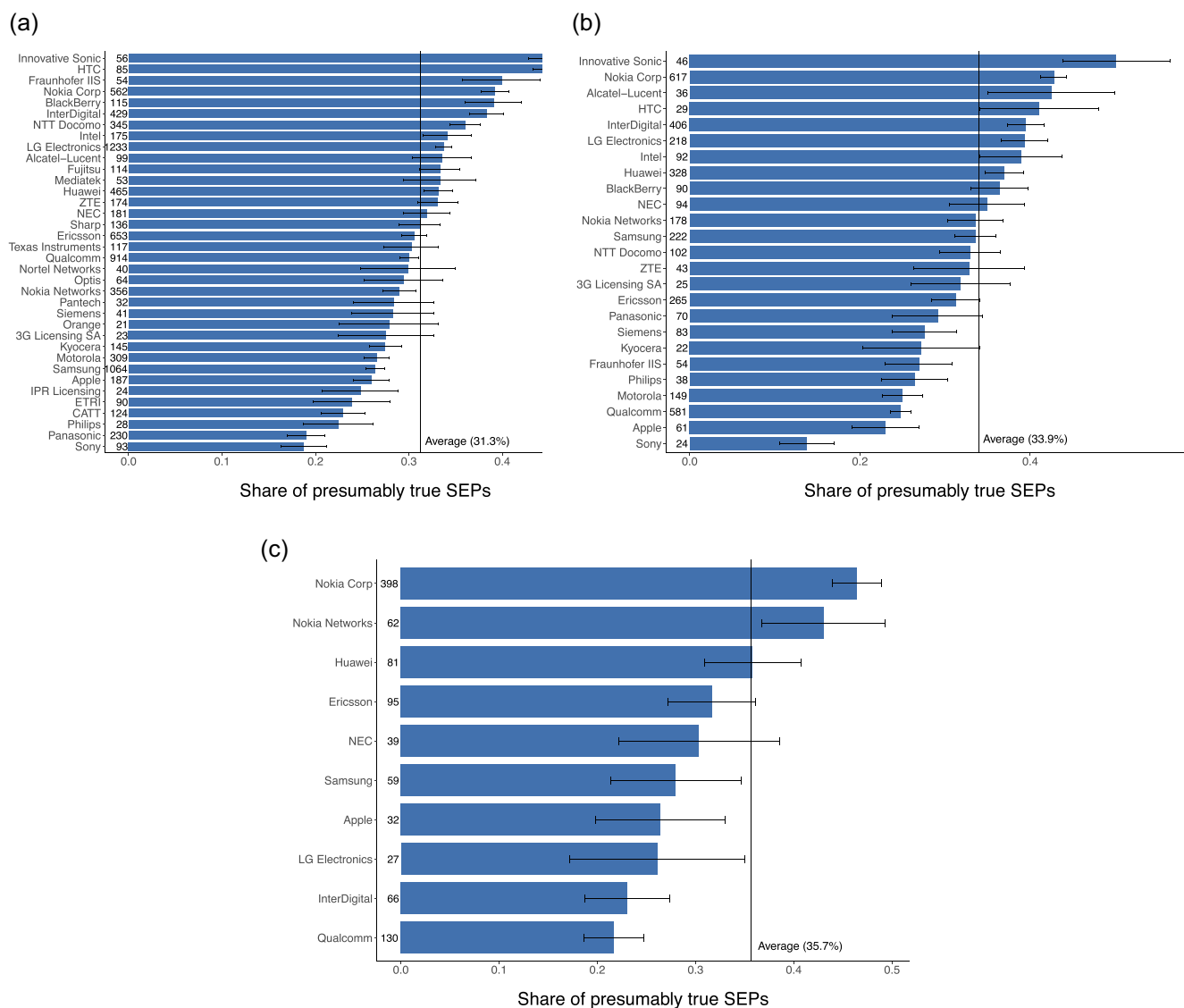


FIGURE 8 Share of truly essential SEPs at portfolio level (out-of-sample predictions). (a) LTE, (b) UMTS, and (c) GSM. The top left-hand graph shows the out-of-sample predictions at the portfolio level for LTE, the top right-hand graph for UMTS, and the lower graph for GSM standards. The numbers on the left-hand side of the bars indicate the count of SEPs declared to LTE/UMTS/GSM standards by the respective patent holder. Only results for patent holders with 20 or more declared SEPs are reported. In all graphs 95% confidence intervals are shown. GSM, Global System for Mobile; LTE, Long-Term Evolution; SEP, standard-essential patent; UMTS, Universal Mobile Telecommunications System.

different business models. This purely descriptive yet intriguing result poses a worthwhile subject for future work on standards and firm behavior.

The validation and application of our semantics-based method focus on patents and standards at ETSI. While ETSI is frequently associated with SEP overdeclaration, we understand the usefulness of our method not limited to one particular SSO or industry.⁴⁶ In light of different IP policies, we consider the systematic investigation of overdeclaration at other SSOs a promising extension of our work.

Beyond that, we see several potential applications of our method in the academic as well as the practical realm. This encompasses use cases that particularly benefit from the method's scalability and do not rely on perfect accuracy in determining standard essentiality. For instance, our method can complement the long-winded search for relevant patents during the standardization process, as it can minimize false negatives; that is, relevant patents that would otherwise remain undetected. Potentially, the continuous measure of semantic similarity underlying our method may provide a cost-effective yet objective way to improve current royalty calculation practices. Instead of determining

