DISCUSSION PAPER SERIES

DP14726

TRULY STANDARD-ESSENTIAL PATENTS? A SEMANTICS-BASED ANALYSIS

Lorenz Brachtendorf, Fabian Gaessler and Dietmar Harhoff

INDUSTRIAL ORGANIZATION

ISSN 0265-8003

TRULY STANDARD-ESSENTIAL PATENTS? A SEMANTICS-BASED ANALYSIS

Lorenz Brachtendorf, Fabian Gaessler and Dietmar Harhoff

Discussion Paper DP14726
Published 07 May 2020
Submitted 05 May 2020

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

TRULY STANDARD-ESSENTIAL PATENTS? A SEMANTICS-BASED ANALYSIS

Abstract

Standard-essential patents (SEPs) have become a key element of technical coordination in 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 exercises. In a first empirical application we illustrate the measure's usefulness in estimating the share of true SEPs in firm patent portfolios for several mobile telecommunication standards. We find firm-level differences that are statistically significant and economically substantial. Furthermore, we observe a general decline in the average share of presumably true SEPs between successive standard generations.

JEL Classification: L24, O33, O34

Keywords: patents, standards, standard essentiality, standard-setting organizations

Lorenz Brachtendorf - lorenz.brachtendorf@ip.mpg.de
Max Planck Institute for Innovation and Competition

Fabian Gaessler - fabian.gaessler@ip.mpg.de
Max Planck Institute for Innovation and Competition

Dietmar Harhoff - dietmar.harhoff@ip.mpg.de
Max Planck Institut for Innovation and Competition, Ludwig-Maximilians-Universität München and CEPR

Acknowledgements
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, and the ZEW/MaCCI Conference on the Economics of Innovation and Patenting for their valuable comments. Furthermore, 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. A previous version of this paper circulated as "Approximating the Standard Essentiality of Patents – A Semantics-Based Analysis."

Truly Standard-Essential Patents? A Semantics-Based Analysis

Lorenz Brachtendorf[a b] Fabian Gaessler[a] Dietmar Harhoff[a b c]

[a] *Max Planck Institute for Innovation and Competition, Munich*

[b] *Munich School of Management, Ludwig-Maximilians-University (LMU), Munich*

[c] *Centre for Economic Policy Research (CEPR), London*

May 5, 2020

ABSTRACT

Standard-essential patents (SEPs) have become a key element of technical coordination in 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 exercises. In a first empirical application we illustrate the measure's usefulness in estimating the share of true SEPs in firm patent portfolios for several mobile telecommunication standards. We find firm-level differences that are statistically significant and economically substantial. Furthermore, we observe a general decline in the average share of presumably true SEPs between successive standard generations.

KEYWORDS: patents, standards, standard essentiality, standard-setting organizations.

JEL Classification: L24, O33, O34

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, and the ZEW/MaCCI Conference on the Economics of Innovation and Patenting for their valuable comments. Furthermore, 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. A previous version of this paper circulated as "Approximating the Standard Essentiality of Patents – A Semantics-Based Analysis."

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 – in particular, if proprietary technologies are to be incorporated (Lerner and Tirole, 2015; Spulber, 2019). Standard-essential patents (SEPs) protect inventions that are part of technical standards. By definition, to avoid infringement any firm implementing the standard will require a license to all standard-related SEPs. 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.[1] 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, i.e., are truly standard-essential. However, there are several factors beyond technical merit that may influence whether a patent is declared standard-essential.[2] Most notably, there are concerns that firms declare patents to be standard-essential due to strategic reasons (Dewatripont and Legros, 2013).[3] Anecdotal evidence from policy reports and case studies strongly suggests that standard essentiality is not necessarily guaranteed by the patent holder's declaration (see Contreras, 2019, for an overview). In fact, the claim of standard essentiality frequently fails to survive scrutiny if the patent is disputed in court (Lemley and 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. With policy-makers interested in a fair and efficient framework for the development and adoption of technical standards, SEP essentiality checks have recently come into regulatory focus (EC, 2017).[4]

[1] See, for instance, the Communication from the European Commission in November, 2017 (EC, 2017).

[2] In this study, we focus on *technical* standard essentiality. We discuss different essentiality definitions in Section 2.

[3] 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.

[4] Several voices have suggested that patent offices should assess the standard essentiality of patents. Consequently, the Japanese Patent Office (JPO) offers a fee-based service comprising an advisory opinion on the standard essentiality of patents starting since April 2018.

This study introduces a semantics-based method to approximate the standard essentiality of patents. This method relies on a novel measure of semantic similarity between patents and standards and is scalable, objective, and replicable. In recent years, text-based measures have proven to be useful for the empirical assessment of similarity and technological relatedness between patents (e.g., Arts et al., 2018; Natterer, 2016; Younge and Kuhn, 2016). Extending this approach, we propose a method for a semantics-based comparison of patent texts and specifications of technical standards. In several validation exercises, we show that the calculated similarity can be used as the core variable to generate a meaningful approximation of standard essentiality. First, we investigate the semantic similarity of patent-standard pairs by comparing SEP declarations with control groups of patents in the same technology class and standard documents from the same standardization project. We observe a significantly higher semantic similarity for pairs of SEP-declared patents and related standards than for random pairs. Second, we replicate the study by Bekkers et al. (2017) regarding the effect of SEP declarations on the number of subsequent patent forward citations. We show that the magnitude of this 'disclosure effect' is considerably larger when focusing on subsets of SEP declarations with particularly high semantic similarity. We then employ a multivariate logit framework to construct a predictor of standard essentiality. This exercise uses manual engineering assessments of patents for three mobile telecommunication standards as they were utilized in the US court case of *TCL v. Ericsson*. Based on these data, we show that our similarity measure is an statistically significant and important predictor of the court-approved SEP assessments.

As recent legal disputes have exemplified, the calculation of licensing fees for standard technologies often involves not just single SEPs but whole SEP portfolios. As Contreras (2017a) 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.*" In a first empirical application, we therefore illustrate our method's usefulness to assess standard essentiality at the portfolio level. Extending our predictions to all declared SEPs of contributors to standards for mobile telecommunication (GSM, UMTS, and LTE), we estimate the share of (presumably) true SEPs in the respective firm patent portfolios. We document the high accuracy of our approach when predicting standard essentiality at the patent portfolio level. Considering the aggregated results, 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 roughly twice as large as the one for the lowest-ranked firm. Moreover, we observe a decline in the share of presumably true SEPs over the three successive generations of mobile telecommunication standards. We discuss possible explanations of this intriguing result in the paper.

So far, analysts seeking to ascertain the true status of an SEP-declared patents had only two choices: to take SEP declarations at face value or to rely on costly expert assessments.[5] By introducing a new method for approximating standard essentiality via an algorithm, this study makes various academic as well as practical contributions. First, we illustrate how a semantics-based tool can be used to measure the essentiality of patents for specific technical standards. The novel method is not just simple and inexpensive in use, it is also scalable, objective and replicable. Prior data on essentiality checks have required substantial technical knowledge and effort. The measure developed in this study, by contrast, can easily be applied to any large set of SEPs. This opens up new avenues of empirical research for scholars interested in standardization, patents, and firm strategy. For instance, the introduced method may help determine the present or historical population of over- as well as under-declared SEPs for a given 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 relationship between patent rights and standards. Section 3 details the methodology of our semantics-based approach. Section 4 then introduces the data used in the subsequent analyses. In Section 5 descriptive results that validate the method are provided. Finally a first use case on determining the share of (presumably) true SEPs in firm portfolios is presented in Section 6, followed by a brief discussion and outlook on further use cases of our essentiality measure.

2 Institutional Background and Prior Literature

2.1 Standard-setting organizations and SEPs

Technical standards typically incorporate a large number of complementary technological solutions owned by various organizations such as firms, research institutes, or universities. To lower transaction costs and gain efficiencies in the development and distribution of standardized technologies, SSOs coordinate the development of such standards (Contreras, 2019). SSOs differ in various dimensions such as technological focus, membership composition as well as policies and practices (Bekkers and Updegrove, 2013; Chiao et al., 2007; EC, 2019). One important and frequently studied aspect of SSO policies concerns the

[5] See, for instance, the study of Goodman and Myers (2005) and, most recently, Stitzing et al. (2017), both drawing on manual assessments of declared SEPs by patent attorneys and engineers. Further publicly available reports include SEP assessments by Cyber Creative Institute, Article One Partners, Jefferies and iRunway. With reference to potential subjectivity and bias in manual evaluations, essentiality assessments by technical experts are not universally considered credible (cf. Mallinson, 2017).

IP-related rules and regulations (Baron and Spulber, 2018; Lemley, 2002) with particular focus on the practiced licensing regime and the disclosure of SEPs.

Rules on the declaration of SEPs are SSO-specific and may address particular aspects, such as upfront patent searches, disclosure content, as well as disclosure timing, and may or may not be binding. For instance, some SSOs *demand* their members to disclose relevant intellectual property whereas other SSOs only *encourage* them to do so. Furthermore, firms may also be required to make reasonable efforts to search for potentially standard-essential IP. SSOs can also differ in terms of the necessary declaration content. At ETSI, for example, the specific disclosure of SEPs is mandatory whereas at other large SSOs, such as IEEE or ITU-T, blanket declarations are allowed. Similarly, requirements on the timing of disclosure might be interpreted as guidelines rather than strict obligations. Most SSOs specify rules that demand a timely disclosure either before the approval of the standard, as soon as possible, or upon an official call for patents. Breaching the duty to disclose relevant intellectual property rights may have serious economic and legal implications.

2.2 Declared SEPs and true standard essentiality

Patents that protect technological solutions required for the implementation of a particular standard are typically referred to as standard-essential patents (SEPs). The status of an SEP is commonly set through the rights holder's own declaration. In practice, however, the determination of standard essentiality proves challenging, and quite frequently, the question whether a patent is truly standard-essential needs to be solved in court.[6] Generally, technical standard essentiality is defined by the patent claims that cover a particular part of the technical standard. That is, the patent is standard-essential if the invention inherent to the implementation of the respective standard falls within the scope of the respective patent's claims. Beyond this definition, SSOs sometimes differentiate between technical and commercial essentiality. Whereas the former refers to purely technical aspects of the patented invention, commercial essentiality includes the additional consideration whether the patented invention is the only commercially feasible solution for the respective standard. Most SSOs focus on the technical essentiality, ETSI even explicitly rules out commercial factors when determining essentiality (Contreras, 2017a). Yet, standards describe a range of technical processes and solutions and may thereby refer to multiple patented inventions. Vice versa, patented inventions can be essential to more than one standard specification.[7] Consequently, the standard essentiality of a patent needs to be understood (and ultimately

[6] See Contreras (2017a) for a thorough summary of different concepts of essentiality, the legal issues arising from those, and the relevant case law on essentiality assessments.

[7] Multiple-Input-Multiple-Output (MIMO) is only one out of many examples for technologies that are part of several standards at different SSOs, as for instance IEEE's WiFi and the 3GPP standard LTE.

assessed) with regard to a particular standard.

Apart from this complex many-to-many relationship between patents and standards, a patent's standard essentiality status can also be time-variant. SSOs aim to include the best available technological solutions into a standard and thus often encourage the timely disclosure of patents covering even *potentially* standard-relevant technologies. Still, standards evolve over time, so that obsolete technologies are removed from the standard and replaced by more recent alternative technologies. Likewise, patent claims are not perfectly static either. 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, which likely affects standard essentiality.

At the time of disclosure, SEP declarations are typically neither verified nor challenged by the respective SSO. Presumably, this is due to cost and liability reasons. Given their non-binding nature, SEP declarations are also rarely withdrawn or updated after the finalization of the standard. As a result, they may represent a poor signal of true standard essentiality which typically remains private information held by the respective rights holder. However, a patent's true standard essentiality becomes public knowledge in some cases. First, results of standard essentiality assessments are disclosed through court decisions.[8] SEP litigation usually deals with selected subsets of SEPs rather than with entire SEP portfolios or, let alone, all SEPs for a particular standard.[9] Second, true standard essentiality of patents can be inferred from SEP assessments by third parties, which do not occur within the context of SEP lawsuits. The costs of such legally non-binding contractual essentiality assessments vary significantly depending on the evaluators' scrutiny.[10] Finally, some patent pools follow the practice to conduct standard essentiality assessments before they include a given SEP (Contreras, 2017a; Quint, 2014). Hence, patent pool inclusion can serve as a signal for true

[8] Although SEP litigation takes place in Europe as well (cf. Contreras et al., 2017), the US remain the hotspot for SEP litigation. Lemley and Simcoe (2019) provide evidence for the presence of non-essential SEPs in the context of SEP litigations before US courts. They examine SEPs brought to court and find, in particular, that SEPs held by non-practicing entities (NPEs) are less likely to be deemed infringed than a set of litigated SEP patents held by operating companies.

[9] The only exception is the recent lawsuit *Ericsson v. TCL* where a fairly large number of SEPs for the mobile telecommunication standards GSM, UMTS and LTE was assessed in order to determine fair, reasonable and non-discriminatory (FRAND) royalty rates.

[10] A report to the European Commission broadly differentiates between three confidence levels of essentiality (EC, 2014). Low-level assessments are estimated to cost around 600-1,800 EUR per patent (corresponding to 1-3 days of work). Industry studies that report on the essentiality of different samples of SEPs may be categorized into this low level assessment. The experts of these studies usually spend only a few hours per patent and would hence be even at the lower bound of this classification. Somewhat more detailed essentiality checks are conducted when patents are to be incorporated into a patent pool. Estimated costs are approximately 5,000-15,000 EUR depending on prior knowledge on the patent and on the number of claims to be assessed. Even more sophisticated assessments start at 20,000 EUR and comprise essentiality checks in the context of lawsuits on smaller subsets of SEPs.

standard essentiality, even though this again applies to a selected set of SEPs only.

2.3 SEPs and firm behavior

Holding patent rights for standard-essential technologies comes along with a range of direct and indirect benefits. SEPs represent revenue-generating opportunities as all standard implementers become potential licensees. Moreover, firms may improve their bargaining position in cross-licensing negotiations if they also hold SEPs.[11] Hence, it seems reasonable to assume that firms follow various strategies to increase the chance of holding standard-relevant patents. First of all, firms may decide to promote their own patented technologies for inclusion in a given standard through engagement in the standardization process.[12] Apart from that, firms may conduct what is commonly known as *just-in-time patenting* (Kang and Bekkers, 2015). Namely, firms intentionally file patents shortly before standardization meetings. The proximity in time allows those firms to increase the standard essentiality of the patented technology by aligning the patent's text to drafts of the standard description that are already in circulation. A similar pattern can be observed even after filing in the form of purposive patent amendments and patent continuations (Berger et al., 2012; Omachi, 2004). Firms tend to amend the claims of their pending patent applications to ensure that they align with the latest version of the standard.[13]

Firms usually enjoy some discretion in their decision whether or not to declare their patent as standard-essential. In this context, one can distinguish between the *over-declaration* and *under-declaration* of SEPs. The *over-declaration* of SEPs refers to the declaration of (ultimately) non-essential patent rights as SEPs. Reasons for over-declaration can be found in over-compliance with SSO disclosure obligations and opportunism. Patent holders may over-declare due to the evident asymmetry in potential sanctions. Typically, SSOs IP policies entail harsher punishments for not disclosing standard-essential patents rather than for disclosing standard-irrelevant patents (Contreras, 2017a). Moreover, SSOs often encourage patent holders to disclose not only patents that are essential, but also patents that *may become* essential to future versions of the standard. Here, the decision to disclose SEPs may be influenced by the patent holder's own opinion about which technological solution will prevail. More opportunistic reasons for over-declaration may lie in the firm's goal to increase licensing revenues and to secure freedom to operate (EC, 2013). The

[11] In fact, there is some empirical evidence that SEPs are on average more valuable (Rysman and Simcoe, 2008) and that SEP ownership correlates with financial performance (Hussinger and Schwiebacher, 2015; Pohlmann et al., 2016).

[12] In line with this, Bekkers et al. (2011) and Leiponen (2008) find that SSO membership and participation in the standardization process play an important role for technology selection. Furthermore, Kang and Motohashi (2015) find a positive correlation between inventor presence and the likelihood of SEP declaration.

[13] Berger et al. (2012) further find that such patents are also more likely to have a higher number of claims and longer grant lags, resulting from those changes to the patent application.

common practice of SEP counting in licensing agreements may incentivize such a behavior, since licensing revenues are often tied to the number of SEPs a firm holds (Dewatripont and Legros, 2013). This is particularly true for top-down approaches, which are frequently used when determining SEP royalty rates in court (Contreras, 2017a). Furthermore, a firm may inflate their SEP portfolio to gain leverage for cross-licensing deals with other SEP holders (Shapiro, 2001).

In contrast, *under-declaration* of SEPs refers to truly essential patents that remain undeclared. The failure to declare can be unintentional, as the patent holder may simply be unaware of its patents' relevance to a particular standard. However, under-declaration can also be the result of willful misconduct to benefit from hold-up situations. Here, patent holders deliberately keep their patents undisclosed up to the point of time when the standard is already implemented. The patent holder can then charge licensing fees, which are not bound to common royalty cap provisions, such as FRAND terms (Lemley and Shapiro, 2006).[14] There is little empirical evidence for under-declaration, but an often-cited example represents the case of Rambus.[15]Note that failing to timely disclose potentially essential patents can directly lead to antitrust liabilities. As a result, SSO policies that are supposed to counter under-declaration may in turn incentivize SEP over-declaration.

3 Methodology

In this section, we introduce a novel approach measuring semantic similarity between patents and technical standards. First, we briefly discuss the current state of the literature on semantic algorithms applied to patent text data and explain the peculiarities concerning the application of such algorithms to patents and standards. We then provide details on the mechanics of our approach and the resulting similarity measures.

3.1 Prior patent text-based measures

Text-based measures have become a popular tool in the empirical assessment of patent similarity (see Abbas et al., 2014, for an overview). Natterer (2016) developed a sophisticated semantic algorithm to search technologically closely related patents. In an application, he shows that similarity density measures are negatively correlated with patent value. The

[14]Depending on the jurisdiction, the patent holder may also be more likely to obtain injunctive relief against infringement if the patent remains undeclared (Larouche and Zingales, 2017). However, non-disclosed standard-essential patents may also be deemed unenforceable, as recently decided in *Core Wireless Licensing v. Apple Inc*.

[15]Rambus failed to disclose its relevant patents and patent applications during a standard-setting process at JEDEC, an SSO in the microelectronics industry. Rambus' subsequent royalty claims against locked-in manufacturers were quickly followed by legal disputes and anti-trust concerns.

author argues that patents with particularly high similarity to many other patents may be located in very dense technological subfields with increasing competitive pressure and therefore, may have lower economic value. Younge and Kuhn (2016) introduce a vector space model to measure patent-to-patent similarity and provide details on significant improvements upon current patent classification schemes. More recently, Arts et al. (2018) used text similarity to measure the technological relatedness between patents and applied their novel approach to prior empirical findings on the localization of knowledge spillovers.

So far, all these applications were restricted to texts within the patent universe. A notable exception is the early study by Magerman et al. (2009). Here, the authors use vector space models and latent semantic indexing to detect similarities between the patents filed and the scientific publications written by a small set of academic inventors. To the best of our knowledge, measuring the similarity between patents and standards has not yet been explored on a scientific and systematic basis.

3.2 Mechanics of the approach

We rely on a sophisticated and field-proven text-mining algorithm to measure the semantic similarity between patents and standards.[16] The algorithm has been specifically developed to handle patent *as well as* patent-related texts and incorporates various text pre-processing techniques and automatic language corrections. The algorithm incorporates various techniques of natural language processing, such as part-of-speech tagging, spelling correction, n-grams, stop words, stemming techniques, entropy-based weighting, and synonym dictionaries. In line with other text-mining algorithms, a vector space model is employed to calculate the semantic similarity between two defined texts. The algorithm measures the similarity between patents, but can also be used to measure the similarity between patents and any other input text (such as product specifications, scientific publications, Wikipedia articles, etc.). Its major advantage is the very efficient implementation which allows for a comparison of any (large) text to the patent universe and yields a list with the most similar patents ranked by their similarity score.[17] Due to performance purposes, semantic similarity scores are integers and scaled between 0 and 1,000. Similarity scores of 0 mean that the two input texts have nothing in common whereas scores of 1,000 imply that they are next to identical.

[16]The algorithm is part of a commercial tool that has been developed by octimine technologies GmbH (now: Dennemeyer Octimine GmbH). The primary use case of this tool is the search for closely related prior art. See Jürgens and Clarke (2018) and Natterer (2016) for more information.

[17]Note that similarity is measured at patent family level, with the most recent publication of a (granted) patent family member used as text input. Only EP, US, WO and DE publications are considered (in this order). German text is machine translated into English. Note that the latest publication is the one most relevant for SEP enforcement.

For illustration purposes, we provide an example of a patent-standard pair with evidently high text similarity. The selected example for a standard is 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 standards projects. According to our semantic 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 to the respective standard specification on June 18, 2009, and appears to have a particularly high textual similarity to the standard. In Figure 1, we exemplarily contrast parts of the technical specification with an excerpt from the patent description. Similar and identical words are highlighted to illustrate the semantic similarity of both.[18]

Figure 1: Text similarity between patents and standards

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* $f(j)$, $i = 0,\ldots,l_{dtx}-1, j=0,\ldots,l_{dtx}-1, i \neq j$ *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* $j=0,\ldots,l_{dtx}-1, j \neq i$ *within the CN averaging period* (l_{dtx})."

In line with the previous literature on text-based similarity between patents, we interpret the semantic similarity between patents and standards as a measure of *their* technological similarity. We consider this a valid extension for the following reasons. First, patent texts as well as standard specifications are highly technical texts and are reasonably comparable to each other as illustrated by the above example. Second, standard documents are utilized by patent examiners, patent attorneys and inventors alike, which underlines their role as

[18] If we deliberately exclude similar terms (e.g., the highlighted parts in the figure above) from the standard text, the measured similarity between standard and this specific patent decreases considerably. This demonstrates that semantic similarity is mostly driven by such technologically similar sections.

informative technology descriptions.[19] In Section 5, we provide evidence for the validity of patent-standard text similarity as a measure of technological similarity and ultimately standard essentiality.

The used text-mining algorithm is proprietary, which renders some aspects of the similarity calculation non-transparent and also complicates replication. To illustrate the general feasibility of semantic algorithms for measuring patent-standard similarity, we apply straightforward techniques implemented in freely available text-mining packages in R and Python. The results based on these open-source algorithms are comparable, yet remain inferior to our similarity measure, in particular for very large text data. Details on this technical exercise can be found in the Online Appendix D.

3.3 Similarity measures

In the following analyses, we apply two different measures to approximate the true standard-essentiality of a patent: 1) the *similarity score* as an absolute value calculated by the algorithm, and 2) the *similarity rank*, which represents the focal patent's rank relative to all other patents in the patent universe (ordered by their *similarity score*). The measures are strongly correlated with each other and can be used individually to quantify patent-standard similarity. However, there are some subtle differences how to interpret them. Whereas the former can be considered as a measure independent from other patents and comparable across standards, only the *similarity rank* provides the standard-specific order of the most similar patents. For each standards text, both similarity measures are retrieved for the most similar 3,000 patent families. Although this allows us to limit the amount of data, it also introduces the need to account for truncation (or censoring, respectively) when interpreting our results.

[19] For instance, Bekkers et al. (2016) find that standard documentations contain relevant prior art that is used to assess a patent's novelty during examination.

4 Data and Descriptives

In this section, we first describe the used data and then provide selected descriptive statistics.

4.1 Data

Standard documents and SEP declarations

We employ two distinct datasets provided by the European Telecommunication Standards Institute (ETSI). ETSI has been established more than thirty years ago and is one of the most important standard-setting organizations in the ICT sector. The most successful standards in telecommunication such as DECT, TETRA, GSM, UMTS, LTE and most recently 5G have been set either directly by ETSI or within the framework of the 3rd Generation Partnership Project (3GPP).[20] In terms of the absolute number of declared SEPs, ETSI is by far the largest and most important SSO (Baron and Pohlmann, 2018).