royalty rates by simply counting SEPs, one may weight each SEP by its similarity score. This small modification may help erode the incentives for licensors to inflate their SEP portfolios.

In general, our method makes information about the true essentiality of SEPs more accessible. Against this backdrop, we hope our method will allow more scholars to empirically study the important yet complex relationship between patents and standards.

ACKNOWLEDGMENTS

We thank Maddalena Agnoli, Timo Ali-Vehmas, Andrea Bonaccorsi, Pere Arque-Castells, Justus Baron, Christian Catalini, Christian Fons-Rosen, Joachim Henkel, Roman Jurowetzki, Pooyan Khashabi, Yann Ménière, Elena Romito, Timothy Simcoe, Robin Stitzing, Lisa Teubner, Vicente Zafrilla, and participants at the Annual Conference of European Policy for Intellectual Property, the DRUID Academy Conference, the EPO Academic Research Programme Workshop, the Annual VHB Conference, the Twelfth Annual Northwestern/USPTO Conference on Innovation Economics, the ZEW/MaCCI Conference on the Economics of Innovation and Patenting, and the Research Roundtable: “Mechanisms, Governance, and Policy Impact of SEP Determination Approaches” for their valuable comments. We would also like to thank Cesare Righi for introducing us to the dSEP database, and Michael Natterer and Matthias Poetzl at octimine technologies for providing us with data used in this study. Financial support through the EPO Academic Research Programme 2018 and the Deutsche Forschungsgemeinschaft through CRC TRR 190 “Rationality and Competition” is gratefully acknowledged. Fabian Gaessler acknowledges financial support from the Spanish Agencia Estatal de Investigación (AEI), through the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S). Open Access funding enabled and organized by Projekt DEAL.