ETSI's IPR database provides detailed information on SEP declarations submitted during the standardization process. Firms and other organizations involved in the standard setting process at ETSI are obliged to make their relevant IPR available. In declaration letters, they disclose information on their relevant patents with regard to particular standards. The level of detail in such declaration letters varies substantially. Whereas some declarations only cite the overall standards project, most others specify the relevant technical specification (TS) and – to some extent – even the specific version of the standard. The IPR data can be readily downloaded and provides most of the information on declarations as listed on the ETSI website.[21]

In addition to the information on declared SEPs and their relevance for standards, the second ETSI database provides details on technical standards. We focus on documents of standards that have been approved and published by ETSI. As of November 11, 2016, the online standards database stores 40,461 documents. The vast majority of documents is available in the portable document format (PDF), is therefore machine readable and can immediately be used for further analyses.[22] The major part of the documents refers to European standards (EN) and technical specifications (TS) for the different generations of mobile telecommunication standards: GSM, UMTS, and LTE. The set of documents covers

[20] 3GPP is a global network of seven standards organizations of which ETSI is one of the key organizations.

[21] As a matter of fact, some declarations are even more fine-grained and indicate the specific sections, figures and tables to which the patent is deemed essential. This information is not part of the IPR data, but can be found on the ETSI website. We retrieved this and further information (e.g., the person responsible for declarations within the organization) and merged them to the IPR database.

[22] However, roughly 9% of these files are encrypted or cannot be accessed for other technical reasons.

all releases and all versions of the approved standards, depicting the evolution of standards over time.

Standard documents are quite distinct documents in several aspects. They provide guidelines on the technologies implemented in a standard in a very detailed and structured manner. Standard documents published by ETSI typically start with the table of contents, references, definitions and abbreviations, followed by the main content, and end with the annex as well as the version history. The length of such documents varies substantially. The average number of pages for all 40,461 documents is 129 pages (median: 44) with some documents comprising thousands of pages. For the subset of standards which are cited in SEP declarations, the average page number is 194 (median: 84) and hence even larger. However, SEPs typically refer to very specific parts within the technical specifications. It should be evident that a semantic comparison of patents with full standard documents comes with considerable noise which may compromise our predictions. Making use of the structured format of standard documents, we developed a routine that automatically identifies the table of contents of a standard document and then compartmentalizes the document into chapters, sections and subsections as stated in the table of contents of the document. Using string matching and similarity metrics, we are able to identify the text of all sections in a structured manner.[23] This allows us to make precise comparisons between patents and specific standard specifications. For the sample of machine-readable documents, we identify 446,666 unique standard document chapters. To keep the task computationally feasible, we restrict the semantic analyses on chapter specific texts to subsamples of all standard documents.

Patents

On patent side, the algorithm draws on full text information, which includes the title, abstract, claims and description of a patent document. Text information is obtained from the databases of the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO) and the World Intellectual Property Organization (WIPO). In total, full text information for approximately 37 million patent documents is used.

We further add bibliographic information on the patents from PATSTAT (autumn 2017 version).[24] We retrieve information on patent families, technology classes, inventor team size, co-applications as well as detailed information on patent claims and furthermore compute various forward and backward citation measures on patent family level necessary for our validity checks.

[23] To this end, we use edit distance functions such as the restricted Damerau-Levenshtein distance.

[24] The Worldwide Patent Statistical Database PATSTAT from the European Patent Office (EPO) covers the entire history of patents worldwide and provides bibliographic information such as patent and inventor information.

Similarity data

All standards that are referenced in SEP declarations are identified which leaves us with a set of 4,796 referenced standard documents. The semantic algorithm described in the previous sections is then used to compare those documents to the approximately 37 million patent documents from the patent database.

We generate two datasets on the similarity between patents and standards. The first dataset includes the 14,388,000 pairs of patent families and standards. Here, the calculation of the similarity scores is done at the *document level*. The second dataset includes a more fine-grained comparison between patents and standards at the *chapter level*. For 4,500 of the 4,796 standard documents, our routine was able to identify the table of contents and to extract the relevant chapters. The compartmentalization of these documents yields a total of 62,482 chapters. Generating the similarity scores for those texts results in 187,398,000 observations at the patent-standard level.

4.2 Sample description

In Table 1, we report summary statistics for the two similarity measures (*similarity score* and *similarity rank*) based on full text as well as chapter-specific data of the standard documents. The measures reveal some distinct differences in similarity across different samples of patent-standard pairs. We provide statistics on all patents and SEPs, where patent-standard pairs are endogenously determined by the highest similarity. Furthermore, we provide statistics on SEP declarations, where patent-standard pairs are predefined. We observe notable differences in the measured similarity. The average *similarity score* of SEPs to their most similar chapters is 377 whereas the average in the full sample of patent-chapter pairs is 216.

Figure 2a shows the similarity score distributions for all patents and the subset of all SEPs.In Figure 2b, the *similarity rank* distribution of SEPs illustrates that SEPs are among the highest ranked patent-standard pairs. Notably, about one third of all SEPs that were declared at ETSI are among the top 20 for the corresponding standards text. Similarly, in Figure 2c, the percentage of SEPs declared at ETSI is plotted against the rank reporting the two samples of SEPs that are included in the chapter dataset (blue line) and full text datasets (red line). For the former, we observe 86% of declared SEPs within the top 3,000 patent families whereas for the sample with full text documents only 66% are observed. Roughly 48% are included within the top 100 patents for chapter, but only 22% for full text information. Altogether, this strongly indicates that comparisons are more precise when shorter texts, i.e., chapters, are used in the analyses.

Table 1: Summary statistics: Similarity data

Sample	Variable	Mean	SD	SE	Min	Max	N
Document level							
All	Score	218	67	0.018	62	818	14388000
	Rank	1500	866	0.228	1	3000	14388000
SEPs	Score	315	96	0.907	71	818	11311
	Rank	926	933	8.774	1	3000	11311
Chapter level							
All	Score	216	69	0.005	37	945	187397890
	Rank	1501	866	0.063	1	3000	187397890
SEPs	Score	377	113	0.935	48	817	14713
	Rank	663	838	6.906	1	3000	14713

Notes: Summary statistics for *similarity score* and *similarity rank* (at document and chapter level) for all patents and the subsample of SEPs. Minimum (maximum) possible score: 0 (1,000). Lowest (highest) possible rank: 3,000 (1).

5 Validation Results and Predictions

We conduct three distinct validation exercises.[25] First, we investigate the technological similarity between patents and standards by comparing SEP declarations with control groups of patents and standards in the same technology class and the same standards project. Second, we replicate the study by Bekkers et al. (2017) about the (positive) effect of SEP declarations on the number of subsequent patent forward citations. Here, we show that the magnitude of this 'disclosure effect' is considerably larger when focusing on subsets of SEP declarations with high similarity ranks. Third, we benchmark our results with a dataset of manually examined SEPs for the mobile telecommunication standards GSM, UMTS and LTE. Based on these data, we test the predictive power of our novel semantics-based similarity measure to determine true standard essentiality.

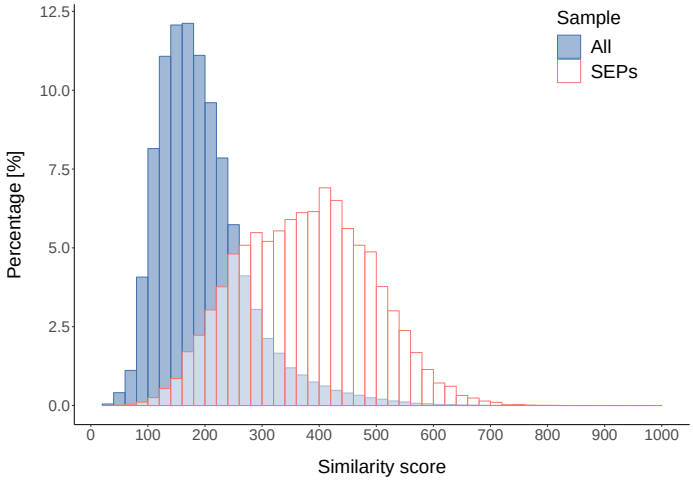
5.1 Comparison of declared SEPs with control groups

The first step to validate our semantic approach involves a comparison of declared SEPs with patents describing technologies from the very same technology class. If our measure has any explanatory value, declared SEPs will be significantly more similar to the respective

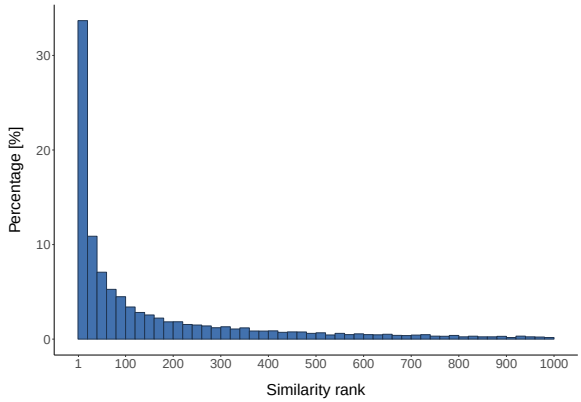
[25] Furthermore, in the Online Appendix C we estimate multivariate (OLS) regression models in which we regress the semantic similarity measure on various patent characteristics and compare our results qualitatively with those reported by Stitzing et al. (2017).

Figure 2: Distribution of SEPs in similarity dataset

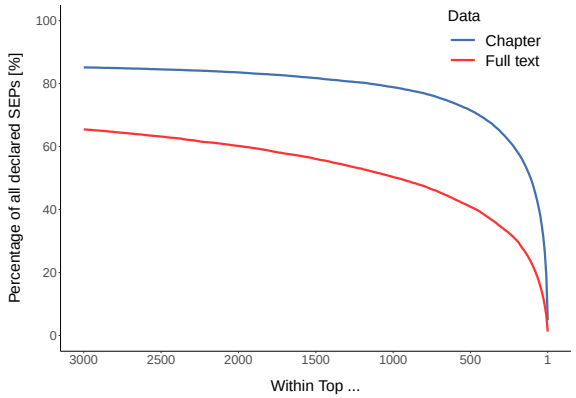
(a) Similarity score distribution: All patents vs. SEPs



(b) Rank distribution for SEPs



(c) Aggregate share of SEPs by rank



Notes: The top figure shows the similarity score distribution for two different sets of patents. All patents in the full sample (blue bars) are compared to the set of SEPs declared at ETSI (white bars). The bottom left-hand graph shows the *similarity rank* distribution for SEPs at chapter level. The bottom right-hand graph shows the aggregate share of SEPs by *similarity rank* at the chapter (blue line) and document level (red line).

standard than the control patents.[26] We exploit the information that SEP declarations at ETSI usually cite the respective standard. We name these predefined pairs of declared SEPs and standards *SEP declarations* and compare those to pairs of the same standard and undeclared patents from the same technology class and cohort. To this end, we select patents with the same CPC-4 codes (e.g., one of the most common technology classes is the *H04W 72* class for local resource managements in wireless communications networks)

[26]As discussed in previous sections, many declared SEPs at ETSI may under scrutiny turn out to be non-essential for the referenced standard. We still expect that the full sample of declared SEPs is significantly more similar to the respective standards as compared to control patents due to the set of correctly declared and hence truly essential patents. We note that the control group comparison with all SEPs renders the average difference in similarity a lower bound.

and same patent priority year. Furthermore, we only take into account patent families that have at least one US or EP publication. Control patents are randomly chosen from this pre-selected group of patents.

Vice versa, we now hold the declared SEP fixed and compare the associated standard document to another randomly chosen standard document from the same ETSI standards project[27] and the same publication year. Selecting the most similar chapter for each patent, we observe 15,000 SEP-standard document pairs (*SEP declarations*) in our data. As explained before, we only observe the 3,000 most similar patent families for each chapter of each standard document cited in SEP declarations and therefore have to deal with either truncation or censoring. Restricting the truncated sample to those patent families with at least one US or EP patent family member, we obtain a total of 29,380 treated and control patents. Note that the control patent is not necessarily within the set of the 3,000 most similar patent families. In this case, we conservatively assign the lowest observed similarity value for the given standard to the control patent. This likely results in a considerable overestimation of similarity scores for control patents.[28]

Figure 3 compares the distribution of similarity scores for each group. On the left-hand side, SEPs are compared with control patents. The mean difference in similarity scores is approximately 59 points. On the right-hand side, standards that are referenced in SEP declarations are compared with control standards. Here, the mean difference in similarity scores is about 135. All differences are statistically significant with t-values greater than 60 (Table B-1 in the Appendix reports the corresponding t-statistics). To summarize, the results of our control group comparison strongly suggest that semantic approaches are appropriate to measure technological similarity between patents and standards.