ORCID

Fabian Gaessler  <http://orcid.org/0000-0001-6178-4097>

ENDNOTES

- ¹ Several other reasons may also play a role (Bekkers et al., 2011). First, standards as well as patents may change in their scope over time. Second, disclosure rules imposed by the SSO may be ambiguous, affecting patent holders in their decision to declare patents as standard-essential. Third, patent holders may simply lack familiarity with the standard and/or their own patent portfolio.
- ² Furthermore, De Grazia et al. (2021) and Kuhn and Teodorescu (2021) measure patent claim text similarity to quantify changes in patent scope during the examination process.
- ³ As Contreras (2017a, Chap. 13) states, the recent case of *TCL v. Ericsson* “[...] highlights the potential importance of essentiality determinations not on a patent-by-patent basis, but on an aggregate basis.”
- ⁴ Moreover, some SSOs take into account the cumulative nature of standards when defining essentiality; that is, a patent essential to a given standard is also essential to subsequent standards that build upon the former one (Baron & Pohlmann, 2018).
- ⁵ Many disclosures, however, happen long after the standard is set and published (Bekkers et al., 2017).
- ⁶ Moreover, SEP holders may face licensing costs themselves if an alternative patented technology finds entrance into the standard instead of their own.
- ⁷ During patent examination, amendments to the claims of the patent application may change the patent’s relevance to a given standard. After the patent has been granted, its scope of protection may be narrowed as a result of validity challenges with the effect that the patent is no longer infringing through standard implementation.
- ⁸ In the context of the recent case *Ericsson v. TCL*, more than 2000 SEPs for the mobile telecommunication standards GSM, UMTS, and LTE were assessed to determine royalty rates. We employ this data in the empirical part of this study.
- ⁹ See, for instance, the publicly available reports by Goodman and Myers (2005), Cyber Creative Institute, Article One Partners, Jefferies, and iRunway.
- ¹⁰ See, for instance, the studies of Stitzing et al. (2017) and Lemley and Simcoe (2019).
- ¹¹ Moreover, the quality of the conducted essentiality checks in terms of thoroughness and objectivity is often hard to evaluate, which has raised some doubts about their credibility (cf. Mallinson, 2017).
- ¹² Over- and underdeclaration are not necessarily mutually exclusive phenomena. In theory, a patent holder may decide to declare standard-irrelevant patents as SEPs, while keeping the truly essential patents in its portfolio undeclared.
- ¹³ deGrazia et al. (2020) apply it to claim text to detect vertically overlapping patent claims, while Kelly et al. (2021) apply it to full-text information to identify patents with particularly high novelty and impact.