5.2 Replicating the ETSI 'disclosure effect'

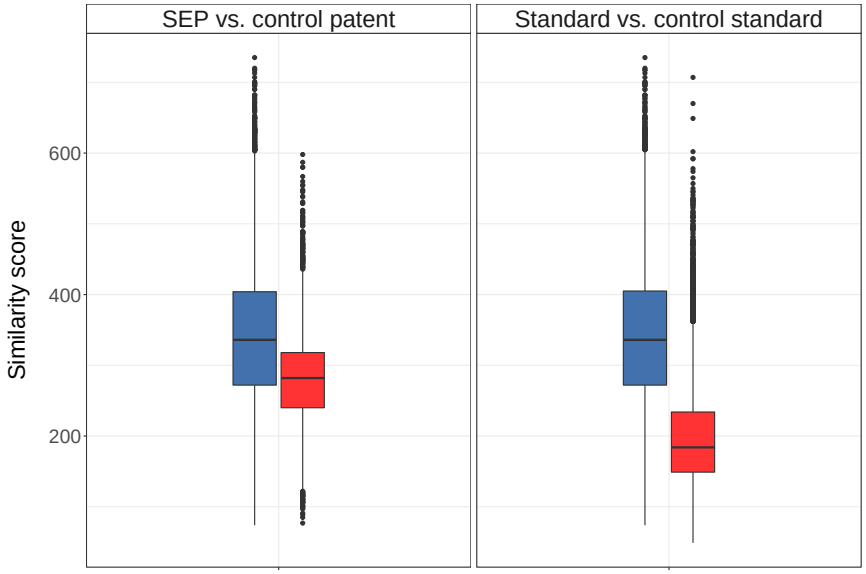
In the second validation exercise we replicate the study of Bekkers et al. (2017) and re-estimate the 'disclosure effect' of SEP declarations on patent forward citations. Bekkers et al. (2017) propose that the disclosure of SEPs should lead to an increase in patent forward citations, reflecting the gain in economic value after the implementation of the patented technology into a standard. While they find this to be true for various other SSOs, the estimated effect is *negative* for SEPs declared at ETSI.[29] Consequently, ETSI may have a

[27] We classify standard documents based on keywords occurring in the title of the standard document. We differentiate between the following groups of standards: LTE, UMTS, GSM, DECT, TETRA, DVB, DAB, ISDN, or any other standard.

[28] We obtain similar results using censored data for both SEPs and controls (see Figure A-1 in the Appendix).

[29] The authors explain this surprising finding with ETSI's special IPR 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.

Figure 3: Comparison of SEP - standard pairs with control groups



Notes: The box plot on the left-hand side shows the difference in similarity scores of SEP declarations (blue) and control patents compared to the same standard (red). On the right-hand side, similarity scores of SEP declarations (blue) are compared to similarity scores of the same SEP and control standards (red). Statistics are shown in Table B-1.

high share of declared SEPs that are in fact never implemented in a standard and therefore not truly essential.

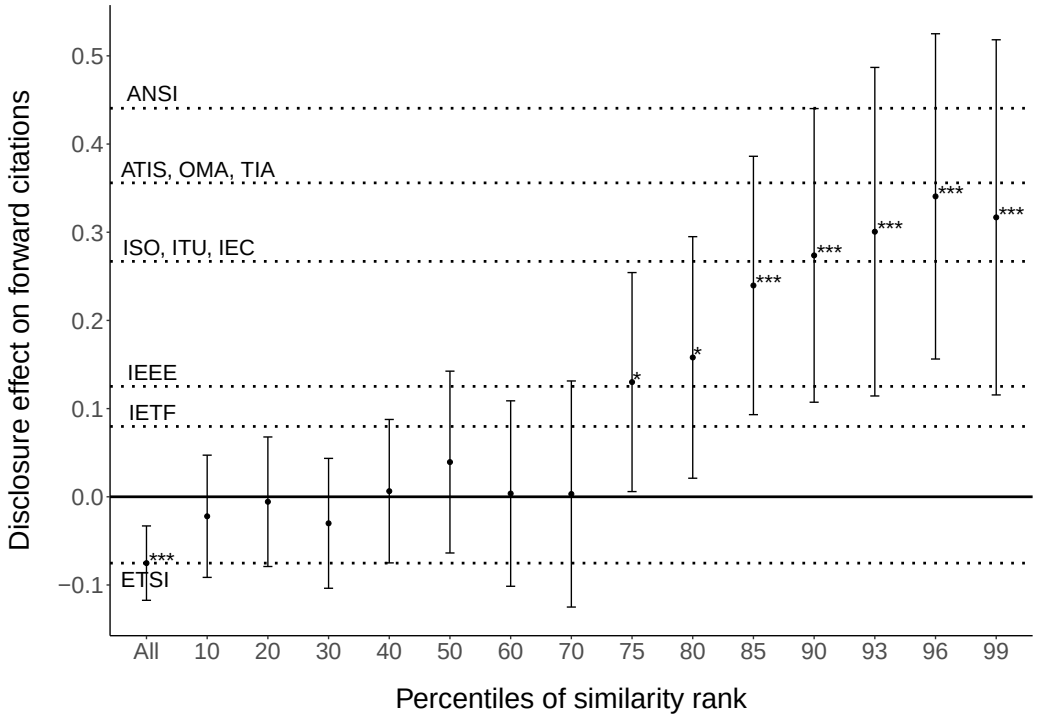
Using our novel measure, we can identify those declared SEPs which are particularly similar to their associated standards relative to other patents. Based on the assumption that the similarity is particularly high for patents which are in fact implemented in a standard, we expect a positive disclosure effect for such a selection of declared SEPs.

To that end, we link our data on semantic similarity to the authors' dataset on declared SEPs, which is publicly available as the 'Disclosed Standard Essential Patents (dSEP) Database'. We identify 1,183 SEPs declared at ETSI that are among the most similar patents for their associated standards. We borrow the empirical design by Bekkers et al. (2017) using a difference-in-differences approach, in which technologically similar patents with the same citation pre-trend before the SEP declaration serve as control patents. We use a Poisson regression model to estimate the following equation:

$$cites_{it} = \sum_{j} PostDisclosure_{ijt}\beta_{ij} + \alpha_i + \gamma_{ay} + \epsilon_{it}. \quad (1)$$

The dependent variable $cites_{it}$, measured on patent-year level, is the count of forward citations received by subsequent patents. The independent variable of interest $PostDisclosure_{ijt}$

Figure 4: Positive disclosure effects of the highest ranked SEPs at ETSI



Notes: This figure shows the estimates of the disclosure effect on forward citations based on the full sample (*All*) and several subsamples defined by the respective percentile of the similarity rank distribution within the full sample. Poisson estimates and 90% confidence intervals are shown. Each point corresponds to a separate regression coefficient estimated as shown in Equation 1. Standard errors are clustered on patent level. Significance levels: * p<0.1, ** p<0.05, *** p<0.01. The dotted horizontal lines reflect the effect sizes at other SSOs, as measured by Bekkers et al. (2017).

is a binary variable that is equal to 1 for each year t after and 0 before SEP declaration to SSO j. Apart from ETSI, the other organizations considered in the analysis are ANSI, IEEE, IETF and the combined groups of ATIS, OMA and TIA as well as ISO, ITU and IEC.

Figure 4 shows Poisson estimates for the effect of disclosure on forward citations. Following the econometric approach by Bekkers et al. (2017), we indeed observe a negative disclosure effect for the full set of SEPs. However, the estimated effect increases substantially for higher percentiles of SEPs by *similarity rank*. Strikingly, at percentiles of 75 and above, the effect sizes become comparable to those which Bekkers et al. (2017) estimated for SEP declarations at other SSOs. We consider this strong evidence for the identification of truly essential patents through our similarity measure.

5.3 Benchmark against manual SEP assessments

In the third validation exercise, we make use of a dataset of manually examined SEPs and test the predictive power of our similarity measure to determine a patent's true standard essentiality. In the following, we briefly introduce the dataset of manual SEP assessments and subsequently present the validation results.

Data origin and overview

The dataset we use was developed by an IP consulting firm involved in the major patent lawsuit *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.[30] The case concerned the calculation of royalty fees for SEPs, but also addressed the question of how many declared SEPs are truly essential for GSM, UMTS and LTE standards. The plaintiff (TCL) recruited the IP consulting firm to assess the essentiality of a selected sample of declared SEPs. This subsample comprises one-third of all SEPs declared for user equipment (UE) standards. Engineers manually evaluated those patents using the respective standard specifications on UE. The experts' essentiality assessments were criticized during the case because of the relatively short time they spent on each patent. In turn, a smaller subsample of patents was cross-checked by an independent expert, who – despite of false positives as well as false negatives – found overall very similar results. The evaluations were ultimately confirmed and accepted in court. We therefore believe that the results are a strong indication for true standard essentiality.

Validation regressions

To validate our measures of semantic similarity, we use logistic regression to predict standard essentiality. We regress the manual SEP assessments on semantic similarity measures using various specifications.[31] Essentiality assessments are reported as binary decision with 1 being actually essential and 0 representing non-essential patents for a corresponding standard. Approximately 36% of patent families were found to be essential for LTE, 40% for UMTS and 39% for GSM standards.[32] The main variable of interest is the *similarity score*, which we report for pairs of patent families and the semantically most similar standard document in the sample. Additionally, several patent characteristics are shown. The number of forward citations is computed on US patent family level. *Length claim 1* refers to the

[30] An elaborate discussion of this case and the decision can be found in Contreras (2017b) and Picht (2018).

[31] Table B-2 in the Appendix provides summary statistics for the full sample of 2,541 evaluated patent families.

[32] This is also within the range of other experts' evaluations such as PA Consulting (35%), Goodman/Myers (2010: 50%) or Cyber Creative Institute (2013: 56%), which all vary in terms of the applied level of scrutiny.

number of words in the first independent claim. Furthermore, the variable *Section-specific declaration* indicates whether the declared SEP cites specific sections, tables or figures of a particular standard document.

In Table 2, we report logistic regression results for correlations between the similarity measure as independent variable and the manually assessed LTE standard essentiality as dependent variable. We find positive and statistically significant correlations for the measure of similarity in all specifications. In the specification without fixed effects in column (2), the effect size for a one standard deviation increase in similarity score (roughly corresponding to 100 points in our data) is 7.8 pp. This coefficient is remarkably similar to the one estimated in the full specification in column (5), which includes controls for patent priority year, declaration year, technology class, technical specification and firm fixed effects. Including this full set of fixed effects alleviates the concern that the correlation of the similarity score with standard essentiality merely reflects different wording styles over time, technologies, standards or patent holders. In fact, we can confirm that our measure has explanatory value even *within* firm SEP portfolios. We further find significant correlations for the length of the first claim suggesting that patents with shorter, i.e., broader, claims are more likely to be essential. The number of citations received from SEPs declared at ETSI are positively correlated with standard essentiality.

We can corroborate the relationship between our similarity measure and standard essentiality for GSM and UMTS standards (see Table B-3 in the Appendix). Although the subsamples of patents evaluated by technical experts are considerably smaller, we again observe statistically significant correlations that are highly similar to our results for LTE patents. If anything, the effect sizes appear to be even larger for UMTS and GSM standards. A one standard deviation increase in similarity scores corresponds to a 15.3 pp increase in essentiality for patents relevant for GSM standards and 14.8 pp for patents relevant for UMTS standards.

To validate predictions of the semantic similarity measure, we consider the sample of LTE patents and employ 10-fold cross validation for all of our predictions. Using weighted precision and recall metrics in a logistic regression setup while confining to simple similarity scores, we obtain scores of 61% and 64%, respectively. Once we control for patent characteristics, precision and recall scores increase to 63% and 65%, respectively. The inclusion of additional patent characteristics therefore does not seem to improve predictions by much.[33] Furthermore, we split the sample of patents evaluated for the LTE standard into a test and training dataset. 70% of the data are used for training and 30% to test our model.[34] These test and training datasets are used in the subsequent SEP portfolio estimations.

[33] We report regression results and discuss the relationship between various patent characteristics and the similarity score in Online Appendix C.

[34] We report the confusion matrix for the test set of 402 SEPs for LTE standards in Table B-4 in the Appendix.

Table 2: Logistic regressions: LTE standard essentiality

	(1)	(2)	(3)	(4)	(5)
Similarity score		0.0738***	0.0501***	0.0461***	0.1032**
		(0.0135)	(0.0162)	(0.0172)	(0.0448)
SEP transferred (d)	−0.1083**	−0.0826	−0.1210*	−0.1350*	−0.1131
	(0.0514)	(0.0534)	(0.0713)	(0.0711)	(0.1385)
# Independent claims	−0.0025	−0.0001	0.0009	0.0022	−0.0079
	(0.0044)	(0.0045)	(0.0050)	(0.0051)	(0.0108)
Length claim 1	−0.0006***	−0.0006**	−0.0006**	−0.0005*	−0.0006
	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0005)
# Inventors	−0.0149*	−0.0116	−0.0210**	−0.0198*	−0.0096
	(0.0086)	(0.0087)	(0.0100)	(0.0103)	(0.0181)
# Applicants	0.0020	0.0037	0.0070	0.0087	−0.0123
	(0.0079)	(0.0079)	(0.0088)	(0.0090)	(0.0145)
Patent family size	0.0040**	0.0042**	0.0043**	0.0055**	0.0077
	(0.0017)	(0.0017)	(0.0021)	(0.0023)	(0.0051)
# Patent references	−0.0004	−0.0001	−0.0001	−0.0001	−0.0012
	(0.0004)	(0.0004)	(0.0005)	(0.0005)	(0.0008)
# NPL references	0.0007**	0.0006*	0.0008*	0.0007	0.0012
	(0.0003)	(0.0003)	(0.0004)	(0.0005)	(0.0008)
# SEP US fwd. cit. (5yrs)	0.0051***	0.0038***	0.0029**	0.0037**	0.0022
	(0.0013)	(0.0013)	(0.0014)	(0.0015)	(0.0023)
Section-specific decl. (d)	0.0975***	0.0935***	0.0869	0.0811	0.3076***
	(0.0293)	(0.0295)	(0.0537)	(0.0568)	(0.0977)
Priority year	No	No	Yes	Yes	Yes
Earliest decl. year	No	No	Yes	Yes	Yes
Firm FE	No	No	Yes	Yes	Yes
CPC-4 FE	No	No	No	Yes	Yes
TS FE	No	No	No	No	Yes
Pseudo R^2	0.04	0.06	0.14	0.16	0.25
AUC	0.64	0.67	0.74	0.76	0.81
Observations	1,290	1,290	1,290	1,290	674

Notes: The dependent variable is a dummy equal to one if the patent family is truly essential for LTE standards. AUC = Area under ROC-Curve. Pairs of SEPs and their most similar standard are selected in the full sample. Similarity scores are divided by 100. Marginal effects of one unit change are reported. For binary variables (d) following the variable name indicates a discrete change from 0 to 1. The sample size drops substantially when fixed effects for technical specifications (TS) are included in the model. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

6 Predicting SEP Portfolio Shares

We use the data from Section 5.3 to predict SEP portfolio shares, i.e., for a given firm the share of declared patents which (we presume to be) truly standard-essential. While our predictor is somewhat noisy at the level of individual patents, the prediction errors would partially cancel out at the portfolio level as long as patents are independent of each other.[35] Based on the logistic regression results, we compute the predicted probabilities of standard-essentiality for a given patent in order to estimate the true share of SEPs $\widehat{P}_F$ on firm-level with the following equation:

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

where n is the number of patents for a given firm F and X_{ij} represent the explanatory variables used in the logistic regression. We confine regressors to those that have shown statistically significant correlations with true essentiality in the case of LTE standards (cf. column (1) in Table B-3 in the Appendix): the semantic similarity score, SEP US forward citations (5yrs), a dummy for section-specific declarations, the number of NPL references and the length of the first independent claim. The regression results are shown in Table B-5 in the Appendix.

We draw random portfolios from the test dataset on LTE patents to determine the error of our prediction, on an aggregated level, as a function of the number of patents in the portfolio.[36] First, we compute the predicted probabilities for the test sample based on the logistic regression results from the training dataset. 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 5 plots these differences in predicted and actual shares of true SEPs against the size of the patent 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. We therefore fit a power law function to the

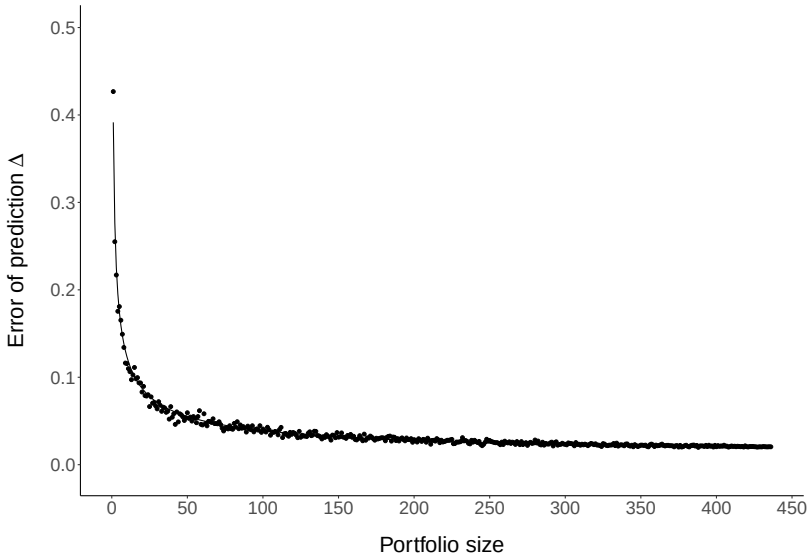
[35] For a discussion of this assumption, see Gambardella et al. (2017).

[36] 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 from portfolio composition and hence provide a general, firm-independent function.

data. The following fitted function describes the error rate for LTE patents:[37]

$$\widehat{\Delta}(N) = \hat{\alpha}\, N^{-\hat{k}}, \quad \text{where} \qquad (3)$$
$$\hat{\alpha} = 0.3916 \quad (\pm\, 0.0025),$$
$$\hat{k} = 0.5008 \quad (\pm\, 0.0019).$$

Figure 5: The error of prediction as a function of portfolio size (LTE)



Notes: The error of prediction Δ is plotted as a function of portfolio sizes where portfolios are randomly drawn from the test sample. Additionally, a non-linear least squares fit is shown for the test sample of LTE patents. The fitted function is a power law function.

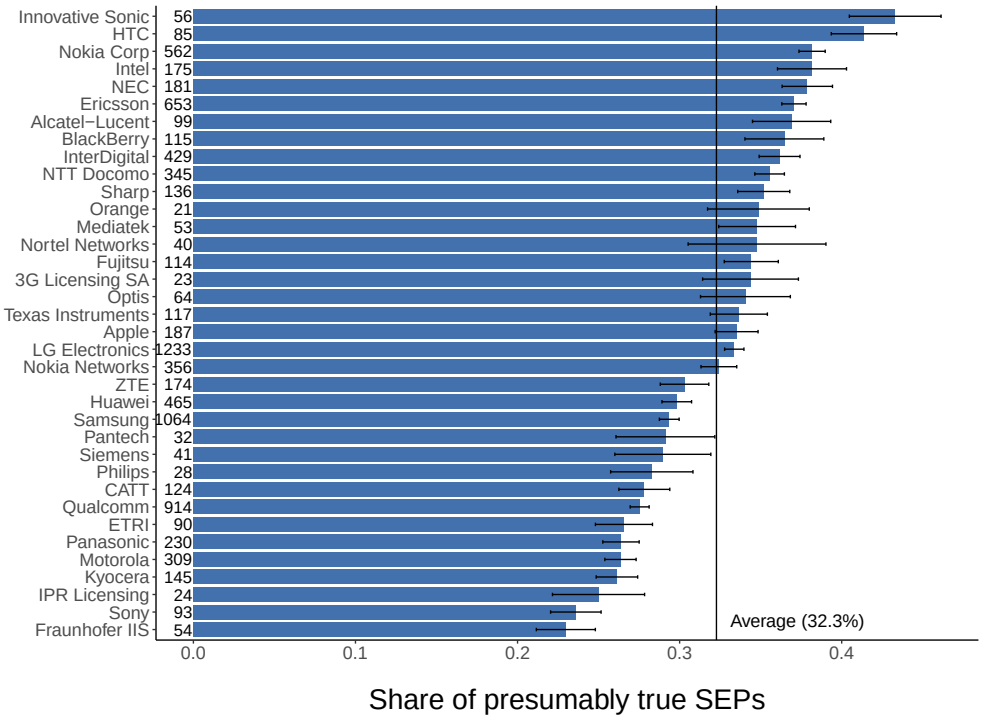
The left-hand side variable $\widehat{\Delta}$ is the difference in the share of true SEPs for actual assessments and predictions and N the portfolio size, i.e., the number of patents for a given patent portfolio. We include no additional constant in the power law function such that the function approaches zero as $N \to \infty$. The fitted function allows us to determine error rates for SEP portfolios of larger size than those in the test dataset. For instance, in a large SEP portfolio of 1,000 declared SEP patent families, the error function yields a prediction error as low as 1.2 pp.

In Figure 6, we present out-of-sample predictions for firm SEP portfolios separately for all three standard generations. In Figure 6a, the overall share of presumably true SEPs for LTE standards is approximately 32.3%, which is 3.6 pp lower than the benchmark evaluations in the manual SEP assessments sample. On firm portfolio level, the share of presumably true SEPs varies substantially from 22.9% to 43.3%. The highest-ranked firm has a

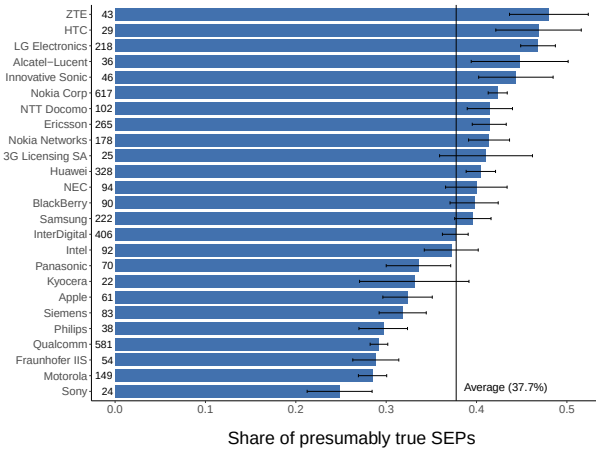
[37] The error functions for UMTS and GSM standards are qualitatively very similar (see Figure A-3 in the Appendix).

Figure 6: SEP firm portfolios for telecommunication standards (out-of-sample predictions)

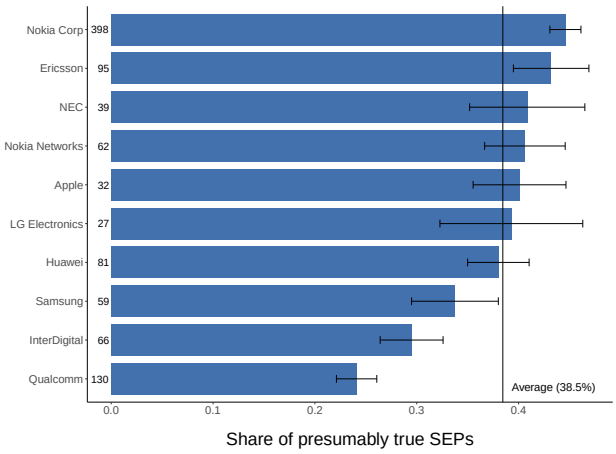
(a) LTE



(b) UMTS



(c) GSM



Notes: The top graph shows the out-of-sample predictions on firm-level for LTE, the lower left-hand graph for UMTS and the lower right-hand graph for GSM patents. The numbers on the left-hand side of the bars indicate the count of patent families declared to LTE/UMTS/GSM standards by the respective firm. Only results for firms with 20 or more declared patents are reported. 95% confidence intervals are shown.

share of presumably true SEPs that is nearly twice as large as the one for the lowest-ranked firm. Notably, there seems no strong correlation between the share of presumably true SEPs and portfolio size. In Figure 6b and Figure 6c, 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 (37.7% for UMTS and 38.5% for GSM).[38] We leave the question as to what causes this trend for future work. However, one candidate explanation might lie in the changing composition of companies contributing technical inventions to standards. First, over time the number of firms holding a portfolio of at least 20 declared SEPs has increased sharply. Second, over time the set of patent holders has become more diverse in terms of their business models. Specifically, there is an increasing number of patent holders that are upstream technology contributors with few, if any, activities in the product market. The average ownership shares of presumably true SEPs are lower (by 3.6 pp or by about 10.2%) for entities that are primarily *upstream* technology contributors, as compared to entities that are primarily *downstream* standard implementers.[39] This finding is consistent with the notion that entities with upstream business models rely more on licensing revenues (for which the number of SEPs matters most) than firms with product market activities (cf. Dewatripont and Legros, 2013; Bekkers et al., 2017).