- ¹⁴ To put the last point into perspective: investigating SEP overdeclaration at ETSI involves the comparison of more than 10,000 patents to thousands of referenced standard documents, which results in millions of patent–standard pairs. For the investigation of underdeclaration, the set of patents is even larger.
- ¹⁵ We extracted the data in late 2018. For more information on the tool (in its current form), see <https://www.octimine.com>
- ¹⁶ Patent documents in a foreign language are machine translated into English. Note that less than 5% of declared SEPs at ETSI do not have a patent family member in English.
- ¹⁷ That said, it is not impossible to use tf-idf algorithms for large data sets. For instance, Younge and Kuhn (2016) compute the pairwise similarities of more than five million US patents with the help of cloud computing services.
- ¹⁸ That is, the semantic structure remains intact through maintaining the location of a term within a sentence, a paragraph, and a document.
- ¹⁹ For instance, standard documents contain relevant prior art that is used to assess a patent's novelty during examination (Bekkers et al., 2016).
- ²⁰ For technical reasons, semantic similarity is also more appropriate to capture standard essentiality at the patent level than at the claim level. Determining “claim essentiality” would require the comparison of standards to a patent's individual claims. However, this implies that the input text on the patent side would be very short: the average claim contains fewer than 200 words altogether (Kuhn & Thompson, 2019). It stands to reason that this would lower the discriminatory power of any common measure of semantic similarity.
- ²¹ SEP declarations and standard documents can be accessed at ipr.etsi.org and, respectively, www.etsi.org/standards-search
- ²² The Worldwide Patent Statistical Database PATSTAT from the EPO covers the entire history of patents worldwide and provides bibliographic information, such as patent and inventor information.
- ²³ We find evidence for the increased precision when plotting the shares of declared SEPs among the most similar patents as calculated at chapter versus document level (Supporting Information Appendix Figure A).
- ²⁴ The raw values of the similarity score calculated by our main algorithm are integers and scaled between 0 and 1000. To ease comparison, we divide these values by 1000.
- ²⁵ The *similarity rank* is highly correlated with the similarity score. However, there are some subtle differences in how to interpret them. Whereas the former can be considered as a measure independent from other patents and comparable across standards, the *similarity rank* provides the standard-specific order of the most similar patents.
- ²⁶ The similarity data are freely available to researchers as part of the *Semantic similarity of patent–standard pairs (ETSI, IEEE, and ITUT) database*: <https://doi.org/10.7910/DVN/B2RJSX>. For more information on the database, see Brachtendorf et al. (2020).
- ²⁷ We focus on all patents with a similarity score of at least 0.6 and a similarity rank of 5 or lower. We further restrict the set of patents to those that have at least one family member granted at the EPO or the USPTO, because patent holders are unlikely to declare rejected patent applications as SEPs.
- ²⁸ At the same time, more than 80% of the non-SEPs are part of patent portfolios with at least one declared SEP. This renders it unlikely that their holders are standard outsiders with no obligation to declare.
- ²⁹ For this exercise, we disregard potential overdeclaration among the declared SEPs. A high overdeclaration rate would render the observed mean difference in similarity conservative.
- ³⁰ We classify standard documents based on keywords from the title of the standard document. We differentiate between the following groups of standards: LTE, UMTS, GSM, Digital Enhanced Cordless Telecommunications (DECT), Terrestrial Trunked Radio (TETRA), Digital Video Broadcasting (DVB), Digital Audio Broadcasting (DAB), Integrated Services Digital Network (ISDN), or any other standard.
- ³¹ As we only observe the 3000 most similar patent families for each standard text, we have to deal with either truncation or censoring. Truncating the sample, we obtain a total of 29,380 treated and control patents.
- ³² We obtain similar results using censoring instead of truncating (see Supporting Information Appendix Figure A4).
- ³³ The authors explain this negative effect with ETSI's IP policy. The early disclosure of potentially essential patents induces competition effects. The disclosure of patents covering poor technological solutions may be followed by the emergence of alternative technological solutions, which then become part of the standard instead.
- ³⁴ Although we can link 91% of all SEPs in the dSEP database to our data, the strict match to *one* specific document further reduces the analysis to about 60% of the dSEP sample.
- ³⁵ An elaborate discussion of this case and the decision can be found in Contreras (2017b) and Picht (2018).
- ³⁶ The similarity score and the assessment outcome can differ at the patent level because some SEPs are declared to standards from multiple generations (GSM, UMTS, and LTE) and thus assessed multiple times.

- ³⁷ The inclusion of a broad set of control variables mitigates the risk of omitted variable bias. The focus on only *one* representative patent per family reduces the risk of a violation of the independence assumption.
- ³⁸ The corresponding results for UMTS and GSM standards can be found in Supporting Information Appendix Tables B8 and B9.
- ³⁹ The corresponding results for UMTS and GSM standards can be found in Supporting Information Appendix Tables B11 and B12.
- ⁴⁰ Note that the dispersion of similarity scores is comparable between the patent text variants.
- ⁴¹ We report the confusion matrix for the test set of 436 SEPs for LTE standards in Supporting Information Appendix Table B13.
- ⁴² An ensemble model, which employs multiple classifiers, may lead to further improvements. We leave this for future research.
- ⁴³ We hereby assume that firms' patent portfolios are randomly composed. The composition of firms' patent or SEP portfolios may be based on strategic decisions. However, the error of prediction should remain largely unaffected by portfolio composition and hence provide a general, firm-independent function.
- ⁴⁴ This share is lower than the one based on the benchmark evaluations in the manual SEP assessments sample. The difference mainly results from the focus on patents in firm portfolios with 20 or more SEPs.
- ⁴⁵ Indeed, the share of presumably true SEPs seems to be particularly low for several patent holders that are *upstream* technology contributors. This observation is consistent with the argument that patent holders with upstream business models tend have stronger incentives to inflate their SEP portfolios than patent holders with product market activities (cf. Dewatripont & Legros, 2013).
- ⁴⁶ In fact, small-scale evidence suggests that semantic similarity also predicts true essentiality for SEPs declared at IEEE and ITU-T (Brachtendorf et al., 2020).