As an alternative explanation, the decrease in the portfolio share of true SEPs may stem from strategic behavior and learning. Firms may simply have adapted their behavior to the observations that owning a large portfolio of SEP-declared patents is advantageous in court proceedings and licensing negotiations.

7 Discussion and Conclusion

In this paper, we propose a novel automated procedure that calculates the semantic similarity between patents and technical standards. We show that this similarity measure serves as a meaningful approximation of standard essentiality.

Three validation exercises confirm our measure's validity. First, we compare pairs of SEPs and the associated standards to control groups of technologically similar patents and standard documents within the same standardization project. We observe a significantly higher semantic similarity for standard-patent pairs defined by SEP declarations. This allows to conclude that the semantic approach is suitable for measuring technological similar-

[38] Some of these firms are primarily known for being both developers and implementers of more recent standards such as UMTS and LTE. Nonetheless, they also made SEP declarations to later releases of the older GSM standard (GSM Phase 2+).

[39] This descriptive finding at the level of (weighted) SEP firm portfolios is corroborated in a multivariate regression at the level of the single SEP (see Table C-4 in the Online Appendix).

ity between patents and standards. Second, we replicate the study of Bekkers et al. (2017) and examine the 'disclosure effect' of SEP declarations on patent forward citations. We build subsamples of SEPs (declared at ETSI) with high semantic similarity to their associated standards. In contrast to Bekkers et al. (2017), we are able to find a positive disclosure effect for these subsamples, as predicted for truly standard-essential patents. Third, we exploit information on available data of manual essentiality checks for a sample of SEPs for mobile telecommunication standards. We find strong and highly significant correlations between the experts' decisions on standard essentiality and our measure of semantic similarity.

Of course, a text-based determination of standard essentiality comes with some limitations. When inventors and patent attorneys draft a patent, they may either use their own words or borrow the terminology from standard documents. The calculated similarity scores will likely differ even if the underlying technology is the same. This introduces potential bias in our measure, especially if patent wording becomes a strategic choice and the processes of patent filing and standard drafting coincide temporally and/or personally. Figure A-2 in the Appendix illustrates the relationship between similarity scores of SEPs by filing year relative to the year of standard publication. Whereas SEPs that are filed shortly before the respective standard gets published have the highest average similarity, patents filed much earlier or after the standard publication have a lower average similarity. We invite future research to build on this stylized finding and further examine the dynamic aspects of standard essentiality.

Moreover, a patent's claims solely define its scope of protection and thus essentiality. Claims are typically written in a highly abstract and generic language, which complicates a semantics-based analysis. The algorithm we deploy makes, by default, use of both patent description and patent claims. However, we explore input-specific differences for our similarity measure in additional robustness checks (see Online Appendix D) and find that this alternative similarity score, which is only based on claim text, also shows a statistically significant relationship with standard essentiality. Even so, the explanatory value of the similarity measure remains higher when we include the description of the patent alongside the patent claims as input text. This is not too surprising; interviewed patent attorneys confirmed that a patent's description and drawings are frequently considered in manual essentiality checks.[40]

In the first use case of our method, we estimate the shares of (presumably) true SEPs in firm patent portfolios. In doing so, we benefit from the high accuracy of our approach when predicting standard essentiality on the aggregate level. Based on manual SEP assessments, we present out-of-sample predictions for firms' true shares of SEPs for GSM, UMTS and

[40] As a matter of fact, patent law explicitly states that the description and drawings shall be used to interpret the claims (see, e.g., Art. 69(1) of the European Patent Convention).

LTE standards. We find statistically and economically substantial differences. The highest-ranked firm has a share of presumably true SEPs which is approximately twice as large as the one for the lowest-ranked firm. Another interesting finding from this analysis is the decline in the average share of presumably true SEPs over the three generations of mobile telecommunication standards. This pattern may be due to an increasing number of *upstream* technology contributors, which on average have a lower share of presumably true SEPs than vertically integrated firms with product market activities. This purely descriptive, yet intriguing result may be a worthwhile subject for future work on standards and firm behavior.

Beyond this first use case, we see several potential applications of our method in the academic as well as the practical realm. Specifically, it may facilitate the assessment of SEPs as well as the search for relevant, but (so far) undeclared patents. Even though our method can hardly replace a thorough manual assessment, its suitability for initial patent screenings can make it a valuable tool for SSOs and firms alike. Furthermore, our approach may help singling out patents relevant for specific parts of the standard. In turn, this would, for instance, allow for a mapping of patents to particular standard technologies such as radio transmission, base stations, or user equipment. Finally, we would like to stress that our approach offers substantial advantages, in particular in terms of scalability as well as time- and cost-efficiency. Moreover, the data generated through our method are arguably more objective and replicable than most of the proprietary datasets on SEP assessments. Against this backdrop, we hope our method will invite more scholars to empirically study the important, yet complex relationship between patents and standards.

References

Abbas, A., L. Zhang, and S. U. Khan (2014). A Literature Review on the State-of-the-Art in Patent Analysis. *World Patent Information 37*, 3–13.

Arts, S., B. Cassiman, and J. C. Gomez (2018). Text Matching to Measure Patent Similarity. *Strategic Management Journal 39*(1), 62–84.

Baron, J. and T. Pohlmann (2018). Mapping Standards to Patents Using Declarations of Standard-essential Patents. *Journal of Economics & Management Strategy 27*(3), 504–534.

Baron, J. and D. F. Spulber (2018). Technology Standards and Standard Setting Organizations: Introduction to the Searle Center Database. *Journal of Economics & Management Strategy 27*(3), 462–503.

Bekkers, R., R. Bongard, and A. Nuvolari (2011). An Empirical Study on the Determinants of Essential Patent Claims in Compatibility Standards. *Research Policy 40*(7), 1001–1015.

Bekkers, R., C. Catalini, A. Martinelli, C. Righi, and T. Simcoe (2017). Disclosure Rules and Declared Essential Patents. NBER Working Paper No. 23627.

Bekkers, R., A. Martinelli, and F. Tamagni (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. and A. S. Updegrove (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, Washington, D.C.

Berger, F., K. Blind, and N. Thumm (2012). Filing Behaviour Regarding Essential Patents in Industry Standards. *Research Policy 41*(2), 2016–225.

Chiao, B., J. Lerner, and J. Tirole (2007). The Rules of Standard-Setting Organizations: An Empirical Analysis. *The RAND Journal of Economics 38*(4), 905–930.

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*, Chapter 13. Cambridge: 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. Available at: https://ssrn.com/abstract=3100976.

Contreras, J. L. (2019). Technical Standards, Standards-Setting Organizations and Intellectual Property: A Survey of the Literature (with an Emphasis on Empirical Approaches). In *Research Handbook on the Economics of Intellectual Property Law Vol. 2 – Analytical Methods*. Cambridge University Press.

Contreras, J. L., F. Gaessler, C. Helmers, and B. J. Love (2017). Litigation of Standards-Essential Patents in Europe: A Comparative Analysis. *Berkeley Technology Law Journal 32*, 1457.

Dewatripont, M. and P. Legros (2013). 'Essential Patents', FRAND Royalties and Technological Standards. *The Journal of Industrial Economics 61*(4), 913–937.

EC (2013). Study on the Interplay between Standards and Intellectual Property Rights. Final report. Tender No. ENTR/09/015 (OJEU S136 of 18/07/2009).

EC (2014). Patents and Standards: A Modern Framework for IPR-based Standardization. Ref. Ares(2014)917720 - 25/03/2014.

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.

Gambardella, A., D. Harhoff, and B. Verspagen (2017). The Economic Value of Patent Portfolios. *Journal of Economics & Management Strategy 26*(4), 735–756.

Goodman, D. J. and R. A. Myers (2005). 3G Cellular Standards and Patents. IEEE WirelessCom.

Hussinger, K. and F. Schwiebacher (2015). The Market Value of Technology Disclosures to Standard Setting Organizations. *Industry and Innovation 22*(4), 321–344.

Jürgens, B. and N. Clarke (2018). Study and Comparison of the Unique Selling Propositions (USPs) of Free-to-Use Multinational Patent Search Systems. *World Patent Information 52*, 9–16.

Kang, B. and R. Bekkers (2015). Just-in-time Patents and the Development of Standards. *Research Policy 44*(10), 1948–1961.

Kang, B. and K. Motohashi (2015). Essential Intellectual Property Rights and Inventors' Involvement in Standardization. *Research Policy 44*(2), 483–492.

Larouche, P. and N. Zingales (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. Available at: 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. and C. Shapiro (2006). Patent Holdup and Royalty Stacking. *Texas Law Review 85*, 1991–2048.

Lemley, M. A. and T. Simcoe (2019). How Essential are Standard-Essential Patents? *Cornell Law Review 104*(3), 607–642.

Lerner, J. and J. Tirole (2015). Standard-Essential Patents. *Journal of Political Economy 123*(3), 547–586.

Magerman, T., B. Van Looy, and X. Song (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. Available at: 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.

Omachi, M. (2004). Emergence of Essential Patents in Technical Standards: Implications of the Continuation and Divisional Application Systems and the Written Description Requirement.

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. Available at: https://ssrn.com/abstract=3177975.

Pohlmann, T., P. Neuhäusler, and K. Blind (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.

Rysman, M. and T. Simcoe (2008). Patents and the Performance of Voluntary Standard-Setting Organizations. *Management Science 54*(11), 1920–1934.

Shapiro, C. (2001). Navigating the Patent Thicket: Cross Licensing, Patent Pools, and Standard Setting. In A. B. Jaffe, J. Lerner, and S. Stern (Eds.), *Innovation Policy and the Economy*, Volume 1, pp. 119–150. M.I.T. Press, Cambridge.

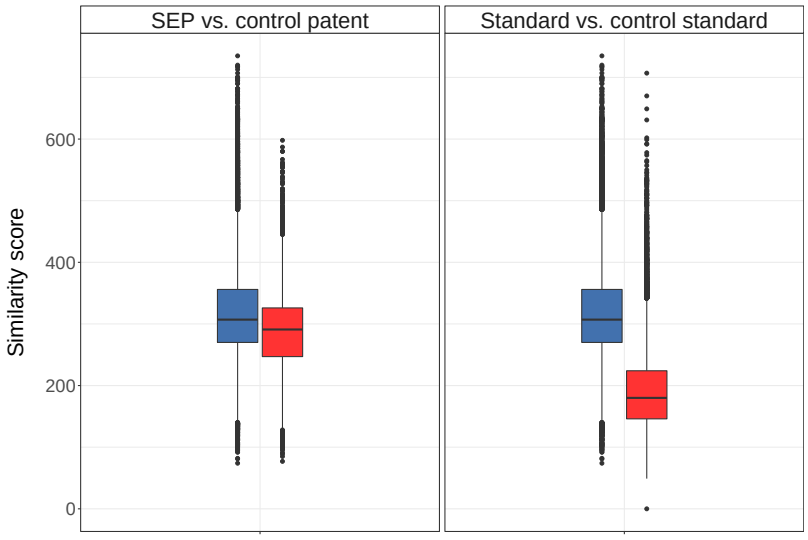
Spulber, D. F. (2019). Standard Setting Organisations and Standard Essential Patents: Voting and Markets. *The Economic Journal 129*(619), 1477–1509.

Stitzing, R., P. Sääskilahti, J. Royer, and M. V. Audenrode (2017). Over-Declaration of Standard Essential Patents and Determinants of Essentiality. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2951617.

Younge, K. and J. Kuhn (2016). Patent-to-Patent Similarity: A Vector Space Model. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2709238.

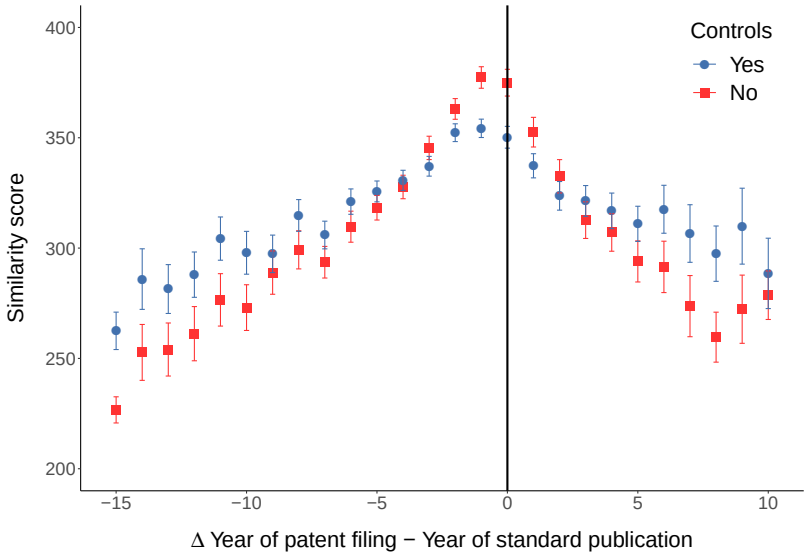
A Appendix: Figures

Figure A-1: Comparison of SEP - standard pairs with control groups (censored data)



Notes: The box plot on the left-hand side shows the difference in similarity scores of SEP declarations (blue) and control patents compared to the same standard document (red). On the right-hand side, similarity scores of SEP declarations (blue) are compared to similarity scores of the same SEP and control standard documents (red). The censored dataset is used in this representation. Differences are significant but considerably less pronounced relative to the results when using the truncated data. Statistics are shown in Table B-1.

Figure A-2: Similarity scores of SEPs by filing year relative to the year of standard publication

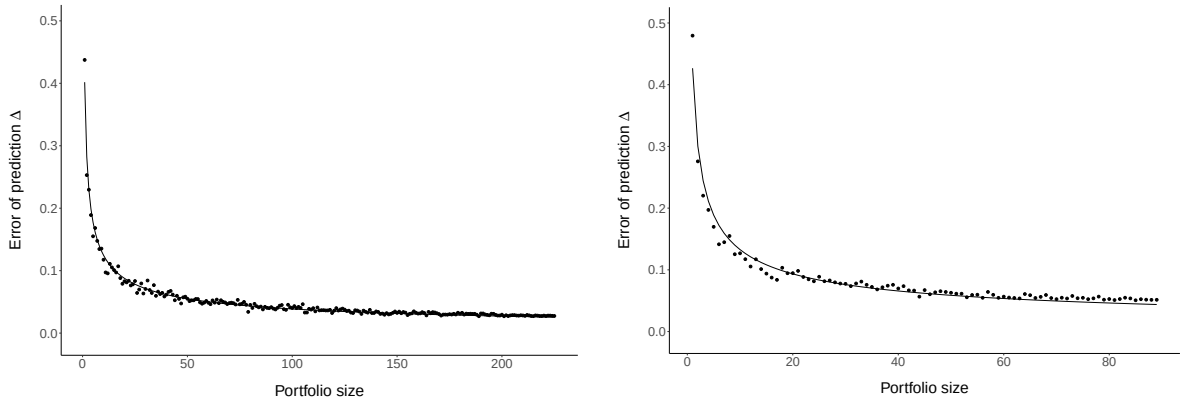


Notes: The graph shows similarity scores as predicted by OLS regression models of similarity scores on fixed effects for the year differences between standard publication and patent filing. Differences are left-censored at −15 and right-censored at 10, respectively. We report results for regressions on patent family level with and without control variables. Controls include patent characteristics as used in column (2) in Table C-3 as well as declaration year, TS and CPC-4 fixed effects. 90% confidence intervals are shown.

Figure A-3: The error of prediction as a function of portfolio size

(a) UMTS

(b) GSM



Notes: The error of prediction Δ is plotted as a function of portfolio size where portfolios are randomly drawn from the test sample of UMTS and GSM patents. Non-linear least squares fits are shown. The fitted functions are power law functions.

B Appendix: Tables

Table B-1: T-statistics for the comparison of SEP - standard pairs with control groups

	t-value	$\Delta Score$
Uncensored		
SEP vs. control patent	61^{***}	59
Standard vs. control standard	127^{***}	135
Censored		
SEP vs. control patent	51^{***}	31
Standard vs. control standard	189^{***}	124

Notes: *** indicate significance levels of $p < 2 \times 10^{-16}$. $\Delta Score$ denotes the differences in mean similarity scores for both groups.

Table B-2: Summary statistics (manual SEP assessments sample)

	Mean	SD	Median	Min	Max	N
LTE Essentiality	0.3590	0.4800	0	0	1	1470
UMTS Essentiality	0.3970	0.4900	0	0	1	794
GSM Essentiality	0.3880	0.4880	0	0	1	304
Similarity score	369.3690	108.9510	373	62	758	2163
Patent family size	12.8580	12.5130	10	1	269	2197
# Inventors	3.0030	1.6970	3	1	13	2197
# Applicants	2.1760	1.8360	1	1	13	2197
# Independent claims	4.1210	2.6050	4	1	18	2197
Length claim 1	134.3880	60.2660	125	1	388	2197
# Patent references	27.1880	36.5770	18	0	911	2197
# NPL references	30.1530	67.5580	11	0	1188	2197
# SEP US fwd. cit. (5yrs)	7.2710	9.9180	4	0	122	2014
Section-specific decl.	0.3690	0.4830	0	0	1	2014
SEP transferred	0.0810	0.2740	0	0	1	2197
Earliest decl. year	2009.9900	3.2690	2010	1998	2016	1951
Priority year	2005.3490	3.8030	2006	1989	2012	2197

Notes: Summary statistics for the sample of patent families which were manually scrutinized by technical experts.

Table B-3: Predicting standard essentiality

	(1) LTE	(2) UMTS	(3) GSM
Similarity score	0.0747***	0.1314***	0.1424***
	(0.0127)	(0.0183)	(0.0327)
SEP transferred (d)	−0.0809	0.0283	0.1182
	(0.0493)	(0.0680)	(0.1071)
# Independent claims	0.0002	0.0051	0.0122
	(0.0043)	(0.0044)	(0.0078)
Length claim 1	−0.0005**	−0.0002	−0.0008
	(0.0002)	(0.0003)	(0.0005)
# Inventors	−0.0089	−0.0090	−0.0255
	(0.0084)	(0.0128)	(0.0245)
# Applicants	−0.0009	−0.0154	0.0029
	(0.0076)	(0.0125)	(0.0211)
Patent family size	0.0031**	0.0069***	0.0049*
	(0.0016)	(0.0018)	(0.0028)
# Patent references	−0.0000	−0.0026***	−0.0016
	(0.0004)	(0.0010)	(0.0016)
# NPL references	0.0007**	0.0002	0.0002
	(0.0003)	(0.0004)	(0.0007)
# SEP US fwd. cit. (5yrs)	0.0033***	−0.0002	−0.0074
	(0.0012)	(0.0024)	(0.0059)
Section-specific decl. (d)	0.0904***	0.0237	0.1070*
	(0.0276)	(0.0390)	(0.0621)
Pseudo R^2	0.05	0.08	0.09
AUC	0.66	0.69	0.70
Observations	1,441	731	280

Notes: The dependent variable is a dummy equal to one if the patent family is truly essential. Regression results for the three telecommunication standards LTE, UMTS and GSM are reported. AUC = Area under ROC-Curve. Similarity scores refer to the most similar chapter for any standard in the dataset. Similarity scores are divided by 100. Marginal effects of one unit change are reported. For binary variables (d) following the variable name indicates a discrete change from 0 to 1. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table B-4: Confusion Matrix

		De facto SEPs	
		No	Yes
Prediction	No	216	126
	Yes	20	40

Notes: Confusion matrix for the test set of LTE SEPs evaluated by the manual SEP assessments data.

Table B-5: Predicting standard essentiality with most relevant characteristics

	(1) LTE	(2) UMTS	(3) GSM
Similarity score	0.0762***	0.1244***	0.1360***
	(0.0125)	(0.0176)	(0.0311)
Length claim 1	−0.0005**	−0.0000	−0.0005
	(0.0002)	(0.0003)	(0.0005)
# NPL references	0.0009***	0.0001	0.0000
	(0.0003)	(0.0003)	(0.0005)
# SEP US fwd. cit. (5yrs)	0.0034***	0.0005	−0.0026
	(0.0012)	(0.0022)	(0.0045)
Section-specific decl. (d)	0.0976***	0.0430	0.1383**
	(0.0269)	(0.0382)	(0.0601)
Pseudo R^2	0.05	0.06	0.07
AUC	0.66	0.66	0.67
Observations	1,441	731	280

Notes: Specifications for out-of-sample predictions presented in Section 6. The dependent variable is a dummy equal to one if the patent family is truly essential. Regression results for the three telecommunication standards LTE, UMTS and GSM are reported. AUC = Area under ROC-Curve. Similarity scores refer to the most similar chapter for any standard in the dataset. Similarity scores are divided by 100. Marginal effects of one unit change are reported. For binary variables (d) following the variable name indicates a discrete change from 0 to 1. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

ONLINE APPENDIX

C Additional Analysis: Characteristics of SEPs

Standard-essential patents, but also patents which are technologically close but not necessarily essential to technical standards, belong to a special group of patents with possibly high economic and technological value. To learn more about these patents, we correlate our novel measure of similarity with various patent characteristics. First, we consider the full sample of patent families which appear in our dataset. Summary statistics for this sample are reported in Table C-1. Secondly, we consider a subsample of declared SEPs in the dataset and examine correlations with various patent characteristics.

Table C-1: Summary statistics (full sample)

	Mean	SD	Median	Min	Max	N
Similarity score	180.5240	72.7920	166	37	945	1762842
Similarity rank	1174.5370	883.7230	1021	1	3000	1762842
Granted US patent	0.4740	0.4990	0	0	1	1708537
# US fwd. cit. (5yrs)	11.9130	28.6770	4	0	3264	1320969
# Independent claims	3.2940	2.3880	3	1	39	920780
Length claim 1	132.5870	78.0280	115	0	499	912474
Patent family size	3.0580	3.1680	2	1	472	1708537
# Patent references	12.4550	22.4960	7	0	4148	1708537
# NPL references	3.5080	24.9550	0	0	12854	1708537
# Applicants	1.5550	1.3220	1	1	77	1687964
# Inventors	2.3780	1.6920	2	1	133	1700801
Priority year	2003.7120	9.5470	2006	1950	2017	1707870

Notes: Summary statistics for patent characteristics of all patents in the dataset. Patent characteristics are on patent family level.

We first consider the full sample of all standards-related patents. In Table C-2, we correlate patent characteristics with the measure *Similarity score* in columns (1) and (2), and the relative measure *Similarity rank* in columns (3) and (4). We include fixed effects for CPC-4 technology classes as well as for technical specifications on document level. In columns (1) and (3), we report significant and positive effects for forward citations and patent family size, which have been used as proxies for patent value in the literature. Furthermore, we find a negative relationship between patent grant and the similarity to a technical standard. We additionally include claim characteristics in the specifications in columns (2) and (4) and find that more independent claims are associated with a higher likelihood of being similar to standards. Furthermore, the length of the first claim is negatively correlated with

similarity suggesting that patents with broader claims are more similar to standards.