REFERENCES

- Aoki, R., & Arai, Y. (2018). *Strategic declaration of standard essential patents* [RIETI Discussion Paper Series 18-E-035]. <https://www.rieti.go.jp/jp/publications/dp/18e035.pdf>
- Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1), 62–84.
- Baron, J., Contreras, J. L., Husovec, M., Larouche, P., & Thumm, N. (2019). *Making the rules: The governance of standard development organizations and their policies on intellectual property rights* (JRC Science for Policy Report). EUR.
- Baron, J., & Pohlmann, T. (2018). Mapping standards to patents using declarations of standard-essential patents. *Journal of Economics & Management Strategy*, 27(3), 504–534.
- Baron, J., & Pohlmann, T. (2021). *Precision and bias in the assessment of essentiality rates in firms' portfolios of declared SEPs*. Mimeo.
- Baron, J., & Spulber, D. F. (2018). Technology standards and standard setting organizations: Introduction to the searle center database. *Journal of Economics & Management Strategy*, 27(3), 462–503.
- Bekkers, R., Bongard, R., & Nuvolari, A. (2011). An empirical study on the determinants of essential patent claims in compatibility standards. *Research Policy*, 40(7), 1001–1015.
- Bekkers, R., Catalini, C., Martinelli, A., Righi, C., & Simcoe, T. (2017). *Disclosure rules and declared essential patents* [NBER Working Paper No. 23627].
- Bekkers, R., Martinelli, A., & Tamagni, F. (2016). *The causal effect of including standards-related documentation into patent prior art: Evidence from a recent EPO policy change* [LEM Working Paper Series 2016/11].
- Bekkers, R., Tur, E. M., Henkel, J., van der Vorst, T., Driesse, M., & Contreras, J. L. (2022). Overcoming inefficiencies in patent licensing: A method to assess patent essentiality for technical standards. *Research Policy*, 51(10), 104590.
- Bekkers, R., & Updegrove, A. S. (2013). *A study of IPR policies and practices of a representative group of standards setting organizations worldwide*. Commissioned by the Committee on Intellectual Property Management in Standard-Setting Processes. National Research Council.
- Bekkers, R., & West, J. (2009). The limits to IPR standardization policies as evidenced by strategic patenting in UMTS. *Telecommunications Policy*, 33(1), 80–97.
- Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In Padó S. & Huang, R. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3615–3620). Association for Computational Linguistics.
- Berger, F., Blind, K., & Thumm, N. (2012). Filing behaviour regarding essential patents in industry standards. *Research Policy*, 41(2), 216–225.
- Brachtendorf, L., Gaessler, F., & Harhoff, D. (2020). *Approximating the standard essentiality of patents—A semantics-based analysis* [Report for the EPO Academic Research Programme].
- Chiao, B., Lerner, J., & Tirole, J. (2007). The rules of standard-setting organizations: An empirical analysis. *The RAND Journal of Economics*, 38(4), 905–930.
- Cho, R. L.-T., Liu, J. S., & Ho, M. H.-C. (2021). Exploring a patent's essentiality to the HEVC standard: A retrospective view. *IEEE Transactions on Engineering Management*.
- Contreras, J. L. (2016). When a stranger calls: Standards outsiders and unencumbered patents. *Journal of Competition Law & Economics*, 12(3), 507–540.