Table C-2: Correlation of standards similarity with patent characteristics

	(1) Score	(2) Score	(3) Rank	(4) Rank
# US fwd. cit. (5yrs)	0.0427***	0.0386***	−0.2872***	−0.1688***
	(0.002)	(0.002)	(0.029)	(0.033)
Granted US patent	−6.8750***		123.6418***	
	(0.112)		(1.680)	
Patent family size	0.5398***	0.8150***	−5.2494***	−8.6861***
	(0.016)	(0.020)	(0.240)	(0.290)
# Patent references	−0.0441***	−0.0426***	0.6177***	0.5362***
	(0.002)	(0.003)	(0.037)	(0.037)
# NPL references	0.0058*	−0.0099***	−0.1032**	0.1269***
	(0.002)	(0.002)	(0.035)	(0.035)
# Applicants	−0.1676***	−0.4248***	4.3477***	7.5940***
	(0.041)	(0.047)	(0.615)	(0.698)
# Inventors	0.0603	0.1327**	0.3340	−0.4327
	(0.033)	(0.042)	(0.490)	(0.618)
# Independent claims		0.0766**		0.0129
		(0.027)		(0.400)
Length claim 1		−0.0283***		0.3343***
		(0.001)		(0.013)
Priority year	Yes	Yes	Yes	Yes
CPC-4 FE	Yes	Yes	Yes	Yes
TS FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.41	0.44	0.14	0.16
Observations	1267261	734587	1267261	734587

Notes: OLS regressions of similarity measures on patent family characteristics. The dependent variables *similarity score* and *similarity rank* are abbreviated as *score* and *rank*, respectively. The sample consists of all patents in the full dataset. Standard errors are in parentheses. Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Table C-3 reports correlations of the similarity score with SEP characteristics that show some differences compared to the full sample of patents. We include fixed effects for CPC-4 technology class and technical specification (TS). Considering column (1), we do not observe an effect of forward citations on similarity. Only after including SEP forward citations, we find a statistically significant, negative effect of patent forward citations, whereas SEP

Table C-3: Correlation of standards similarity with SEP characteristics

	(1) Score	(2) Score	(3) Rank	(4) Rank	(5) Score	(6) Score
# US fwd. cit. (5yrs)	0.0278	−0.1595***	−0.0255	0.4681***	0.1199***	−0.1570***
	(0.019)	(0.029)	(0.098)	(0.148)	(0.028)	(0.044)
Granted US patent	1.2951	0.8413	35.8341***	37.0302***	4.8581	3.3348
	(2.074)	(2.069)	(10.422)	(10.418)	(4.818)	(4.769)
Patent family size	0.0470	0.1074	−0.1480	−0.3070	−0.3194*	−0.2312
	(0.110)	(0.110)	(0.552)	(0.553)	(0.188)	(0.186)
# Patent references	−0.1902***	−0.1931***	1.1873***	1.1950***	−0.0570	−0.0736*
	(0.034)	(0.034)	(0.172)	(0.172)	(0.044)	(0.044)
# NPL references	0.0815***	0.0768***	−0.5208***	−0.5084***	0.0516	0.0403
	(0.026)	(0.026)	(0.129)	(0.128)	(0.036)	(0.036)
# Applicants	−0.3917	−0.5430	6.8290***	7.2278***	−0.1772	−0.2425
	(0.497)	(0.496)	(2.499)	(2.499)	(0.839)	(0.830)
# Inventors	−1.0402**	−1.2051**	2.9709	3.4056	−1.0890	−1.4472*
	(0.498)	(0.497)	(2.501)	(2.501)	(0.871)	(0.863)
Section-specific decl. (d)	1.4308	1.7074	−32.9312***	−33.6603***	11.5435***	11.2724***
	(1.822)	(1.817)	(9.156)	(9.151)	(3.398)	(3.360)
# SEP US fwd. cit. (5yrs)		1.3646***		−3.5967***		1.8782***
		(0.160)		(0.807)		(0.234)
Priority year	Yes	Yes	Yes	Yes	Yes	Yes
Earliest decl. year	Yes	Yes	Yes	Yes	Yes	Yes
CPC-4 FE	Yes	Yes	Yes	Yes	Yes	Yes
TS FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.47	0.47	0.16	0.17	0.57	0.58
Observations	13,330	13,330	13,330	13,330	3,239	3,239

Notes: OLS regressions of similarity measures on patent family characteristics. The dependent variables *similarity score* and *similarity rank* are abbreviated as *score* and *rank*, respectively. The sample consists of SEPs declared at ETSI. Standard errors are in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

forward citations are positively related to standards similarity. SEPs that are declared to specific sections of a standard are relatively more similar to the standard (see columns (3) and (4)). For the analyses in columns (5) and (6), we reconstructed the sample used in Stitzing et al. (2017), which comprises 3,239 US SEPs declared to LTE standard documents until 2013. We observe small effects for forward citations and significantly larger effects for

SEP forward citations. This result mirrors the results of Stitzing et al. (2017), in particular the correlation between their patent forward citation measure and LTE standard essentiality. They also state that SEPs declared to specific technical specifications are more likely to be essential – a result that we find as well.

In Table C-4, we furthermore correlate our measure of similarity with firm business models. We hereby rely on the classification by Bekkers et al. (2017) and differentiate between up- and downstream firms. These descriptive results reveal substantial differences in the level of semantic similarity between patents and standards across business models. This finding is in line with theoretical predictions that the incentives to generate revenues through SEP licensing is higher for upstream firms (Dewatripont and Legros, 2013).

Table C-4: Correlation of standards similarity with firm business models

	(1) Score	(2) Score	(3) Score	(4) Score	(5) Score	(6) Score
Downstream	35.2386***	33.4471***	37.3720***	32.0548***	19.9433***	19.5235***
	(2.527)	(2.608)	(2.580)	(2.558)	(2.245)	(2.238)
Patent controls	No	Yes	Yes	Yes	Yes	Yes
Earliest filing year	No	No	Yes	Yes	Yes	Yes
Earliest decl. year	No	No	Yes	Yes	Yes	Yes
CPC-4 FE	No	No	No	Yes	No	Yes
Standard doc. FE	No	No	No	No	Yes	Yes
Adjusted R^2	0.02	0.05	0.14	0.20	0.51	0.52
Observations	11,181	11,181	11,181	11,181	11,181	11,181

Notes: OLS regressions of *similarity score* on firm business models. The dependent variable *similarity score* is abbreviated as *score*. Patent controls include the variables reported in column (2) in Table C-3. We differentiate between upstream (baseline) and downstream firms according to the classification by Bekkers et al. (2017). The sample consists of declared SEPs. Robust standard errors are in parentheses. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

D Robustness Checks

In this paper we rely on the semantic algorithm that has the major advantage of searching for the most similar patents (in the entire patent universe with more than 37 million documents) for any input text you enter to the machine. Whereas it is not trivial to replicate such an efficient algorithm, we can test the validity of our main result developing a simple text mining algorithm that is generally used in the literature (see Section 3.1).

For a small subset of our data, we show that measuring standard essentiality using the common text-based approaches is relatively simple. The text mining package 'tm' that is implemented in R helps us to convert the text data into a corpus of documents. We remove any kind of special characters, punctuation, numbers and English stop words. To stem the words in our corpus, we rely on the stemming algorithm by Porter. The pre-processed data are then converted into a (sparse) document-term-matrix. Words are weighted by term frequency-inverse document frequency (tf-idf). We additionally remove highly sparse terms and compute the text similarity between patents and standards using cosine similarity. The comparison conducted for this exercise includes US full text data for patents and full text data for ETSI's LTE standards on chapter level. Furthermore, we also use the text of patent claims (excluding patent description, abstract and title). For both cases, we compare LTE patents assessed by technical experts to their corresponding standard documents identified by its engineers. This procedure yields 117,282 text-based comparisons for only 657 patent families. In Table D-1, we report logistic regression results for the full text comparison using the alternative similarity measures as described before. Table D-2 reports results when semantic similarity calculations are confined to patent claim texts only. By comparing the effect sizes of the similarity score measures in both tables, we find that the coefficients are larger when also patent title, abstract and descriptions are taken into account. This speaks in favor of considering all patent text information.

Furthermore, we compute micro-average precision and recall scores. We obtain 63.4% precision and recall using patent full text data. In comparison, we obtain 62.7% using only claim texts. These values are comparable, yet slightly inferior to the similarity measure based on the proprietary algorithm used in this study.

Table D-1: Logistic regressions: Standard essentiality (alternative measures with patent full text)

	(1)	(2)	(3)	(4)	(5)
Similarity score (alt)		0.1032***	0.1034***	0.1088***	0.0981***
		(0.0197)	(0.0229)	(0.0243)	(0.0272)
SEP transferred (d)	−0.1617*	−0.1234	−0.1362	−0.1932*	−0.2531***
	(0.0837)	(0.0897)	(0.0935)	(0.1088)	(0.0954)
# Independent claims	0.0092	0.0141*	0.0057	0.0057	0.0045
	(0.0082)	(0.0084)	(0.0092)	(0.0104)	(0.0110)
Length claim 1	−0.0006	−0.0005	−0.0007*	−0.0009**	−0.0009*
	(0.0004)	(0.0004)	(0.0004)	(0.0005)	(0.0005)
# Inventors	−0.0110	−0.0020	−0.0072	−0.0049	−0.0013
	(0.0171)	(0.0176)	(0.0190)	(0.0213)	(0.0228)
# Applicants	−0.0103	−0.0156	−0.0096	−0.0116	−0.0078
	(0.0151)	(0.0157)	(0.0172)	(0.0179)	(0.0197)
Patent family size	0.0126***	0.0104***	0.0095**	0.0113**	0.0129**
	(0.0035)	(0.0036)	(0.0042)	(0.0048)	(0.0053)
# Patent references	−0.0004	−0.0002	0.0010	0.0002	−0.0003
	(0.0017)	(0.0017)	(0.0018)	(0.0020)	(0.0021)
# NPL references	−0.0002	−0.0000	0.0000	−0.0002	−0.0001
	(0.0007)	(0.0007)	(0.0008)	(0.0008)	(0.0010)
# SEP US fwd. cit. (5yrs)	0.0039	0.0031	0.0029	0.0028	0.0045
	(0.0029)	(0.0029)	(0.0030)	(0.0033)	(0.0036)
Section-specific decl. (d)	0.1648***	0.1440**	0.1041	0.0642	0.0559
	(0.0556)	(0.0576)	(0.0702)	(0.0982)	(0.1084)
Priority year	No	No	Yes	Yes	Yes
Earliest decl. year	No	No	Yes	Yes	Yes
Firm FE	No	No	No	Yes	Yes
CPC-4 FE	No	No	No	No	Yes
Pseudo R^2	0.05	0.10	0.16	0.20	0.24
AUC	0.66	0.70	0.76	0.79	0.80
Observations	480	480	480	480	480

Notes: The dependent variable is a dummy equal to 1 if the patent family is truly essential for LTE standards. AUC = Area under ROC-Curve. Pairs of SEPs and their most similar standard in the sample of manual SEP assessments are selected for the regressions. For patents the full text is used. The alternative similarity scores are multiplied by 10. Marginal effects of one unit change are reported. For binary variables (d) following the variable name indicates a discrete change from 0 to 1. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table D-2: Logistic regressions: Standard essentiality (alternative measures with claim text only)

	(1)	(2)	(3)	(4)	(5)
Similarity score (alt)		0.0726***	0.0598**	0.0602**	0.0458
		(0.0229)	(0.0253)	(0.0267)	(0.0289)
SEP transferred (d)	−0.1617*	−0.1505*	−0.1467	−0.1953*	−0.2586***
	(0.0837)	(0.0855)	(0.0915)	(0.1074)	(0.0928)
# Independent claims	0.0092	0.0145*	0.0038	0.0014	0.0000
	(0.0082)	(0.0084)	(0.0092)	(0.0103)	(0.0109)
Length claim 1	−0.0006	−0.0006	−0.0008*	−0.0010**	−0.0010**
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0005)
# Inventors	−0.0110	−0.0085	−0.0155	−0.0097	−0.0052
	(0.0171)	(0.0172)	(0.0186)	(0.0210)	(0.0225)
# Applicants	−0.0103	−0.0136	−0.0037	−0.0063	−0.0022
	(0.0151)	(0.0153)	(0.0168)	(0.0176)	(0.0194)
Patent family size	0.0126***	0.0104***	0.0090**	0.0115**	0.0140***
	(0.0035)	(0.0036)	(0.0041)	(0.0048)	(0.0054)
# Patent references	−0.0004	−0.0004	0.0007	0.0001	−0.0005
	(0.0017)	(0.0017)	(0.0018)	(0.0020)	(0.0021)
# NPL references	−0.0002	−0.0003	−0.0001	−0.0002	−0.0002
	(0.0007)	(0.0007)	(0.0007)	(0.0008)	(0.0010)
# SEP US fwd. cit. (5yrs)	0.0039	0.0032	0.0030	0.0033	0.0057
	(0.0029)	(0.0029)	(0.0030)	(0.0033)	(0.0036)
Section-specific decl. (d)	0.1648***	0.1570***	0.0982	0.0632	0.0523
	(0.0556)	(0.0563)	(0.0699)	(0.0967)	(0.1078)
Priority year	No	No	Yes	Yes	Yes
Earliest decl. year	No	No	Yes	Yes	Yes
Firm FE	No	No	No	Yes	Yes
CPC-4 FE	No	No	No	No	Yes
Pseudo R^2	0.05	0.07	0.14	0.18	0.22
AUC	0.66	0.68	0.74	0.77	0.79
Observations	480	480	480	480	480

Notes: The dependent variable is a dummy equal to 1 if the patent family is truly essential for LTE standards. AUC = Area under ROC-Curve. Pairs of SEPs and their most similar standard in the sample of manual SEP assessments are selected for the regressions. The alternative similarity scores are multiplied by 10. Marginal effects of one unit change are reported. For binary variables (d) following the variable name indicates a discrete change from 0 to 1. Standard errors in parentheses. Significance levels: * p<0.1, ** p<0.05, *** p<0.01.