- Contreras, J. L. (2017a). Essentiality and standards-essential patents. In J. L. Contreras (Ed.), *Cambridge handbook of technical standardization law—Antitrust, competition and patent law*. Cambridge University Press.
- Contreras, J. L. (2017b). *TCL v. Ericsson: The first major U.S. top-down FRAND royalty decision* [University of Utah College of Law Research Paper No. 245]. <https://ssrn.com/abstract=3100976>
- Contreras, J. L. (2019). Technical standards, standards-setting organizations and intellectual property: A survey of the literature. In Menell, P. and Schwartz, D. (Eds.), *Research handbook on the economics of intellectual property law Vol. 2—Analytical methods*. Cambridge University Press.
- De Grazia, C. A. W., Pairolo, N. A., & Teodorescu, M. H. (2021). Examination incentives, learning, and patent office outcomes: The use of examiner's amendments at the USPTO. *Research Policy*, 50(10), 104360.
- deGrazia, C. A., Frumkin, J. P., & Pairolo, N. A. (2020). Embracing invention similarity for the measurement of vertically overlapping claims. *Economics of Innovation and New Technology*, 29(2), 113–146.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. <https://arxiv.org/abs/1810.04805>
- Dewatripont, M., & Legros, P. (2013). 'Essential patents', FRAND royalties and technological standards. *The Journal of Industrial Economics*, 61(4), 913–937.
- EC. (2014). *Patents and standards: A modern framework for IPR-based standardization*. Ref. Ares(2014)917720—25/03/2014.
- EC. (2016). *Transparency, predictability, and efficiency of SSO-based standardization and SEP licensing—A report for the European Commission*.
- EC. (2017). *Communication from the commission to the European Parliament, the council, and the European Economic and Social Committee: Setting out the EU approach to standard essential patents*. Brussels, 29.11.2017 COM(2017) 712 final.
- EC. (2019). *Making the rules—The governance of standard development organizations and their policies on intellectual property rights* [JRC Science for Policy Report].
- Farrell, J., Hayes, J., Shapiro, C., & Sullivan, T. (2007). Standard setting, patents, and hold-up. *Antitrust Law Journal*, 74(3), 603–670.
- Gaessler, F., Harhoff, D., & Sorg, S. (2019). *Bargaining failure and freedom to operate: Re-evaluating the effect of patents on cumulative innovation* [Max Planck Institute for Innovation & Competition Research Paper].
- Gątkowski, M., Dietl, M., Skrok, Ł., Whalen, R., & Rockett, K. (2020). Semantically-based patent thicket identification. *Research Policy*, 49(2), 103925.
- Gerken, J. M., & Moehle, M. G. (2012). A new instrument for technology monitoring: Novelty in patents measured by semantic patent analysis. *Scientometrics*, 91(3), 645–670.
- Goodman, D. J., & Myers, R. A. (2005). *3G cellular standards and patents*. IEEE WirelessCom.
- Hain, D. S., Jurowetzki, R., Buchmann, T., & Wolf, P. (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, 177, 121559.
- Hussinger, K., & Schiebacher, F. (2015). The market value of technology disclosures to standard setting organizations. *Industry and Innovation*, 22(4), 321–344.
- Jürgens, B., & Clarke, N. (2018). Study and comparison of the unique selling propositions of free-to-use multinational patent search systems. *World Patent Information*, 52, 9–16.
- Kang, B., & Bekkers, R. (2015). Just-in-time patents and the development of standards. *Research Policy*, 44(10), 1948–1961.
- Kang, B., & Motohashi, K. (2015). Essential intellectual property rights and inventors' involvement in standardization. *Research Policy*, 44(2), 483–492.
- Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3), 303–320.
- Kuhn, J. M., & Teodorescu, M. H. (2021). The track one pilot program: Who benefits from prioritized patent examination? *Strategic Entrepreneurship Journal*, 15(2), 185–208.
- Kuhn, J. M., & Thompson, N. C. (2019). How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business*, 26(1), 5–38.
- Larouche, P., & Zingales, N. (2017). *Injunctive relief in FRAND disputes in the EU—Intellectual property and competition law at the remedies stage* [Tilburg Law School Research Paper No. 01/2017]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2909708
- Leiponen, A. E. (2008). Competing through cooperation: The organization of standard setting in wireless telecommunications. *Management Science*, 54(11), 1904–1919.
- Lemley, M. A. (2002). Intellectual property rights and standard-setting organizations. *California Law Review*, 90, 1889–1980.
- Lemley, M. A., & Shapiro, C. (2006). Patent holdup and royalty stacking. *Texas Law Review*, 85, 1991–2048.
- Lemley, M. A., & Simcoe, T. (2019). How essential are standard-essential patents? *Cornell Law Review*, 104(3), 607–642.
- Lerner, J., Tabakovic, H., & Tirole, J. (2016). *Patent disclosures and standard-setting* [NBER Working Paper No. 22768].
- Lerner, J., & Tirole, J. (2015). Standard-essential patents. *Journal of Political Economy*, 123(3), 547–586.
- Liu, D.-R., & Shih, M.-J. (2011). Hybrid-patent classification based on patent-network analysis. *Journal of the American Society for Information Science and Technology*, 62(2), 246–256.
- Magerman, T., Van Looy, B., & Song, X. (2009). Exploring the feasibility and accuracy of latent semantic analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, 82(2), 289–306.

- Mallinson, K. (2017). *Do not count on accuracy in third-party patent-essentiality determinations*. Blog post at IP Finance. <http://www.ip.finance/2017/05/do-not-count-on-accuracy-in-third-party.html>
- Natterer, M. (2016). *Ähnlichkeit von Patenten: Entwicklung, empirische Validierung und ökonomische Anwendung eines textbasierten Ähnlichkeitsmaßes*. Verlag für Nationalökonomie, Management und Politikberatung.
- Picht, P. G. (2018). *FRAND determination in TCL v. Ericsson and unwired planet v. Huawei: Same Same But Different?* [Max Planck Institute for Innovation & Competition Research Paper No. 18-07]. <https://ssrn.com/abstract=3177975>
- Pohlmann, T., Neuhäusler, P., & Blind, K. (2016). Standard essential patents to boost financial returns. *R&D Management*, 46(S2), 612–630.
- Quint, D. (2014). Pooling with essential and nonessential patents. *American Economic Journal: Microeconomics*, 6(1), 23–57.
- Righi, C., & Simcoe, T. (2020). *Patenting inventions or inventing patents? Strategic use of continuations at the USPTO* [NBER Working Paper No. w27686].
- Royall, M. S., Tessar, A., & Vincenzo, A. D. (2008). Deterring patent ambush in standard setting: Lessons from Rambus and Qualcomm. *Antitrust*, 23, 34.
- Rysman, M., & Simcoe, T. (2008). Patents and the performance of voluntary standard-setting organizations. *Management Science*, 54(11), 1920–1934.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.
- Shapiro, C. (2001). Navigating the patent thicket: Cross licensing, patent pools, and standard setting. In A. B. Jaffe, J. Lerner, & S. Stern (Eds.), *Innovation policy and the economy* (Vol. 1, pp. 119–150). M.I.T. Press.
- Simcoe, T. S., Graham, S. J., & Feldman, M. P. (2009). Competing on standards? Entrepreneurship, intellectual property, and platform technologies. *Journal of Economics & Management Strategy*, 18(3), 775–816.
- Spulber, D. F. (2019). Standard setting organisations and standard essential patents: Voting and markets. *The Economic Journal*, 129(619), 1477–1509.
- Stitzing, R., Sääskilähti, P., Royer, J., & Audenrode, M. V. (2017). *Over-declaration of standard essential patents and determinants of essentiality*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2951617
- Tansey, R., Neal, M., & Carroll, R. (2005). “Get rich, or die trying”: Lessons from Rambus’ high-risk predatory litigation in the semiconductor industry. *Industry & Innovation*, 12(1), 93–115.
- Whalen, R., Lungeanu, A., DeChurch, L., & Contractor, N. (2020). Patent similarity data and innovation metrics. *Journal of Empirical Legal Studies*, 17(3), 615–639.
- Younge, K., & Kuhn, J. (2016). *Patent-to-patent similarity: A vector space model*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2709238

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher’s website.

How to cite this article: Brachtendorf, L., Gaessler, F., & Harhoff, D. (2023). Truly standard-essential patents? A semantics-based analysis. *Journal of Economics & Management Strategy*, 32, 132–157. <https://doi.org/10.1111/jems.12500>